Article

Navigating haystacks and discovering needles

Introducing the new topic map standard

Steve Pepper
Senior Information Architect
STEP Infotek A.S
Gjerdrums vei 12
N-0486 Oslo
Norway

TEL +47 22 02 16 87
FAX +47 22 02 16 81
EMAIL pepper@infotek.no
WEB http://www.infotek.no

This article provides an introduction to the new topic map standard (ISO/IEC 13250) with particular reference to the domain of encyclopaedia publishing, and discusses the relationship between topic maps and the W3C recommendation Resource Description Framework (RDF). It is based on the author's participation in the development of the topic map standard (representing Norway in SC34, the ISO committee responsible for SGML and related standards), and two years' collaboration with leading reference works publishers in Norway, Denmark, Sweden, Poland and Germany.

A *needle in a haystack* is, according to the Concise Oxford Dictionary, "something almost impossible to find because it is concealed by so many similar things"; *to search for a needle in a haystack* is defined in another version of the same dictionary as "to attempt a hopeless task". It is a metaphor that is more than apposite to the task of information retrieval in the age of infoglut, as anyone who uses web search engines regularly knows only too well. As the amount of information available on the WWW and elsewhere continues to grow at an almost exponential rate, it becomes increasingly difficult to locate the particular piece of information we need: precious time and resources are consumed navigating haystacks of information and those sought-after needles of information become ever more difficult to discover.

Two recent standards are designed to provide ways of coping with this problem: *ISO/IEC 13250, Information technology – SGML Applications – Topic maps* [ISO, *Topic Maps*] and the *Resource Description Framework (RDF)* [W3C, *RDF model and syntax*], [W3C, *RDF schema*]. This article aims to provide a simple introduction to the basic concepts underlying the first of these, the topic

Markup Languages: Theory & Practice 1.4 (1999): 41–68

map standard, and to discuss the relationship between topic maps and RDF. In the first section we introduce the topic map standard itself and describe its background, rationale and current status. The second section presents the topic map model along with its key concepts. This is followed by a discussion of some areas of applicability of the topic paradigm to the domain of encyclopaedia publishing. Finally we give a brief overview of RDF and discuss its relationship with topic maps.

Introducing topic maps

Current status

Topic maps (formerly known as "topic navigation maps") are the subject of a new international standard (ISO/IEC 13250) developed by what is now Working Group 3 of Subcommittee 34 (the "SGML committee") of ISO/IEC's Joint Technical Committee (JTC 1).

At the time of writing (September 1999), this standard is undergoing final balloting and is expected to be published early in the year 2000. The full text of the final committee draft is freely available at `http://www.ornl.gov/sgml/sc34/document/0058.htm`.

Background

The topic map standard has had a long and convoluted history. Its genesis, almost 10 years ago, is described by Steve Newcomb, one of the prime movers, at the time co-editor of the (then) soon-to-be-published HyTime standard and now a co-editor of the Topic Map standard itself:

> At ACM Hypertext '91 in San Antonio, the emerging "Davenport" group met to decide how to go about the development of a standard for software documentation. HyTime was being considered. I agreed to participate, and for the first few meetings I served as convenor. A primary contributor was O'Reilly & Associates, whose X-Windows documentation was being shared among several computer vendors.
>
> My personal technical views (dyed-in-the-wool HyTime bigot that I am) were ultimately regarded as "futuristic", and the group split into two groups, one of which went on to develop DocBook, while the other became Conventions for the Application of HyTime (CApH) under GCARI (GCA Research Institute). I continued to convene CApH and serve as its editor.
>
> Just before the split, Fred Dalrymple (who was then in charge of documentation at the Open Software Foundation), Michel Biezunski and I were thinking about the problem of how to merge indexes. Digital Equipment Corporation wanted to merge the index of O'Reilly's X-

> Windows documentation with all the other indexes of all the other manuals that DEC would ship with its computers.
>
> That first inspiration, which occurred at OSF in Cambridge, Massachusetts, 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 modeled indexes together. But how to express such models? I made the first stab at writing it down in an early CApH draft, but the structural ideas didn't stabilize for years to come. It was always clear, from the very beginning, that hyperlinks were heavily involved. Beyond that, it was not clear. The solution, when found, should be obvious.
>
> With Fred, Michel, Wayne Wohler, and others, CApH went on to develop several ways of modeling what we called "Topic Maps". Then Michel carried the banner, almost alone, for a long time – several years, in fact. His faith in the concept never wavered, and he committed virtually all his resources to implementing and demonstrating its power. The rest of the story you know.[1]

The "rest of the story" is that, through the perseverance of Michel Biezunski, what was then called "Topic Navigation Maps" was accepted as a new work item by ISO's SGML working group in Munich in 1996. Michel was the original editor and architect; he was joined by Martin Bryan in 1997, and by Steve Newcomb the year after.

Topic maps were the subject of intense debate through 1997 and 1998 at meetings in Washington, Paris and Chicago, and on the topic map mailing list, and finally, the standard was submitted to the members of ISO for its final committee draft ballot, with a four month ballot period, in October 1998.

During this long period of gestation the model changed many times and swung back and forth from an extremely high level of generality (at one point in time the standard consisted of just two architectural forms) to much more specific models designed to be used solely for navigation.

The final result is a compromise which the working group believes offers the optimal balance (at the present point in time) between extreme power and flexibility on the one hand, and sufficiently well-defined semantics on the other. In other words, it is a standard which will allow us to do pretty much anything we can think of today, without being impossible to implement either in part or in toto.

---

[1] Private communication to the author.

Purpose

So what can the standard be used for?

As Steve Newcomb points out in the quote given above, the original interest for topic maps related to the need to be able to merge indexes. This was later extended to other forms of navigational aid: that is, to the electronic equivalents of not only printed indexes, but also tables of contents, glossaries, thesauri, cross references, etc. Common to all these applications is the attempt to provide access to information based on a model of the *knowledge* it embodies. At the heart of that model lies the concept of the *topic*.

Today it is becoming apparent that the topic paradigm can have even broader applicability. Not only can it serve as the basis for more effective navigation; in many information management contexts it can constitute the *fundamental organizing principle* for the creation and maintenance of information. In this article we will focus on the domain of reference works publishing, but it is clear that the same approach and techniques are applicable for most branches of commercial and technical publishing, including software and hardware documentation, legal publishing, financial information and many others.

The topic map model

The topic map standard defines both an abstract data model for topic maps and an SGML-based serialization syntax.[2] In order to provide maximum flexibility, the standard interchange representation is actually defined in terms of an *SGML architecture*, or "meta document type", as specified in the HyTime standard [ISO, *HyTime*]. A topic map in its interchange form is therefore an SGML (or XML) document (or set of documents) in which different element types, derived from a base set of architectural forms, are used to represent topics, occurrences of topics, and relationships (or "associations") between topics. The key concepts, then, are:

- topic (and topic type)
- topic occurrence (and occurrence role)
- topic association (and association type)

Other concepts which extend the expressive power of the topic map model are those of:

- scope
- public subject
- facets

The next section describes each of these in turn.

