The TAO of Topic Maps

finding the way in the age of infoglut

Steve Pepper
Infostream, Oslo, Norway
pepper@infotek.no
http://www.infotek.no

Abstract:

Topic maps are a new ISO standard for describing knowledge structures and associating them with information resources. As such they constitute an enabling technology for knowledge management. Dubbed "the GPS of the information universe", topic maps are also destined to provide powerful new ways of navigating large and interconnected corpora.

While it is possible to represent immensely complex structures using topic maps, the basic concepts of the model – Topics, Associations, and Occurrences (TAO) – are easily grasped. This paper provides a non-technical introduction to these and other concepts (the IFS and BUTS of topic maps), relating them to things that are familiar to all of us from the realms of publishing and information management, and attempting to convey some idea of the uses to which topic maps will be put in the future. [1]

Introduction

Someone once said that "a book without an index is like a country without a map".

However interesting and worthwhile the experience of driving from A to B without a map might be in its own right, there can be no doubt that when the goal is to arrive at one's destination as quickly as possible (or at least without undue delay), some kind of a map is indispensable.

Similarly, if you are looking for a particular piece of information in a book (as opposed to enjoying the experience of reading it from cover to cover), a good index is an immense asset. The traditional back-of-book index can be likened to a carefully researched and hand-crafted map, and the task of the indexer, as Larry Bonura puts it , "to chart[ing] the topics of the document and [presenting] a concise and accurate map for readers."

In **Troilus and Cressida** Shakespeare used a different metaphor:

And in such indexes (although small pricks

To their subsequent volumes) there is seen

The baby figure of the giant mass

Of things to come at large

but also here there is the same sense of the index replicating, in miniature, the structures of its subject, in order to provide a more manageable view of the whole.

Perhaps it isn't surprising that Shakespeare chose not to use the map metaphor. After all, the art of cartography was still in its infancy in his time ... and so too were communications. Today the situation

[1] *Portions of this paper are based on earlier conference proceedings by the same author and the article .*

is quite different, the sheer speed of modern communications makes accurate and advanced mapping techniques of major importance. One answer to this problem in the realm of transportation is the GPS (Global Positioning System) . The answer in the realm of publishing and information management is the new international standard, Topic Maps .

Up until now there has been no equivalent of the traditional back-of-book index in the world of electronic information. True enough, people have marked up keywords in their word processing documents and used these to generate indexes "automatically", but the resulting indexes have remained firmly within the paradigm of single documents destined to be published on paper. The world of electronic information is quite different, as the World Wide Web has taught us. Here the distinction between individual documents vanishes and the requirement is for indexes to span multiple documents, and in some cases, to cover vast pools of information. In this situation, old-fashioned indexing techniques are pitifully inadequate.

The problem has been recognized for several decades in the realm of document processing, but the methodology used to address it – full text indexing – has only solved part of the problem, as anyone who has used search engines on the internet knows only too well.

The main problem with full text indexes is their lack of discrimination. They index **everything**: Imagine creating a traditional back-of-book index by taking **every single word** in the book, removing a couple of hundred of the most obviously useless ones, and then including **every single usage** of those that remain. Even with some intelligence to allow for inflected forms and synonyms the result would be of no practical use whatsoever. And yet this is basically how a web search engine works (no wonder you always get thousands of irrelevant hits and **still** manage to miss the thing you are looking for!).

That is why new methodologies are called for, and topic maps provide an approach that marries the best of several worlds, including those of traditional indexing, library science and knowledge representation, with advanced techniques of linking and addressing. It is our firm conviction that they will become as indispensable for tomorrow's information providers as maps for the traveller. And once topic maps have become ubiquitous, they will indeed constitute the GPS of the information universe.

Knowledge structures and information management

Before looking at the topic map model itself, let us begin by trying to understand the essence of those navigational aids that are familiar to us from the realm of books and paper-based publishing. We'll start with indexes, and go on to consider glossaries and thesauri.

What is an index, really?

The word "index" means many things, most of which relate to pointing in some way ("index", plural "indicis", is Latin for forefinger, informer, or sign). The sense we are interested in is that given in the **Concise Oxford Dictionary** as:

6. *Alphabetical list, usu. at end of book, of names, subjects, etc., with references;*

A traditional index is in fact a map of the **knowledge** contained in a book; it lists the topics covered, by whatever name users might be expected to want to look them up, and includes salient (and **only** salient) references to those topics. The following example, adapted from , illustrates the basic features:

```
Madama Butterfly, 70-71, 234-236, 326
Puccini, Giacomo, 69-71
soprano, 41-42, 337
Tosca, 26, 70, 274-276, 326
```

The main constituents of this and any index are:

1. an (alphabetical) list of **names of topics**, and

2. references to **occurrences** of those topics

Since our example is taken from a book on opera, the range of topics covers composers, works, and other relevant subjects; in a book on another subject, the **kinds** of topics included would be different, but the principle would be the same. Also, the occurrences in the example above are referred to by page number, but other mechanisms (e.g. section numbers) might also be envisaged (in fact, references are often called "locators" among indexers).

A more complex example illustrates more features of a typical index:

```
La Bohème, 10, 70, 197-198, 326
Cavalleria Rusticana, 71, 203-204
The Girl of the Golden West, see La fanciulla del West
Leoncavallo, Ruggiero
   I Pagliacci, 71-72, 122, 247-249, 326
Madama Butterfly, 70-71, 234-236, 326
Manon Lescaut, 294
Mascagni, Pietro
   Cavalleria Rusticana, 71, 203-204
Puccini, Giacomo, 69-71
   La Bohème, 10, 70, 197-198, 326
   La fanciulla del West, 291
   Madama Butterfly, 70-71, 234-236, 326
   Manon Lescaut, 294
   Tosca, 26, 70, 274-276, 326
   Turandot, 70, 282-284, 326
Rustic Chivalry, see Cavalleria Rusticana
singers, 39-52,
   See also individual names
   baritone, 46
   bass, 46-47
   soprano, 41-42, 337
   tenor, 44-45
soprano, 41-42, 337
tenors, 44-45
Tosca, 26, 70, 274-276, 326
Turandot, 70, 282-284, 326
```

There are a number of new features here:

Typographical conventions are used to distinguish between different **types** of topic (the names of operas are shown in italic);

Similarly, typographical conventions are used to distinguish between different **types** of occurrence (references to synopses are shown in bold);

The use of **see** references allows multiple points of entry (by different names) to the same topic;

See also references point to associated topics;

Subentries provide an alternative mechanism for pointing out associations between different topics (e.g. between a composer and his works, or between supertypes and subtypes).

Indexes can also have yet other features that are not illustrated by our example:

A book may contain **multiple indexes**, for example an index of names, an index of places, and an index of subjects. This mechanism provides an alternative to the use of typographic conventions for distinguishing between topics of different types in one and the same index.

Different topic types might also be distinguished through the use of explanatory labels following the names, e.g. “Tosca (opera)” and “Tosca (character)”.

The locators (page numbers) may contain modifiers that help distinguish between different types of occurrence, for example “54n” for a footnote on page 54. Again, this is an alternative to the use of different typefaces.

The role played by different types of occurrence might also be shown using a subentry mechanism (for example makes heavy use of subentries for roles such as “clause”, “defined in”, “defined in glossary”, “used in production”, etc.).

The key features of a typical index are thus: topics (identified by their names, of which there may be more than one); associations between topics; and occurrences of topics (pointed to via locators). For each of these constructs it is useful to be able to say something about the type, in order to convey more information to the user.

Topics, Associations and Occurrences are also the key constructs in the topic map model (hence the title of this paper). But before discussing that model in more detail, let us look briefly at some related navigational aids (glossaries and thesauri), and at one common method of knowledge representation in the domain of artificial intelligence (semantic networks), since those will broaden our understanding of the kind of structures we are dealing with.

Glossaries and thesauri

A glossary is basically a list of terms and definitions. It can be thought of as a kind of index in which only one type of occurrence is of interest (the one that plays the role of “definition”), and which therefore includes the occurrence inline (instead of pointing to it via a locator). Here is part of the glossary from (slightly amended for the purpose of illustration):

bass: *The lowest of the male voice types. Basses usually play priests or fathers in operas, but they occasionally get star turns as the Devil.*

diva: *Literally, “goddess” – a female opera star. Sometimes refers to a fussy, demanding opera star. See also* ***prima donna****.*

first lady: *See* ***prima donna****.*

Leitmotif *(German, “LIGHT-mo-teef”): A musical theme assigned to a main character or idea of an opera; invented by Richard Wagner.*

prima donna *(“PREE-mah DOAN-na”): Italian for “first lady”. The singer who plays the heroine, the main female character in an opera; or anyone who believes the world revolves around her.*

soprano: *The female voice category with the highest notes and the highest paycheck.*

Like an index, a glossary may also contain **see** and **see also** references to associated topics. It can also (as in this case) contain additional information relevant to the term itself, such as its language or pronunciation, but the key elements are the topic names and their definitions.

A thesaurus, on the other hand, emphasizes other aspects of an index. It is basically a network of interrelated terms within a particular domain, and although it will often contain other information (such as definitions, examples of usage, etc.), the key feature of a thesaurus is the interrelations, or associations, between terms. Given a particular term, a thesaurus will indicate which other terms mean the same, which terms denote a broader category of the same kind of thing, which denote a narrower category, and which are related in some other way. To continue our example from the world of opera, a thesaurus entry might look as follows:

Soprano

definition	The highest category of female (or artificial male) voice.
broader term(s)	vocalist, singer
narrower term(s)	lyric soprano, dramatic soprano, coloratura soprano
related term(s)	mezzo-soprano, treble

The special thing about associations in a thesaurus (as compared to associations found in a typical index or glossary) is that they are **typed**. This is important because it makes it possible not only to say that two terms are related, but also **how** or **why** they are related. It also makes it possible to group together terms that are associated in the same way, thus making navigation much easier. Commonly used association types like “broader term”, “narrower term”, “used for” and “related term” are defined in standards for thesauri such as , and .

Semantic networks

Indexes, glossaries and thesauri are all ways of mapping the knowledge structures that exist implicitly in books and other sources of information. In the field of AI (Artificial Intelligence) there also exists the need to be able to represent knowledge (and meaning), in order to support communication between people and machines. One widely used mechanism is that of **conceptual graphs**, whose building blocks are **concepts** and **conceptual relations**.

In the following conceptual graph for the phrase “man biting dog” , square brackets denote concepts (’man’, ’bite’, ’dog’), and parentheses denote relations (’agent’, ’object’):

```
[man] <- (agent) <- [bite] -> (object) -> [dog]
```

Under various names, such as “semantic nets”, “associative nets”, “partioned nets” and “knowledge” (or “conceptual”) “maps”, such graphs have been implemented in many AI systems. The earliest forms, called **existential graphs**, were invented by the philosopher Charles Sanders Peirce at the end of the 19th century as a graphical notation for symbolic logic. One of the most completely worked out schemes, the conceptual graphs developed by John Sowa and his collaborators , is claimed to be completely isomorphic with first order logic.