[2] An XML-based serialization syntax will be defined once the W3C's recommendations for XML-based linking and addressing have been finalized.

Topics and their occurrences

First of all, what is a topic?

Topics and topic types: 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 abstraction 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:[3] 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 will often be used more loosely to denote both of these things together. Whenever there is a need to distinguish between the two, we will use the terms "topic link" and "subject".

So, in the context of an *encyclopaedia*, a topic might represent subjects such as "Spain", "Andalusia", "Granada", "La Alhambra", the poet "Federico García Lorca", or a piece of music by Manuel de Falla: that is, anything that might have an entry in the encyclopaedia – but also much else besides.

Any individual topic is an instance of zero or more **topic types**.

Thus, Spain would be a topic of type "country", Andalusia a topic of type "region", Granada and Sevilla topics of type "city", García Lorca a topic of types "poet" and "playwright", 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 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 "country", "city", "poet", etc. as topics in your topic map if

[3] This image is borrowed from Rafal Ksiezyk's paper quoted below, but it also fits our purpose rather nicely.

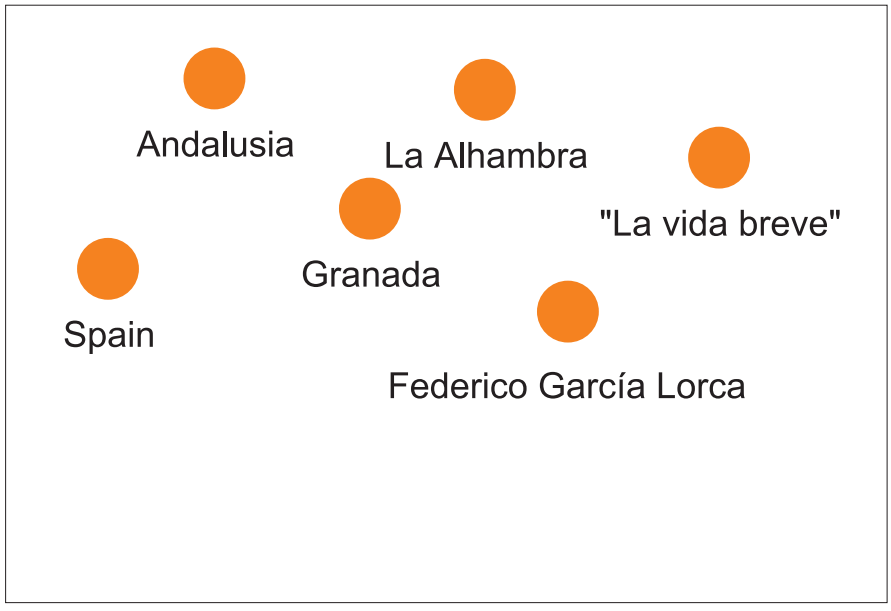

Figure 1 Topics

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).

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.[4] 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 attempt 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.

[4] 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!