Since the basic model of semantic networks is very similar to that of the topics and associations found in indexes, combining the two approaches should provide great benefits in both information management and knowledge management, and this is precisely what the new topic map standard achieves. By adding the topic/occurrence axis to the topic/association model, topic maps provide a means of “bridging the gap” between knowledge representation and the field of information management.

“Knowledge management” is of course one of today’s buzzwords and a term that often involves not a little marketing hype. For the big consulting companies, knowledge management is essentially about new business management techniques designed to address the fact that people (and the expertise they possess) are the primary assets in an increasingly knowledge-based economy. Others equate knowledge management with information management (especially some vendors of information management tools, who are only too happy to slap a new label on their boxes).

But knowledge is fundamentally different from information: the difference is that between knowing a thing versus simply having information about it. And if, as one writer claims , "knowledge management covers three main knowledge activities: generation, codification, and transfer", then topic maps can be regarded as the standard for codification that is the necessary prerequisite for the development of tools that assist in the generation and transfer of knowledge.

The TAO of topic maps

The genesis of topic maps is to be found back in the early 1990's when what later became known as the Davenport Group was discussing ways of enabling the interchange of computer documentation. The group went on to develop DocBook , one of the most widely used DTDs for authoring SGML and XML documents.

One of the problems the Davenport Group faced was that of merging the indexes of different sets of documentation, and the insight they arrived at was that:

indexes, if they have any self-consistency at all, conform to models of the structure of the knowledge available in the materials that they index. But the models are implicit, and they are nowhere to be found! If such models could be captured formally, then they could guide and greatly facilitate the process of merging modelled indexes together. [2]

What eventually became the topic map standard almost ten years later was thus based from the start on the basic concepts embodied in indexes, that is Topics, Associations, and Occurrences – the TAO of topic maps. The following sections explain these and the additional concepts of Identity, Facets and Scope – the IFS of topic maps.

T is for Topic

Topics are clearly the most fundamental concept in all the structures discussed in the first part of this paper, and so too in topic maps.

Topics

What then is a topic? 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.

You can't get much more general than that!

In fact, this is almost word for word how the topic map standard defines **subject**, the term used for the real world "thing" that the topic itself stands in for. We might think of a "subject" as corresponding to what Plato called an **idea**. A topic, on the other hand, is like the shadow that the idea casts on the wall of Plato's cave: It is an object within a topic map that represents a subject. In the words of the standard: "The invisible heart of every topic link is the subject that its author had in mind when it was created. In some sense, a topic reifies a subject..."

Strictly speaking, the term "topic" refers to the element in the topic map document (the **topic link**) that represents the **subject** being referred to. However, in this article it is used more loosely to denote both of

[2] *Personal communication from Steven R. Newcomb, one of the original developers of the topic map model.*

these things together. Whenever there is a need to distinguish between the two, we use the terms “topic link” and “subject”.

So, in the context of a **dictionary of opera**, a topic might represent subjects such as “Tosca”, “Madame Butterfly”, “Rome”, “Italy”, the composer “Giacomo Puccini”, or his birthplace, “Lucca”: that is, anything that might have an entry in the dictionary – but also much else besides.

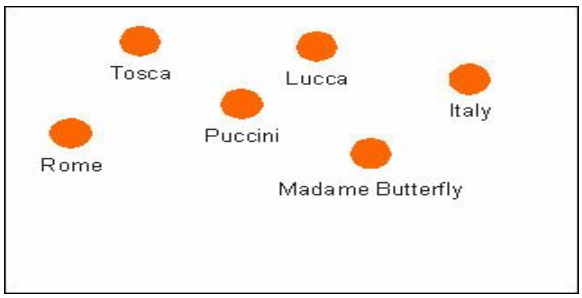

Figure 1. Topics

Topic types

Topics can be categorized according to their kind. In a topic map, any given topic is an instance of zero or more **topic types**. This corresponds to the categorization inherent in the use of multiple indexes in a book (index of names, index of works, index of places, etc.), and to the use of typographic and other conventions to distinguish different types of topics.

Thus, Puccini would be a topic of type “composer”, Tosca and Madame Butterfly topics of type “opera”, Rome and Lucca topics of type “city”, Italy a topic of type “country”, etc. In other words, topic types represent a typical **class-instance** relationship.

Exactly what one chooses to regard as topics in any particular application will vary according to the needs of the application, the nature of the information, and the uses to which the topic map will be put: In a **thesaurus**, topics would represent terms, meanings, and domains; in **software documentation** they might be functions, variables, objects, and methods; in **legal publishing**, laws, cases, courts, concepts, and commentators; in **technical documentation**, components, suppliers, procedures, error conditions, etc.

Topic types are themselves defined **as topics** by the standard. You must explicitly declare “composer”, “opera”, “city”, etc. as topics in your topic map if you want to use them as types (in which case you will be able to say more about them using the topic map model itself).

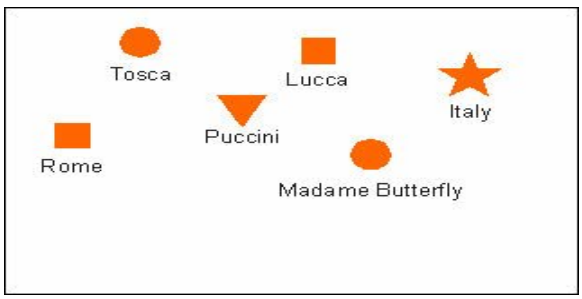

Figure 2. Topic types

Topics have three kinds of characteristics: names, occurrences, and roles in associations.

Topic names

Normally topics have explicit names, since that makes them easier to talk about. [3] However, topics don't **always** have names: A simple cross reference, such as "see page 97", is considered to be a link to a topic that has no (explicit) name.

Names exist in all shapes and forms: as formal names, symbolic names, nicknames, pet names, everyday names, login names, etc. The topic map standard doesn't pretend to try to enumerate and cover them all. Instead, it recognizes the need for some forms of name (that have particularly important and universally understood semantics) to be defined in a standardized way, in order for applications to be able to do something meaningful with them, and at the same time the need for complete freedom and extensibility to be able to define application-specific name types.

The standard therefore provides an element form for **topic name**, which it allows to occur zero or more times for any given topic, and to consist of one or more of the following types of name:

base name (required)

display name (optional)

sort name (optional)

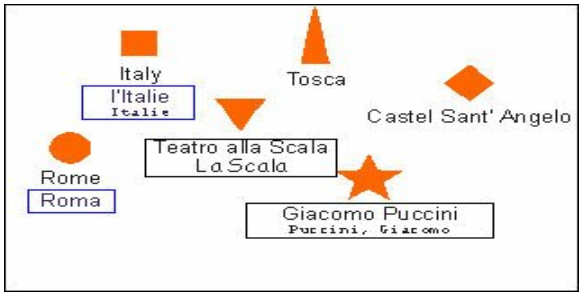

Figure 3. Topic names

The ability to be able to specify more than one topic name can be used to indicate the use of different names in different contexts or **scopes** (about which more later), such as language, style, domain, geographical area, historical period, etc. A corollary of this feature is the **topic naming constraint**, which states that no two subjects can have exactly the same name in the same scope.

O is for Occurrence

Occurrences

A topic may be linked to one or more information resources that are deemed to be relevant to the topic in some way. Such resources are called **occurrences** of the topic.

An occurrence could be a monograph devoted to a particular topic, for example, or an article about the topic in an encyclopaedia; it could be a picture or video depicting the topic, a simple mention of the topic in the context of something else, a commentary on the topic (if the topic were a law, say), or any of a host of other forms in which an information resource might have some relevance to the subject in question.

[3] *It should be clear that the preceding paragraphs would have been rather more difficult to understand if we hadn't given names to our topics and topic types!*

Such occurrences are generally outside the topic map document itself (although some of them could be inside it), and they are "pointed at" using whatever mechanisms the system supports, typically HyTime addressing or XPointers. Today, most systems for creating hand-crafted indexes (as opposed to full text indexes) use some form of embedded markup in the document to be indexed. One of the advantages to using topic maps, is that the documents themselves do not have to be touched.

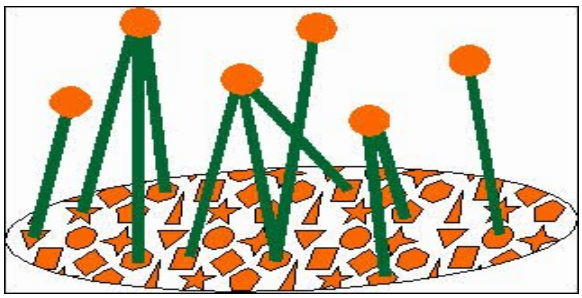

Figure 4. Occurrences

An important point to note here is the **separation into two layers** of the topics and their occurrences. This separation is one of the clues to the power of topic maps and we shall return to it later.

Occurrence roles

Occurrences, as we have already seen, may be of any number of different types (we gave the examples of "monograph", "article", "illustration", "mention" and "commentary" above). Such distinctions are supported in the standard by the concepts of **occurrence role** and **occurrence role type**.

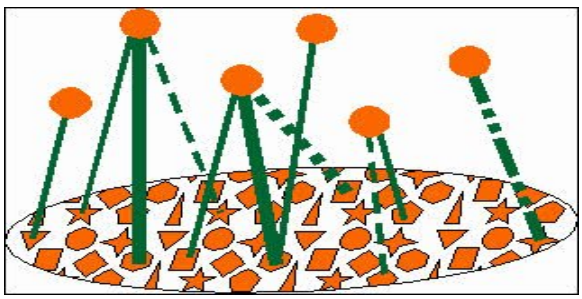

Figure 5. Occurrence roles

The distinction between an occurrence role and its type is subtle but important. In general terms they are both "about" the same thing, namely the way in which the occurrence contributes information to the subject in question (e.g. through being a portrait, an example or a definition). However, the role (indicated by the **role** attribute) is simply a mnemonic; the type (indicated by the **type** attribute), on the other hand, is a reference to a topic in the map which further characterizes the relevance of the role. In general it makes sense to specify the type of the occurrence role, since then the power of topic maps can be used to convey more information about the role.

A is for Association

Up to now, all the constructs that have been discussed have had to do with topics as the basic organizing principle for information. The concepts of "topic", "topic type", "name", "occurrence" and "occurrence

role" allow us to organize our information resources according to topic, and to create simple indexes, but not much more. [4]

The really interesting thing, however, is to be able to describe **relationships** between topics, and for this the topic map standard provides a construct called the **topic association**.