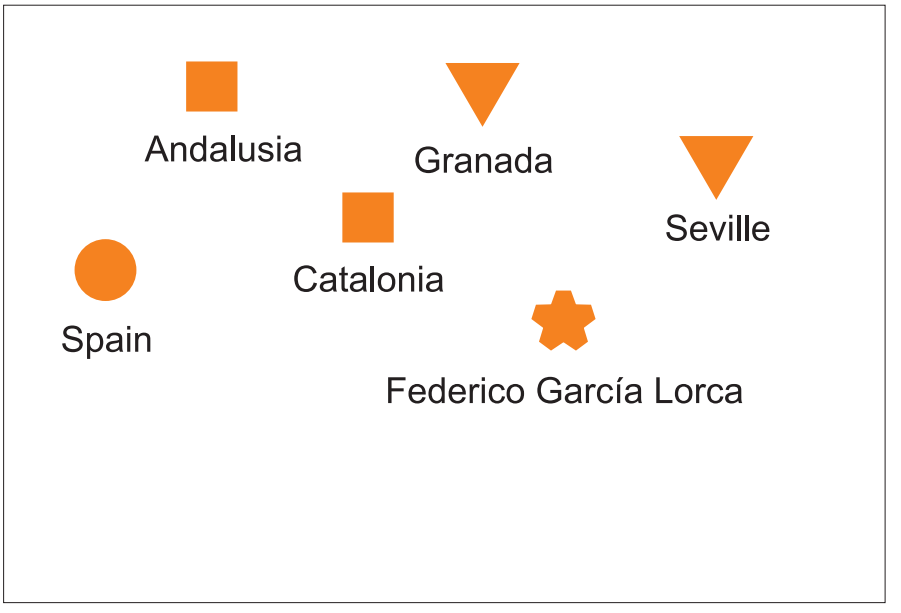

Figure 2 Topic types

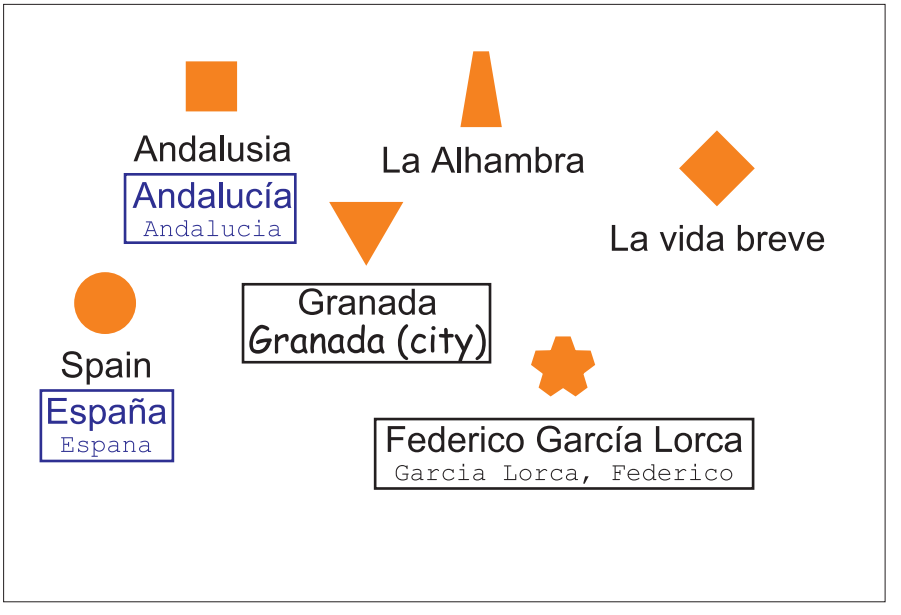

Figure 3 Topic names

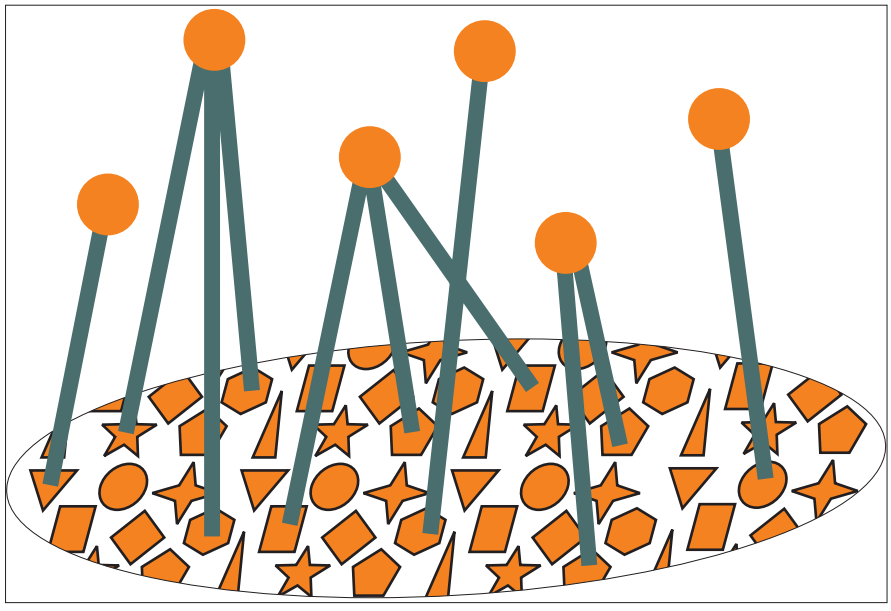

Figure 4 Occurrences

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)

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.

Occurrences and occurrence roles: 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.

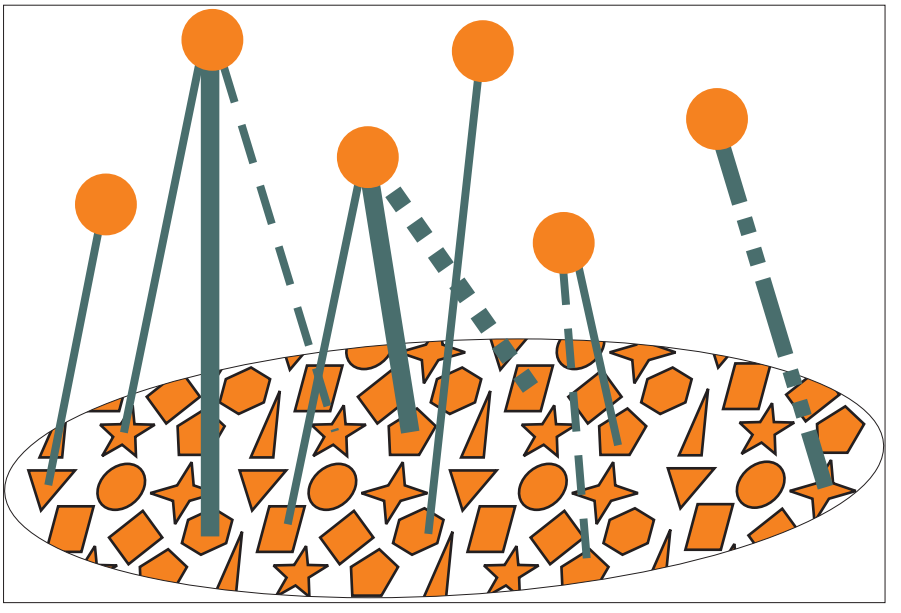

Figure 5 Occurrence roles

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.

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.

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 concept of the **occurrence role**.

As with topic types, occurrence roles are really topics, and you can therefore use the facilities of topic maps to say useful things about them (such as their names, and the relationships they partake in).

Indexes and glossaries: As described so far, topics and occurrences provide a model for explicitly stating which subjects a pool of information pertains to and how. That is basically what an index also does:

```
Andalusia          17, 77
Catalonia              72
Granada                49
Seville                22
```

But topic maps offer more. Through the concept of occurrence roles, they generalize and extend the conventions used to distinguish different kinds of references from one another:

```
Andalusia          17, 77
```

The use of different typefaces here indicates different roles played by the occurrences on pages 17 and 77 (perhaps a main description and a mention).

Some books contain more than one index (index of names, index of places, etc.). Topic types provide the same facility, but extend it in several directions to enable the creation of multiple, dynamic, user-controlled indexes organized as taxonomic hierarchies.

Glossaries can also be implemented using just the bare bones of the topic map standard that has been described so far. After all, a glossary is nothing more than a set of topic definitions, ordered by topic name:

```
España, see Spain.
...
Spain: Constitutional monarchy in southern Europe...
```

The definitions are just one particular kind of occurrence (those that play the role of "definition"). With a topic map it is easy to create and maintain much more complex glossaries than this; for example, ones that use different kinds of definitions (perhaps suited to different kinds of users).

Topic associations

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.[5]

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**.

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

- "Andalusia *is in* Spain"
- "La Alhambra *is in* Granada"
- "García Lorca was *born in* Granada"
- "*La vida breve* was *written by* Manuel de Falla"
- "Lorca *collaborated with* de Falla"

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

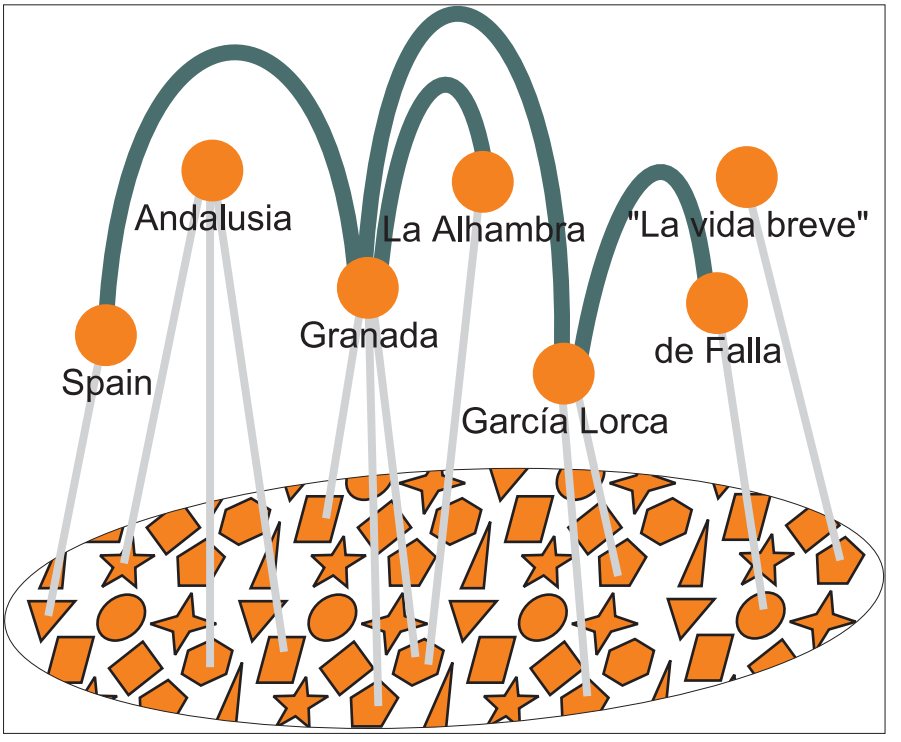

Figure 6 Topic associations

Association types: Just as topics can be grouped according to type (country, city, poet, 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**s for the relationships mentioned above are ***is-in*** (or geographical containment), ***born-in***, ***written-by***, and ***collaborated-with***. 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 Granada being a city) could quite easily be expressed through an association (of type ***instance-of***) between the topic "Granada" and the topic "city". 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

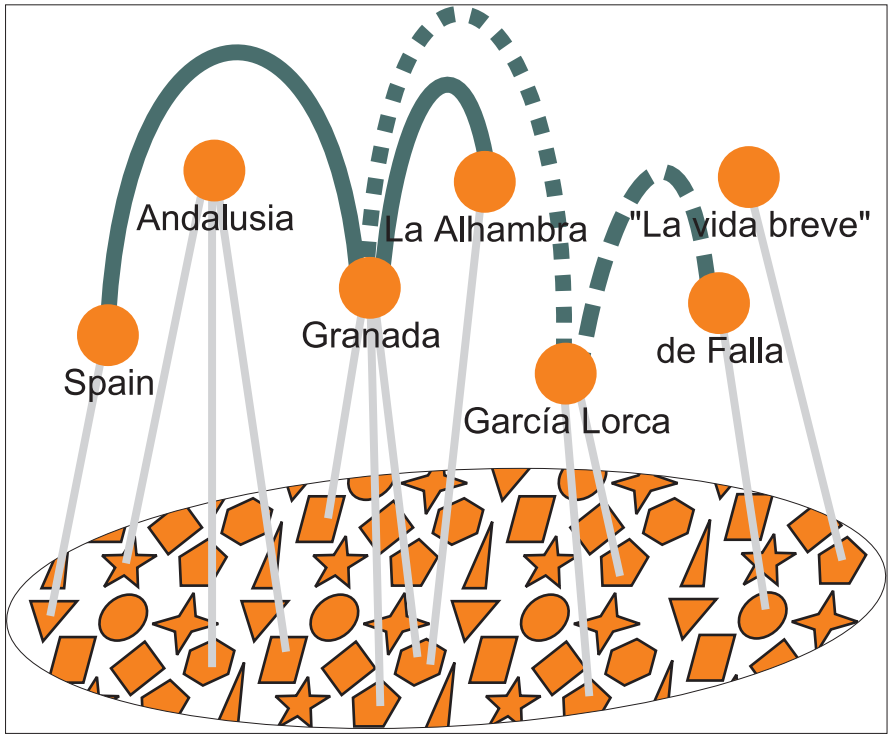

Figure 7 Association types

(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.

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 Granada is in Andalusia, that *La vida breve* was written by de Falla and is set in Granada, 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,

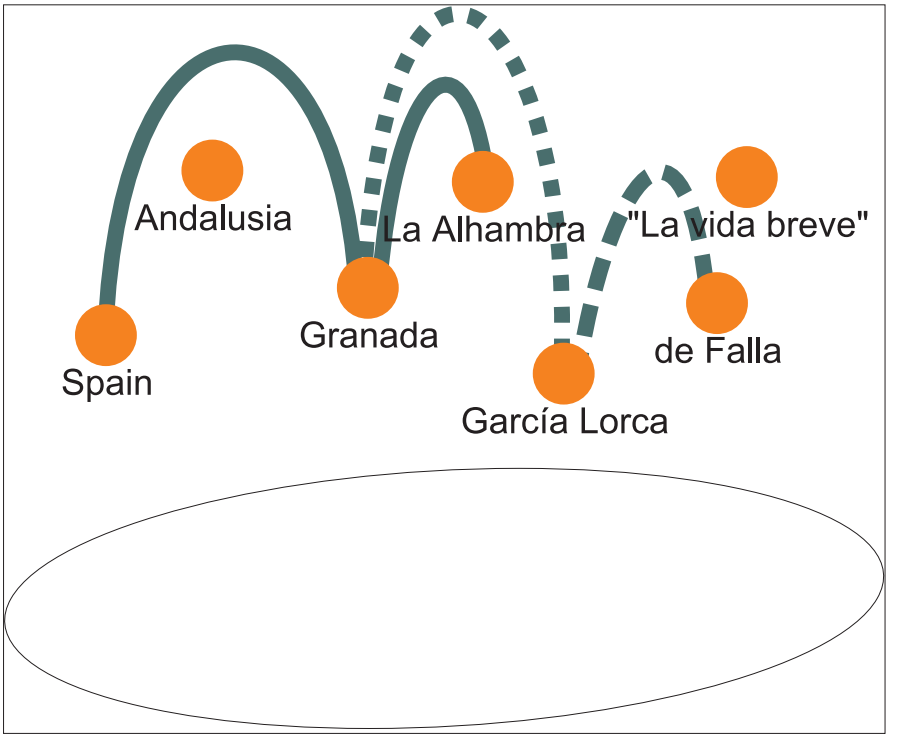

Figure 8 Topic maps as portable semantic networks

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.[6]

Association roles: Each topic that participates in an association has a corresponding **association role** which states the role played by the topic in the association. In the case of the relationship "García Lorca was born in Granada", expressed by the association between García Lorca and Granada, those roles might be "person" and "birthplace"; for "*La vida breve* was written by Manuel de Falla" they might be "opera" and "composer". It will come as no surprise now to learn that also association roles are regarded as topics in the topic map standard!

Another aspect of topic associations that is worth noting, is that they are not one-way. The ***born-in*** relationship between García Lorca and Granada implies

[6] 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.

what might be called a ***fostered*** relationship between the province and the poet (Granada ***fostered*** García Lorca), and the ***written-by*** relationship between *La vida breve* and de Falla is also a ***composed*** relationship between the composer and his opera (de Falla ***composed*** *La vida breve*).

Sometimes associations are "symmetrical", in the sense that the nature of the relationship is the same whichever way you look at it. For example, the corollary of "Lorca collaborated with de Falla" would (presumably) be that "de Falla collaborated with Lorca". Sometimes the anchor roles in such symmetrical relationships are the same (as in this case: "collaborator" and "collaborator"), sometimes they are different (as in the case of the "husband" and "wife" roles in a ***married-to*** relationship).

Other association types, such as those that express superclass/subclass and some part/whole (meronymy/holonymy) relationships,[7] are transitive: If we say that Lorca is a poet, and that a poet is a writer, we have implicitly said that Lorca is a writer. Similarly, by asserting that Granada is in Andalusia, and that Andalusia is in Spain, we have implicitly asserted that Granada is in Spain and any topic map-aware search engine should be able to draw the necessary conclusions without the need for making the assertion explicitly.[8]

Thesauri, semantic networks, and knowledge management: The addition of typed associations to the basic topic paradigm allows topic maps to model thesauri and other networks of information and knowledge.