Associations

A topic association is (formally) a link element that asserts a relationship between two or more topics. Examples might be as follows:

"Tosca was **written by** Puccini"

"Tosca **takes place in** Rome"

"Puccini was **born in** Lucca"

"Lucca **is in** Italy"

"Puccini was **influenced by** Verdi"

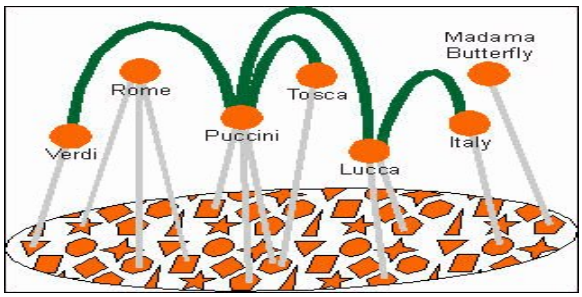

Figure 6. Topic associations

Association types

Just as topics can be grouped according to type (composer, opera, country, etc.) and occurrences according to role (mention, article, commentary, etc.), so too can associations between topics be grouped according to their type. The **association type** for the relationships mentioned above are `written_by`, `takes_place_in`, `born_in`, `is_in` (or geographical containment), and `influenced_by`. As with most other constructs in the topic map standard, association types are themselves defined in terms of topics.

The ability to do typing of topic associations greatly increases the expressive power of the topic map, making it possible to group together the set of topics that have the same relationship to any given topic. This is of great importance in providing intuitive and user-friendly interfaces for navigating large pools of information.

It should be noted that topic types are regarded as a special (i.e. syntactically privileged) kind of association type; the semantics of a topic having a type (for example, of Tosca being an opera) could equally well be expressed through an association (of type "type-instance") between the topic "opera" and the topic "Tosca". The reason for having a special construct for this kind of association is the same as the reason for having special constructs for certain kinds of names (indeed, for having a special construct for names at all): The semantics are so general and universal that it is useful to standardize them in order to maximize interoperability between systems that support topic maps.

[4] *The principle exception to this statement is the topic type, as we shall see shortly.*

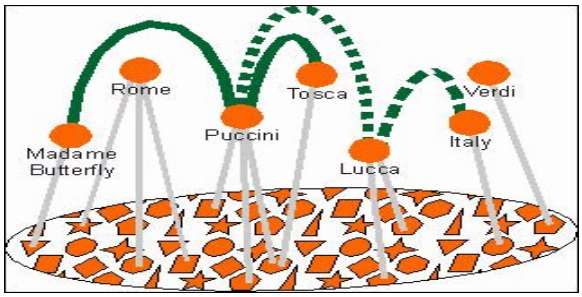

Figure 7. Association types

It is also important to note that while both topic associations and normal cross references are hyperlinks, they are very different creatures: In a cross reference, the anchors (or end points) of the hyperlink occur **within the information resources** (although the link itself might be outside them); with topic associations, we are talking about links (between topics) that are **completely independent** of whatever information resources may or may not exist or be considered as occurrences of those topics.

Why is this important?

Because it means that topic maps are information assets in their own right, irrespective of whether they are actually connected to any information resources or not. The knowledge that Rome is in Italy, that **Tosca** was written by Puccini and is set in Rome, etc. etc. is useful and valuable, whether or not we have information resources that actually pertain to any of these topics.

Also, because of the separation between the information resources and the topic map, the same topic map can be overlaid on different pools of information, just as different topic maps can be overlaid on the same pool of information to provide different "views" to different users. Furthermore, this separation provides the potential to be able to interchange topic maps among publishers and to merge one or more topic maps. [5]

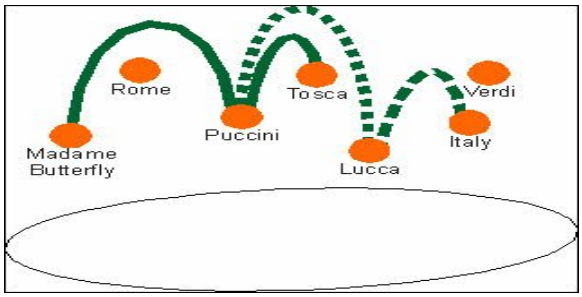

Figure 8. Topic maps as portable semantic networks

Association roles

Each topic that participates in an association plays a **role** in that association called the **association role**. In the case of the relationship "Puccini was born in Lucca", expressed by the association between Puccini and Lucca, those roles might be "person" and "place"; for "**Tosca** was composed by Puccini" they might be "opera" and "composer". It will come as no surprise now to learn that the type of an association role (indicated by the **type** attribute) is also a topic!

[5] *However, in order to be able to merge topic maps successfully, the additional concepts of **scope** and **public subject** are required. These are discussed below.*

Unlike relations in mathematics, associations are inherently multidirectional. In topic maps it doesn't make sense to say that A is related to B but that B isn't related to A: If A is related to B, then B **must**, by definition, be related to A. Given this fact, the notion of association roles assumes even greater importance. It is not enough to know that Puccini and Verdi participate in an "influenced-by" association; we need to know who was influenced by whom, i.e. who played the role of "influencer" and who played the role of "influencee".

This is another way of warning against believing that the names assigned to association types (such as "was influenced by") imply any kind of directionality. **They do not!** This particular association type could equally well (under the appropriate circumstances) be characterized by the name "influenced" (as in "Verdi influenced Puccini"). (See section "Scope" for an example (involving **Tosca** and Rome) of how the scope feature might be used to make this happen in practice.)

The IFS of topic maps

Identity (and public subjects)

Sometimes the same subject is represented by more than one topic link. This can be the case when two topic maps are merged. In such a situation it is necessary to have some way of establishing the identity between seemingly disparate topics. For example, if reference works publishers from Norway, France and Germany were to merge their topic maps, there would be a need to be able to assert that the topics "Italia", "l'Italie" and "Italien" all refer to the same subject.