A thesaurus is a network of interrelated terms (along with their definitions, examples, etc.) within a particular domain. There exist various standards for thesauri [ANSI, *Guidelines*], [ISO, *Guidelines 1985*], [ISO, *Guidelines 1986*] that predefine relationship types such as "broader term", "narrower term", "used for", and "related term", all of which correspond directly to association types in a topic map. Other thesaurus constructs, such as "source", "definition", and "scope note" would be modeled as occurrence roles in a topic map.

One advantage of applying the topic map model to thesauri is that it becomes possible to create hierarchies of association types that extend the thesaurus schema without deviating from accepted standards (for example, by subclassing "used for" as "synonymous for", "abbreviation for", and "acronym for"). Further advantages would be gained from using the facilities for scoping, filtering and merging described in the next section.

"Semantic networks", "associative networks" and "knowledge (or conceptual) maps" are terms used within the fields of semantics and artificial

---

[7] For a discussion of the various kinds of part/whole relationship and their properties, see [Iris/Litowitz/Evens, "Problems"].

[8] The current version of the topic map standard does not have "built in" support for expressing transitivity, but this would not prevent applications from providing such capabilities.

intelligence to describe various models for representing knowledge structures within a computer. Many of these already correspond closely to the topic/association model. Adding the topic/occurrence axis provides a means for "bridging the gap" between knowledge representation and the field of information management.

"Knowledge management" is 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",[9] 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.

Additional concepts

Scope: From the preceding discussion we see that topics can have various characteristics assigned to them: they can have *names*, they might have *occurrences*, and for every association in which they partake, they have a *role*. These different kinds of assertions that can be made about a topic are collectively known as **topic characteristics**.

In the topic map standard, any assignment of a characteristic to a topic, be it a name, an occurrence or a role, is considered to be valid within certain limits, which may or may not be specified explicitly. The limit of validity of such an assignment is called its **scope**, and scope – as you might expect – is defined in terms of topics.

For example, when I refer to "Granada", it is clear that I am referring to the city in Spain. Or is it? How can someone know that I am not talking about the town of the same name in Nicaragua, or the song by Agustín Lara that Carreras sang in the first Three Tenors concert? Presumably because of the context set by my use of examples so far in this paper.

With topic maps, there is machinery for specifying that kind of scope explicitly, and also for handling situations (for example, when merging topic maps) in which the use of implicit scoping might otherwise lead to errors or ambiguities.

---

[9] [Ruggles, *Knowledge management tools*]

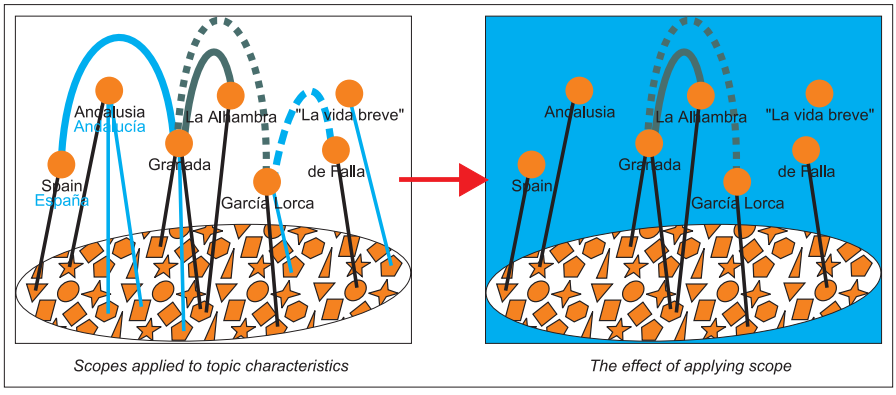

Figure 9 Scoping topic names, occurrences and associations

One part of this machinery, is the concept of the **theme**, which 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. So, in a topic map where the scope was set in terms of the themes "Spain" and "popular music", the name "Granada" could be unambiguously used to denote the song referred to above.

Public subject: 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, Poland and Germany were to merge their topic maps, there would be a need to be able to assert that the topics "Spania", "Hiszpania" and "Spanien" 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.[10]

[10] 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.

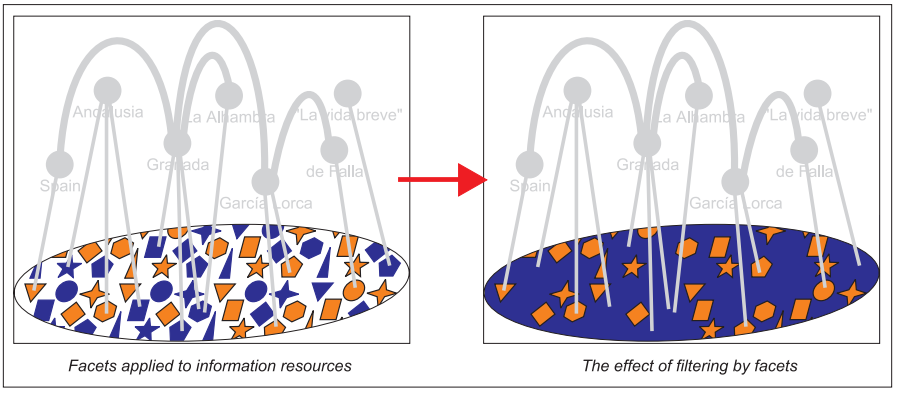

Figure 10 Applying facets for filtering

Facets: The final feature of the topic map standard to be considered in this introduction is 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. This could include properties such as "language", "security", "applicability", "user profile", etc. Facets could also cover the kinds of properties used in faceted classification systems (hence the name); for example, typical facets within the domain of medicine might be "disease", "therapy" and "age group".[11]

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 "Spanish" and user profile is "secondary school student". This provides a complement to scoping; whereas the latter 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*.

In a sense, facets are orthogonal to the topic map model itself (except to the extent that both facets and facet values, like most other things in the topic map standard, are regarded as topics). Despite this, facets provide a useful mechanism that complements and significantly extends the power of topic maps.

Classes of topic maps: Earlier in this paper it was mentioned that the interchange syntax for topic maps is defined as an SGML architecture according to the HyTime standard. The meta-DTD given in Annex A provides declarations for element types such as *topic*, *occurs*, *assoc*, etc. that represent the concepts

[11] For a short description of faceted classification see the article by Bob Streich in an earlier issue of this journal [Streich, "Techniques"]. For more detailed expositions, see [Ranganathan, *Prolegomena*], [Vickery, *Faceted classification*] and [Vickery, *Faceted classification schemes*].

described above. It can be used "as is", or it can be used as the base DTD from which application-specific DTDs are derived. For example, it would be possible to create such a DTD that declared element types for *term*, *definition*, *use-for*, *hypernym*, *broader-term*, *narrower-term*, etc. and thus could be used to mark up topic maps that represent standard thesauri.

Another way of creating classes of topic maps is through the concept of *topic map templates*, which is described in a paper to be presented at the GCA's Markup Technologies '99 conference [Rath/Pepper, "Introduction and Allegro"]. Essentially, a topic map template is a set of topics that is used to declare base constructs that are reused across multiple topic maps. This would comprise all those topics that are used as themes and as types for

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

To return to our examples from the domain of encyclopaedia publishing, such a template might contain topics for topic types such as *country*, *region*, *city*, *poet* and *playwright*; for occurrence roles such as *article*, *illustration* and *mention*; and for association types such as *is-in*, *born-in*, *written-by* and *collaborated-with*.

This leads us back to the question of how topic maps can be applied in a specific domain such as reference works publishing, which is the subject of the next section.

Topic maps and reference works publishing

In the age of digital information all commercial publishers are facing major new challenges, but perhaps none more so than publishers of reference works, especially encyclopaedias and dictionaries. Not only has the advent of the World Wide Web finally forced all of them – even the laggards – to think seriously about moving into electronic publishing; it has also turned out to be perhaps their biggest and most threatening competitor.

The reason for this, of course, is that the raw material from which reference works are fashioned consists for the most part of "hard facts" that cannot be owned. The knowledge that Lorca was born in Granada, that the population of Spain is about 39 millions, or that the Alhambra was built by the Moors is not copyrightable. You cannot take out a patent on the information that de Falla wrote *La vida breve*! Almost every piece of information to be found in any modern, commercial encyclopaedia can be found somewhere on the Internet for free, so how is a reference works publisher to compete?

Once again the answer lies in the fact that most users today do not need *more* information – if anything, they need *less*, because they are already drowning in enormous quantities of it. At the very least, they need the ability to be able to find their way to relevant information as quickly as possible and to be able to filter out the "noise" created by all the information for which they have no use. They also need to be able to trust the information they receive, to know that it is reliable and up-to-date.

When writing this article, I wanted to know who wrote the song *Granada* in order to be able to make my point about the scope of names. So I did a search using AltaVista and eventually, after several attempts to narrow the number of hits, found the following:

> Agust'n Lara, one of Mexico's greatest songwriters, wrote popular songs about Spain and Spanish life. The Spanish tenor Plácido Domingo – who grew up and began singing in Mexico – returns the compliment in his new Sony Classical Recording of Lara's songs entitled *Under the Spanish Sky (Bajo el Cielo Español)*.
>
> Best known for "Granada", a song Plácido Domingo has recorded before and performed on the first of the "Three Tenors" concerts, Lara was so prolific and successful as a songwriter that his name is synonymous with the popular song in Mexico, yet many of his songs describe Spain, among them the 12 songs of his *Suite Española* that Domingo performs on *Under the Spanish Sky (Bajo el Cielo Español)* (SK/ST/SM 62625)
>
> `http://www.sonyclassical.com/releases/62625.htm`

Just in this short extract there were two errors that even *I* managed to spot;[12] how many others might lurk there undetected?

Thus, two of the most important "value-adds" that commercial publishers can provide are

- tools and methods for finding the required information in a timely manner; and
- the confidence that the information so found can be trusted.

Another way for publishers to meet the challenges imposed by the new age of information is the ability to be able to customize, re-purpose and reuse existing information efficiently, by providing new products at short notice based on an existing body of information assets. One prerequisite for this is that information assets are organized as a central pool of knowledge rather than as a set of unrelated "works" or "publications". Another is that redundancy is kept to a

[12] The composer's first name was "Agustín", not "Agust'n", and it was Carreras, not Domingo, who sang *Granada* in Rome!

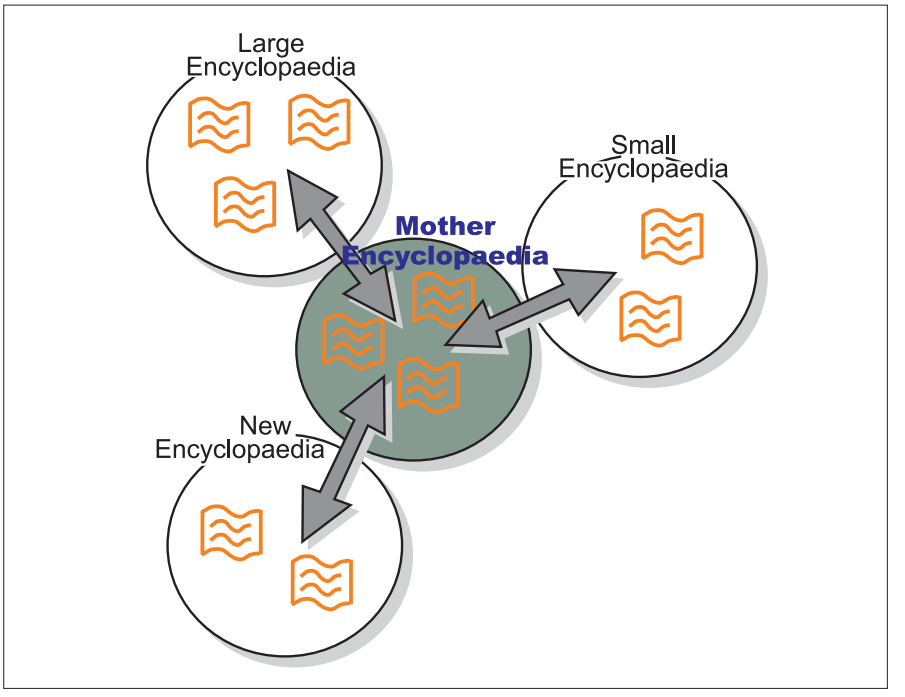

Figure 11 PWN's "Mother Encyclopaedia"

minimum. Both of these impose new requirements on the editorial system used to manage those assets.

Before looking at how topic maps can help solve these problems, here is some background drawn from the ideas of two leading European publishers of reference works, PWN and KF.

The "Mother Encyclopaedia"

Polish Scientific Publishers (or PWN) is the largest publisher of encyclopaedias in Eastern Europe. Their concept of the "Mother Encyclopaedia" was described by Rafal Ksiezyk in a paper given at XML Europe '98. The basic idea is as follows:

> The idea of ME (Mother Encyclopaedia) comes after Plato. In ME we place SGML instances of ideas of all articles which can appear or already appeared in the real encyclopaedia. The real articles are the shades on the wall of the Plato's cave cast by ideas from ME. They can differ from publication to publication but [the] original is the same. Since articles in ME have no standard body (they are pure ideas) they are linked to their children in particular publications. So children define them. [Ksiezyk, "Plato"]