The concept that enables this is that of **public subject**, and the mechanism used is an attribute (the **identity** attribute) on the topic element. This attribute addresses a resource which identifies the subject in question as unambiguously as possible. That resource could be some official, publicly available document (for example, the ISO standard that defines 2- and 3-letter country codes), or it could simply be a definitional description within (or outside) one of the topic maps.

Any two topics that reference the same subject by means of their identity attributes are considered to be semantically equivalent to a single topic that has the union of the characteristics (the names, occurrences and associations) of both topics. In the topic map grove, a single topic node results from combining the characteristics of the two topics. [6]

Public subjects are a necessary precondition for the widespread use of portable topic maps, since there is no point in offering a topic map to others if it is not guaranteed to "match up" with relevant occurrences in the receiver's pool of information resources. Activities are therefore underway, under the aegis of ISO, OASIS and the GCA, to develop directories of public subjects.

Facets

Sometimes it is convenient to be able to assign metadata to the information resources that constitute the occurrences of a topic **from within the topic map**. To provide this capability, the standard includes the concept of the **facet**.

Facets basically provide a mechanism for assigning property-value pairs to information resources. A facet is simply a property; its values are called **facet values**. Facets are typically used for supplying the kind of metadata that might otherwise have been provided by SGML or XML attributes, or by a document

[6] *Of course, the fact that the identity attributes of two topics are not identical is not sufficient to prove that the topics do not refer to the same subject; the only thing that can be proven is that there **is** identity, not that there **is not** identity.*

management system. This could include properties such as “language”, “security”, “applicability”, “user level”, “online/offline”, etc.

Once such properties have been assigned, they can be used to create query filters producing restricted subsets of resources, for example those whose language is “Italian” and user level is “secondary school student”.

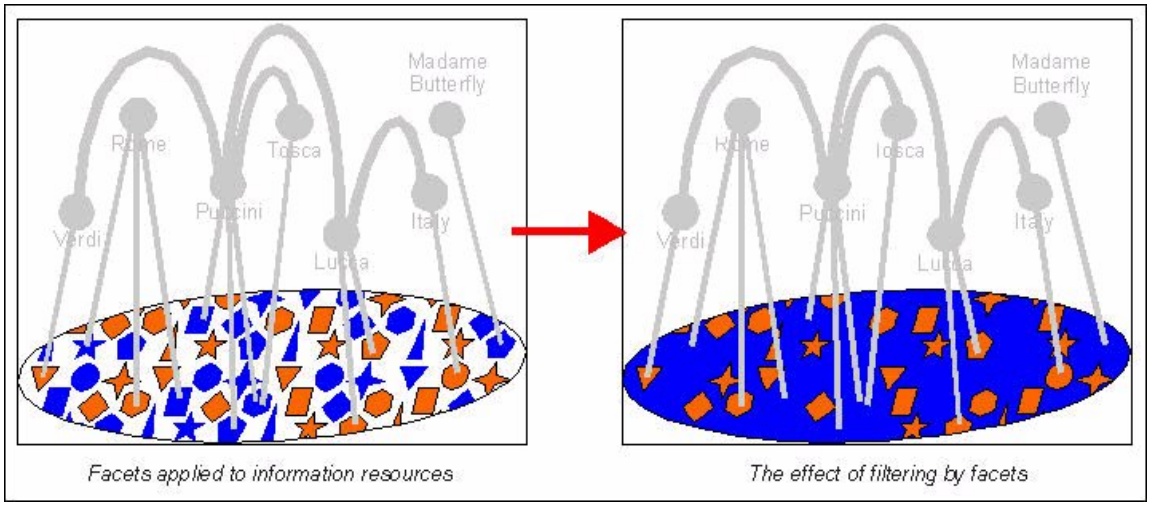

Figure 9. Applying facets for filtering

It is important not to confuse facets with scope (about which more in the next section). Facets are generally speaking **not** used to qualify the objects in the “topic domain” part of the topic map (i.e. the topics, topic names and associations). Their purpose is simply to add attributes to information resources. In a sense, facets are orthogonal to the topic map model itself (except to the extent that both facet types and facet value types, like most other things in the topic map standard, are regarded as topics). Despite this, they provide a useful mechanism that complements and significantly extends the power of topic maps.

The distinction between a facet value name and its type, like that between an occurrence role and its type, is subtle but important. In one way they are both “about” the same thing, namely the value of a property exhibited by an information resource. However, the facet value name (indicated by the **facetval** attribute) is simply a token; the type (indicated by the **type** attribute), on the other hand, is a reference to a topic in the map which further characterizes the relevance of the value. As with occurrence role types, it generally makes sense to specify the type of the facet value, since then the power of topic maps can be used to convey more information about it. One situation in which it often will not make sense to instantiate a topic in order to provide a value for a facet is when the value is an integer, date, or some such about which there is no more to be said. [7]

Scope

The topic map model allows three things to be said about any particular topic: What names it has, what associations it partakes in, and what its occurrences are. These three kinds of assertions are known collectively as **topic characteristics**.

[7] *Unfortunately, the **facetval** attribute, as defined in the original version of the standard, has the declared value NAME, which prevents the use of numeric and string values. This is regarded as a bug which will be fixed in the near future.*

Assignments of topic characteristics are always made within a specific context, which may or may not be explicit. For example, if I (yet again) mention “tosca”, I should expect my readers to think of the opera by Puccini (or its principle character), because of the context that has been set by the examples used so far in this paper. For an audience of bakers, however, the name “tosca” has quite other and sweeter connotations: it denotes another topic altogether.

Although we seldom notice it in everyday life, the problem of context is with us all the time. According to a sentence is derived from six different kinds of information, four of which (**tense** and **modality**; **presupposition**; **focus**; and **emotional connotations**) are in way or another related to context.

Humans are remarkably good at dealing with context. It is that ability that enables them to make sense of two such similar statements as **John Smith to marry Mary Jones** on the one hand, and **Retired priest to marry Bruce Springsteen** on the other, or to parse and interpret the two sentences **Time flies like an arrow** and **Fruit flies like an apple**. [8]

Computers, however, are not yet that smart. Given two such simple statements as **Tosca takes place in Rome** and **Tosca kills Scarpia**, most of today’s computers would not be able to infer which of the topics named “Tosca” was involved. In order to avoid this kind of problem, topic maps consider any assignment of a characteristic to a topic, be it a name, an occurrence or a role, 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**.

Scope is defined in terms of **themes**, and a theme is defined as “a member of the set of topics used to specify a scope”. In other words, a theme is a topic that is used to limit the validity of a set of assignments. Thus, the name “tosca” might be assigned to three different topics in scopes defined by the themes “opera”, “opera”+“character”, and “baking” respectively, thereby removing any ambiguity and reducing the chance of errors, for example when merging topic maps.

In fact, the well-designed, consistent and imaginative use of scope in topic maps does much more than simply remove ambiguity. It can also aid navigation, for example by dynamically altering the view on a topic map based on the user profile and the way in which the map is used. For example, any user that declares a specific interest in opera (or a specific lack of interest in baking!) can have the various toscas ranked accordingly.

Similarly, anything that is known about the general background of the user might be regarded as presuppositions that can affect the behaviour of the map. For example, in a topic map devoted to presenting the tourist attractions of a country, scope might be used to qualify topics such that different views on the information were presented to prospective visitors and professional tour operators.

Scope can also be used to dynamically determine which name to use for a topic based on how the topic was arrived at. For example, the association between **Tosca** (the opera) and Rome might be labelled “takes place in” in the scope of the association role type “action” (the opera) and “setting for” in the scope of the association role type “location” (the city).

[8] *A short explanation for non-English native speakers:*

In the first example, the verb “to marry” is used once in its normal sense (to get married to), and once is the sense of “to perform the marriage ceremony for”.

The multiple ambiguity is the second example can best be explained using two topic associations:

```
[time]-----(flies like)---[an arrow]
[fruit flies]---(like)----[an apple]
```

The ambiguity is caused, as can be seen, by the different roles played by the words “flies” (verb, noun) and “like” (adverb, verb).

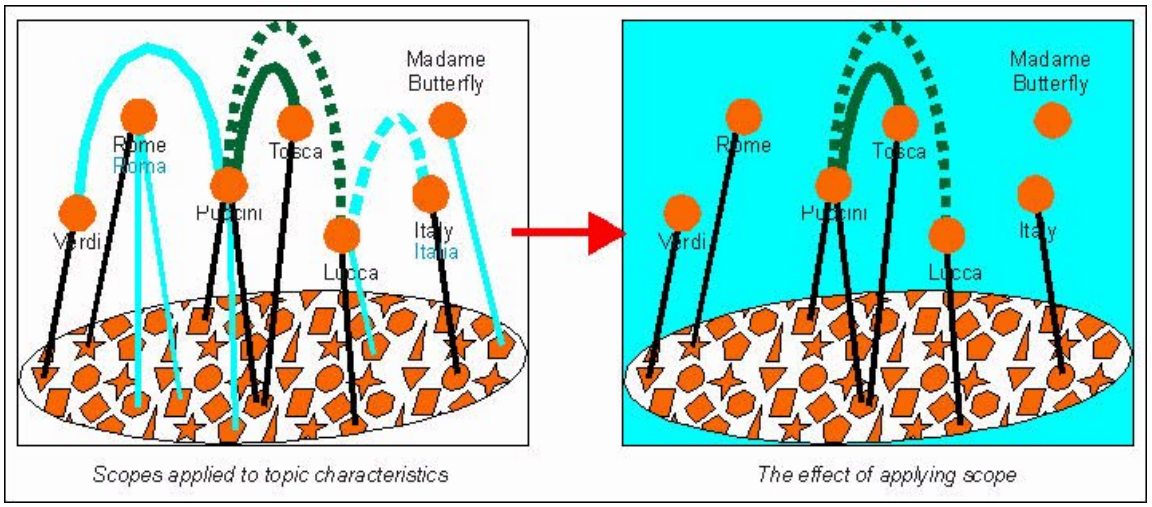

Figure 10. Scoping topic names, occurrences and associations

As mentioned above, scope should not be confused with facets. The two mechanisms are different and complementary. Whereas scope can be seen as a filtering mechanism that is based on **properties of the topics**, facets provide for filtering based on **properties of the information resources themselves**.

The BUTS of topic maps

Topic maps have no buts, but... well, there are a couple more things to be said and this seemed like as good a title as any.

It is sometimes claimed that "everything in a topic map is a topic". This is almost true, but not quite. Specifically, all types (topic types, association types, occurrence role types, facet types and facet value types) are defined as topics. In addition, scope is defined in terms of themes which are themselves topics. [9]

This design gives tremendous power to the model, allowing among other things for the topic map to be self-documenting. Since the ontology of a topic map (the kinds of things it consists of) is defined in terms of topics in the same map, the map can be used to described its own ontology and provide more functionality and flexibility when used for navigation or querying.