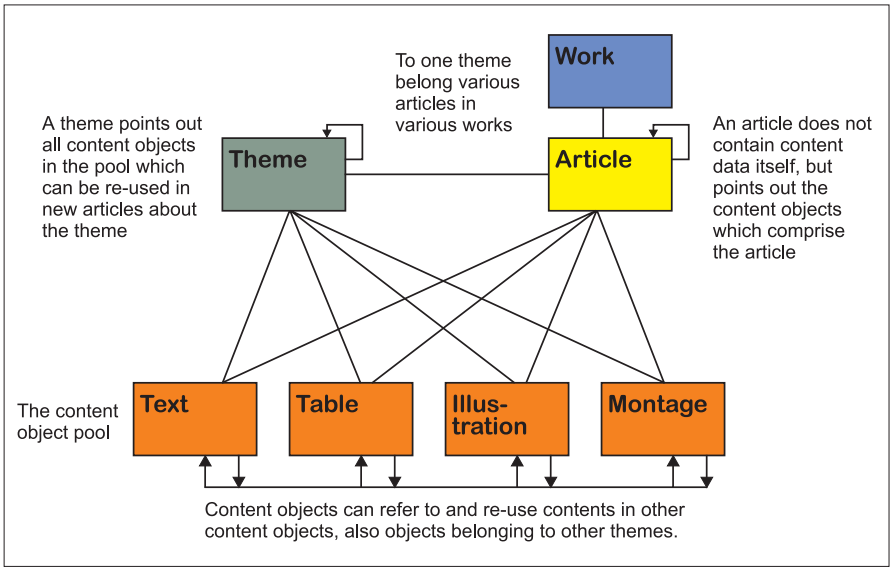

Figure 12 KF's "Theme model"

The benefits of this approach were identified as being: *identification* (i.e. establishing the connections between articles about the same subject in different works, and *ease of maintenance* (core data relating to a particular subject is stored in one place and therefore easier to update and reuse).

The "Theme" Model

Kunnskapsforlaget (or KF – the name translates as "Knowledge Publishers") is the biggest reference works publisher in Norway. They developed their own model of how their information should be organized in order to promote reuse and the construction of new works from existing resources:

> In our logical data model the central data objects are *Theme*, *Article* and *Work*, in addition to the so called *content objects: Text, Illustration, Montage, Table*. The figure shows simplified how we envisage these put together in the database.
>
> The data model's fundamental aim is to enable the use in many works of one set of contents (texts, illustrations etc.). These contents are to be kept and updated in one place, a pool adapted for re-use. Here we gather all the contents in Kunnskapsforlaget's various publications, each allotted its particular *Theme*. Through the *Theme* the editor can overview the content elements available when composing f.ex. a new article on a theme. [Henriksen, *Kunnskapsforlagets*]

Once again, the goal is to ease the task of maintenance, and enable more reuse and the faster creation of new works from the existing pool of information.

Applying topic maps

It should be clear that both the Mother Encyclopaedia and the Theme model fit very well with the topic paradigm. The base set of topics in an encyclopaedia are precisely those subjects about which PWN would have an article in the ME, or that KF would designate as a "theme". So merely organizing the information pool according to topic provides most of the immediate benefits that these publishers are looking for.

However, the story doesn't end there. Once the information is organized according to topic map principles, other benefits accrue, in areas such as navigation, semantic validation, and the reuse of "hard facts".

New navigational paths: A traditional encyclopaedia typically provides a number of different navigational mechanisms:

1. alphabetical list of headwords
2. cross references within the text
3. "see also" references
4. alphabetical index of keywords

As we have seen, topic maps make the creation and maintenance of traditional indexes much easier. In the same way, cross references that point to their target indirectly, via its topic, provide more flexibility than those that point directly to an information resource. Finally, "see also" references are infinitely more powerful when based on topic maps, since the related topics that they point to can be grouped, prioritized and presented based on the nature of their association with the current topic (perhaps related to the path by which the current topic was arrived at), the occurrence roles of the resources in question, topic types, the user's context (or scope), etc.

In addition (and this is perhaps the most significant advantage), these powerful indexing tools can be maintained independently from the set of information resources that make up any particular publication: The maps constitute portable semantic networks that can be re-used across products.

"Hard facts": Any article in an encyclopaedia is said to be "about" a certain topic. Associated with that topic there will very often be some hard facts, and the articles on that topic will consist of a mixture of such facts and other, more subjective material that is more or less influenced by the opinions of the editor and the goals of the publisher. The hard facts are repeated many times, in different works and on different media; the other material is usually unique to a particular work or a small number of works.

Examples of what we are calling "hard facts" are the following:

- Keywords and associated information: sort keys, pronunciation, etymology, parts of speech, etc.
- Parts of the article header: alternative forms of the keyword (pseudonyms, synonyms, etc.), subject areas, dates of birth and death, etc.
- Statistical data of various kinds: population figures, geographical data, economic data, etc.

Sometimes such facts are best stored as attributes of the topic itself (or its name); sometimes it makes more sense to store them along with similar information about other topics of the same type – not least because they will then be available for use in comparative tables and diagrams. (This applies particularly to population figures and other kinds of statistics, that also become easier to maintain when stored together in relational tables.)

The following example shows how information stored in such a way might be referenced from within an SGML or XML document by means of an SQL query defined using architectural forms:

```
<tbody
  SLquery="tptable"
  tptype="country"
  tpnames="DK DE NO PL"
  tpprops="name area population gdp"
  key="ISO3166_A2"
  order="name"
>
```

The result of such a query would be a table that is automatically kept up-to-date as the statistical information in the relational table is updated:

Country	Area (km^2)	Population	GDP (USD)
Denmark	43,094	5,305,048	118.2 bill.
Germany	356,910	82,071,765	1.7 trill.
Norway	324,220	4,399,993	114.1 bill.
Poland	312,683	38,615,239	246.3 bill.

Semantic validation: Finally, the holy grail of total factual accuracy becomes a little less distant when semantic validation mechanisms are implemented, and these become immeasurably easier to maintain when they are linked to the topic paradigm.

Thus editorial guidelines might stipulate that the set of allowable values for language attributes, used when providing etymological or other language-related information, is that specified in ISO 639; that countries should have a geographical code taken from ISO 3166; etc. These guidelines can be enforced by storing the controlled vocabularies as attributes of the topic types "language" and "country" respectively and once again using architectural mechanisms to invoke the validation:

```
<!NOTATION iso639-1 PUBLIC
           "ISO 639:1988//NONSGML Two letter language codes//EN" >
<!ELEMENT term - - (#PCDATA) >
<!ATTLIST term
          language CDATA #IMPLIED
          SLvalid CDATA #FIXED "CONVOC language cvlocatt"
          cvlocatt CDATA #FIXED "iso639-1 icase"
>
```

Topic maps and RDF

RDF (Resource Description Framework) is a specification developed by the W3C in two parts: the "Model and Syntax Specification" [W3C, *RDF model and syntax*] and the "Schema Specification" [W3C, *RDF schema*]. It describes itself as "a foundation for processing metadata [providing] interoperability between applications that exchange machine-understandable information on the Web."

RDF is intended to be applied in a number of application areas, including:

> in *resource discovery* to provide better search engine capabilities, in *cataloging* for describing the content and content relationships available at a particular Web site, page, or digital library, by *intelligent software agents* to facilitate knowledge sharing and exchange, in *content rating*, in describing *collections of pages* that represent a single logical "document", for describing *intellectual property rights* of Web pages, and for expressing the *privacy policies* of a Web site.

This prompts the question: What is the relationship between RDF and topic maps, since they patently address some of the same fundamental problems of the age of infoglut? This final section attempts to provide some of the answers.