It also turns out that because of the power of the topic map model, topic maps can also be used to define the control information used for much topic map processing. The committee that developed topic maps has already coined the term "topic map template" for the declarative part of a topic map, and this is itself a topic map. Current research also indicates that queries on topic maps, schemas for constraining classes of topic maps, and user profiles for interacting with topic maps all can be expressed as topic maps. Finally, interesting work is being done on the use of topic maps for even more esoteric purposes, from the standardized representation of other graph structures, such as social networks, to the management of multiple schema languages.

Conclusion

[9] *Some people believe it would be useful if **absolutely everything** in a topic map were a topic, including associations, association roles, and even the topic map itself, but we won't go into that here.*

Topic maps started life as a way of representing the knowledge structures inherent in traditional back of book indexes, in order to solve the information management problems involved in creating, maintaining and processing indexes for complex documentation. As the model evolved, their scope was broadened to encompass other kinds of navigational aid, such as glossaries, thesauri and cross references.

One of the ground-breaking aspects of topic maps, made possible by the use of the HyTime standard, was the use of independent (or out-of-line) linking and addressing mechanisms. This frees the index from the resource it indexes and made it possible to create indexes for resources to which the indexer does not have write access.

However, instead of simply replicating the features of a printed index, the topic map model **generalizes** them, extending them in many directions at once and thereby enabling navigation in hitherto undreamt of ways. With topic maps a user can wander at leisure through a multidimensional topic space of knowledge before deciding which information resources are relevant, instead of wading through volumes or megabytes of data in order to find what he or she is looking for. Similarly, queries based on topic maps can be much more accurate than simple full text searching. From being a useful but often underused adjunct to the main body of information, indexes (when based on topic maps) look set to become the **sine qua non** of information delivery and consumption.

The generality and expressive power of the topic map model bring with it other advantages that go far beyond those traditionally associated with indexes. The close similarity to semantic nets gives an idea of how topic maps, even without any occurrences connecting them to an information pool, can become valuable resources in their own right. This in turn opens up new business opportunities for creating and selling "portable topic maps" that can be overlaid on multiple information pools. For traditional commercial publishers, producing well-crafted topic maps could be a new way of leveraging their existing knowledge and experience and combating the threat to their existence posed by the vast amounts of information now available for free.

The ability to encode arbitrarily complex knowledge structures and link them to information assets indicates a major role for topic maps in the realm of knowledge management: Topic maps can be used to represent the interrelation of roles, products, procedures, etc. that constitute corporate memory, and link them to the corresponding documentation.

The list seems to be endless and we can expect new usage scenarios to continue to turn up as new tools are developed, new projects conceived and new insights gained. And if a book without an index is like a country without a map, then perhaps one day a world without topic maps will seem like a head without a brain.

Acknowledgments

Thanks to my former colleagues (now collaborators) in the STEP topic map conspiracy, in particular Rafal Ksiezyk, Graham Moore and Hans Holger Rath, for many inspiring discussions; to Geir Ove Gr⊘nmo for developing all manner of topic map software for me to play with; to the editors of the topic map standard, Steve Newcomb, Martin Bryan and (in particular) Michel Biezunski, and all the others involved in its development, for doing such an excellent job; and to Sylvia Schwab for her advice and encouragement.

For examples, demos and more information about topic maps, see the website www.topicmaps.com.

Bibliography

[Bonura 94] Bonura, L. S.: **The Art of Indexing** (John Wiley, New York 1994)

[DocBook 99] Walsh, N. and Muellner, L.: **DocBook: The Definitive Guide** (O'Reilly, Sebastopol 1999)

[Goldfarb 90] Goldfarb, Charles F.: **The SGML Handbook** (OUP, Oxford 1990)

[ISO 13250] International Organization for Standardization, **ISO/IEC 13250, Information technology – SGML Applications – Topic Maps** (ISO, Geneva 2000)

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

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

[Pepper 99] Pepper, S.: "Navigating Haystacks, Discovering Needles", **Markup Languages: Theory and Practice**, Vol. 1 No. 4 (MIT Press, 1999)

[Pogue 97] Pogue, D. and Speck, S.: **Opera for Dummies** (IDG Books, Chicago 1997)

[Ruggles 97] Ruggles, Rudy L., ed. **Knowledge management tools** (Butterworth-Heinemann, Boston 1997)

[Sowa 84] Sowa, J.: **Conceptual Structures** (Addison-Wesley, Reading 1984)

[Sowa 2000] Sowa, J.: **Knowledge Representation: Logical, Philosophical and Computational Foundations** (Brooks-Cole, Pacific Grove 2000)

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

Author

Steve Pepper

Senior Information Architect
Infostream
STEP Infotek
Postal Address:
Gjerdrums vei 12
0486 Oslo
Norway
Telephon: +47 22021680
Fax: +47 22021681
E-mail: pepper@infotek.no
Web: www.infotek.no

Steve Pepper — Steve Pepper is the Senior Information Architect at STEP Infotek, a division of Infostream specializing in standards-based information reengineering. He represents Norway on JTC 1/SC 34, the ISO committee responsible for the development of SGML and related standards, and is convenor of WG 3 (Information association), whose responsibilities include the HyTime and Topic Map standards. A frequent speaker at SGML and XML events around the world, he is the author and maintainer of the “Whirlwind Guide to SGML and XML tools” and co-author (with Charles Goldfarb and Chet Ensign) of the “SGML Buyer’s Guide” (Prentice-Hall, 1998).