Overview of RDF

The RDF specification has three main components:

- an abstract data model
- two XML-based syntaxes for encoding instances of that model (the basic serialisation syntax and the basic abbreviated syntax)

- an XML-based schema language for describing definitions and constraints that are common to a class of RDF instances

The basic data model consists of three object types:

Resources: Everything that RDF seeks to describe are called resources. A resource may be "an entire Web page; ... part of a Web page; ... [or] a whole collection of pages." It may also be "an object that is not directly accessible via the Web; e.g. a printed book".

Properties: A property is a "specific aspect, characteristic, attribute, or relation used to describe a resource."

Statements: A statement is the assignment of a named property plus the value of that property to a specific resource.

The basic model, in other words, is one in which property-value pairs are assigned to resources. This apparent simplicity however is belied by the fact that the *value* of a property can be another resource (or collection of resources). The recursion introduced by this feature greatly increases the expressibility of the model.

The basic model, as can be seen, is very general. The role of the RDF Schema specification is to permit the creation of more specific models to suit the common requirements of communities of interest. Using RDF schema it is possible to define controlled vocabularies (and corresponding semantics) for properties and their values.

Comparison with topic maps

Topic maps and RDF clearly exhibit a number of similarities:

- at the core of both standards is a data model whose primary purpose is to facilitate in some way the annotation of information resources
- both standards thereby provide a mechanism for making it easier to find relevant information and thus alleviate the main problem of infoglut
- both standards define interchange syntaxes based on XML and/or SGML
- both standards provide mechanisms for subclassing the base model (derived DTDs and templates for topic maps; schemas for RDF)

Looking more closely it is tempting to equate RDF resources with topic occurrences, and RDF statements (the assignment of property-value pairs to resources) with the topic map concept of facets, in which case RDF could be regarded as an alternative to the facet mechanism.

However the matter is more complicated, since an RDF resource, as we have seen, can be "an object that is not directly accessible via the Web". Property values can be "structured entities", which themselves are represented as further resources. The example given in the specification is that the value of the "creator"

property of a particular resource is another resource that represents the author and that itself has properties such as a name, an email address, or an employer.

In other words, RDF resources are not restricted to information resources in the form of web pages (or parts or collections of web pages) or even information resources that are not accessible via the Web (like a book): An RDF resource can also represent entities of the kind that could be considered to be topics in a topic map, such as people, organizations, etc.

Topics could therefore be represented as RDF resources (and topic types could be expressed via the `rdf:type` property). Topic names would be user (or schema) defined properties of those resources. Furthermore, since properties express relations,[13] topic associations (i.e. relations between two or more topics) may also be represented using RDF statements in which the property values are resources that represent topics.

The link between topics and their occurrences could in turn be expressed through RDF statements assigning properties like "occurrenceOf" or "definedBy" (representing different occurrence roles) to the information resources, and giving these properties values that are resources representing topics.

From this it is clear that the RDF model is capable of subsuming large parts (if not all) of the topic map model, *but without retention of the semantics*. Or, to put it another way:

A topic map may be encoded in RDF, but an RDF processor would not be able to do anything useful with it because of the loss of semantics. There would be no way for the processor to know the difference between topics and information resources, or between occurrences, associations, and facets; neither would it be able to discern topic names, recognize scope, automatically merge topics that represented the same subject, etc.

It is the same problem that faces any generalized processor (including a HyTime engine), and is in fact the very reason why the topic map standard was originally proposed by CApH as an "application of HyTime".

Of course, none of this detracts from the value of RDF for the purpose for which it was primarily designed, namely the application of (complex) metadata to information resources on the Web. However, the fact that RDF uses a very powerful and expressive model should not tempt us to use it for purposes for which it is far from optimized. RDF is best used for what we might call "resource-centric" applications (i.e. describing information resources), whilst topic maps excel at describing knowledge structures and linking them to information pools.

---

[13] Any property assignment can also be expressed as a relation between two concepts, and vice-versa. For example, the statement "the car is black" can be expressed (in RDF) as a resource (the car) with a property (its color) whose value is black, or (in a topic map) as an association of type "has-color" between two topics (the car and the color black).

Conclusion

The new topic map standard provides a standardized way of modeling the structure of the knowledge contained in information resources in such a way as to enable new means of navigation and retrieval, and ultimately also new means of organization of that information.

The applicability of topic maps extends to all spheres of information management, not least commercial reference works, and effectively "bridges the gap" between knowledge representation and information management.

Support for topic maps is currently being implemented in a number of information management tools, including STEP's document management and editorial system, *SigmaLink*.

For more information about topic maps, join the topic map discussion list by sending email to <mb@infoloom.com>.

References

Delcambre, Lois M.L., David Maier, Radhika Reddy, Lougie Anderson, "Structured Maps: modeling explicit semantics over a universe of information" in *International Journal of Digital Librarie*s, Vol. 1, No. 1, Berlin/Heidelberg: Springer, 1997.

Henriksen, Petter, *Kunnskapsforlagets publishing system – PUS: Specification of requirement*s, Oslo: KF, 1997.

Iris, M., B. Litowitz, and M. Evens, "Problems of the part-whole relation" in Evens, M., ed. *Relational models of the lexicon*, Cambridge: Cambridge University Press, 1988.

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

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

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

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

Ksiezyk, Rafal, "Plato, SGML and Revolution" in *Proceedings of the SGML/XML Europe '98 Conference*, Alexandria: GCA, 1998.

Pepper, Steve, "Euler, Topic Maps, and Revolution" in *XML Europe '99 Conference Proceedings*, Alexandria: GCA 1999.

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

Rath, Hans Holger, and Steve Pepper, "Topic maps: Knowledge navigation aids" in Goldfarb, Charles F., and Paul Prescod, *XML Handbook*, 2nd Edition, Upper Saddle River: Prentice Hall, 1999.

Rath, Hans Holger, and Steve Pepper, "Topic maps: Introduction and Allegro" in *XML '99 Conference Proceedings*, Alexandria: GCA [forthcoming].

Brickley, Dan, and R.V. Guha, ed. *Resource Description Framework (RDF) Schema Specification*, Cambridge, Paris, Tokyo: W3C, 1999. W3C Proposed Recommendation 03 March 1999. The latest version is available at http://www.w3.org/TR/PR-rdf-schema.

Lassila, Ora, and Ralph R. Swick, ed. *Resource Description Framework (RDF) Model and Syntax Specification*, Cambridge, Paris, Tokyo: W3C, 1999. W3C Recommendation 22 February 1999. The latest version is available at http://www.w3.org/TR/REC-rdf-syntax.

Ruggles, Rudy L., ed. *Knowledge management tools*, Boston: Butterworth-Heinemann, 1997.

Streich, Robert, "Techniques for managing collections of interrelated text modules", *Markup Languages*, Vol.1:2, Cambridge: MIT Press 1999.

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

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

ANSI/NISO, *Z39.19. Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, Bethesda: ANSI/NISO, 1966.