Euler, Topic Maps, and Revolution

Steve Pepper, *Senior Information Architect*

STEP Infotek A.S
Gjerdrums vei 12
N-0486
Oslo
Norway
+47 22 02 16 87
+47 22 02 16 81
pepper@infotek.no
http://www.infotek.no

Biography:

Steve Pepper is the Senior Information Architect at STEP Infotek, a company in the STEP group, based in Norway, Germany and Hungary, that specialises in information reengineering.

Originally trained as a typographer, Steve has worked with structured information since 1988 and participated in a wide range of SGML-related projects in both the public and private sectors in Norway and abroad. He is currently heavily involved in the implementation of document management solutions for leading European reference works publishers and is a principal architect of the Reference Works Module functionality of STEP's flagship product, SigmaLink.

Steve represents Norway on JTC 1/SC 34, the ISO committee responsible for the development of SGML and related standards, where he has played an active role in the development of the new Topic Navigation Map standard (ISO/IEC 13250) and is currently acting convenor of WG3 with responsibility for HyTime, Topic Maps, ISO-HTML, ISMID (International Standard Metafile for Interactive Documents) and SMDL (Standard Music Description Language).

A frequent speaker at SGML events around the world, he is the author and maintainer of the popular *Whirlwind Guide to SGML and XML tools*, which is freely available on the internet at `http://www.infotek.no/sgmltool/`, and co-author (with Charles Goldfarb and Chet Ensign) of the *SGML Buyer's Guide*, a comprehensive guide to choosing SGML and XML products and services.

Abstract:

This paper aims to provide a short introduction to the new topic map standard, and to illustrate some potential applications of topic maps, particularly in the area of encyclopaedia publishing. 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, Poland and Germany.

The title owes its inspiration to a paper given by Rafal Ksiezyk of Polish Scientific Publishers at this conference in Paris last year (1998).[1]

---

1. Introduction to Topic Maps

1.1. Current status

Topic maps (formerly called “topic navigation maps”) are the subject of a new international standard (ISO/IEC 13250) developed by what is now Working Group 3 of Sub-Committee 34 of ISO/IEC's Joint Technical Committee (JTC 1).

At the time of writing (early March 1999), this standard has just passed its final ballot and is now being edited prior to publication in late spring 1999. During the editing process, it is expected that some minor changes will be made as the result of comments received from member bodies of ISO. Those changes may slightly affect the terminology and disposition of the standard, but not its basic underlying model.

This paper is based on a version of the text as sent out for ballot that is already partially revised and it should therefore correspond in most, if not all details with the final standard.

1.2. Background

The topic map standard has had a long and convoluted history. Its genesis, almost 10 years ago, is described here by Steve Newcomb, one of the prime movers, then co-editor of the 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 modelled 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 stabilise 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 modelling 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.[2]

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 in 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, with additional major contributions from Eliot Kimber, Peter Newcomb and Sam Hunting. 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.

1.3. Purpose

So what can the standard be used for?

As Steve Newcomb points out in the quote given above, the original motivation for topic maps related to the need to be able to merge indexes. This was later extended to other forms of navigational aid: 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 contains. At the heart of that model lies the concept of the *topic*.

Today it seems that the topic paradigm can have even broader applicability. Not only can it serve as the basis for navigating information; in many contexts it may also be seen as the *fundamental organising principle* for the creation and maintenance of information. This is very definitely the case in reference works publishing and in legal publishing, but it seems that the ideas can be equally useful in all branches of commercial and technical publishing. We will return to some practical applications of this towards the end of this paper.

Formally speaking, the standard interchange representation of topic maps is defined in terms of an *SGML architecture*. A topic map is basically an SGML (or XML) document (or set of documents) in which different element types, derived from a basic set of architectural forms, are used to represent topics, occurrences of topics, and relationships (or "associations") between topics. The key concepts, then, are the **topic** (and **topic type**), the **topic occurrence** (and **occurrence role**), and the **topic association** (and **association type**). Other concepts which extend the expressive power of the topic

map model are those of **scope**, **public subject** and **facets**.

1.4. Topics and their occurrences

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 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 an earlier draft of the standard: “The invisible heart of every topic is the subject that its author had in mind when it was created. In effect, 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 paper 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: anything that might have an entry in the encyclopaedia – but also much else besides.

Fig. 1. Topics

A topic has a **topic type** – or perhaps multiple 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 (variously called hyponymy/hypernymy, subordination/superordination, or the IS-A relation).

Just what one regards as topics and topic types will vary depending on the kind of information in question: 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 can explicitly declare “country”, “city”, “poet”, etc. as topics in your topic map if you want (in which case you will be able to say more about them using the topic map model itself); otherwise a topic map processor will tacitly interpret them as topics and instantiate them as such in the internal data structure it uses to represent the topic map (called the “topic map grove”).

Fig. 2. Topic types

What are the characteristics of a topic?

First of all, a topic can have a **name** – or more than one.

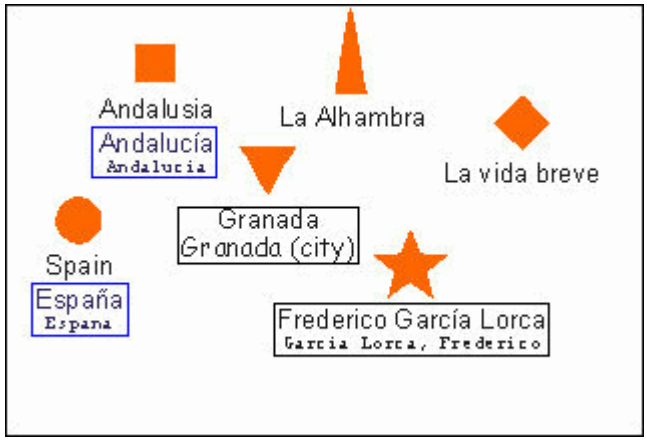
Normally topics have explicit names, since that makes them easier to talk about. (It should be clear that the preceding paragraphs would have been rather more difficult to understand if we hadn't given names to our topics and topic types!) However, topics don't *always* have names: A simple cross reference, such as “see page 97”, is considered to be a link to a topic that has no (explicit) name.

There are various kinds of names: 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 recognises the need for some forms of names that have particularly important and universally understood semantics to be defined in a standardised 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 **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)

Fig. 3. Topic names



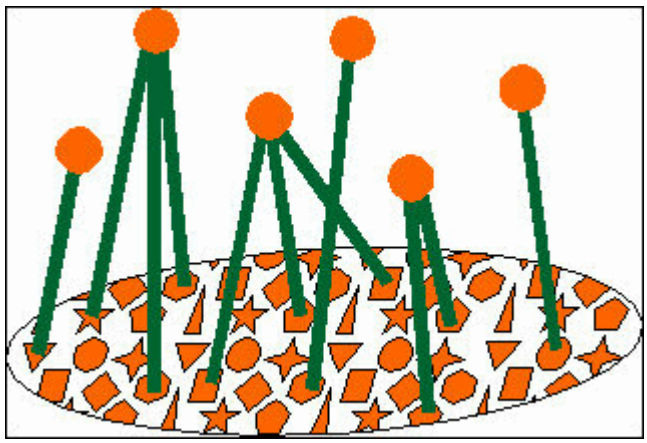
The ability to be able to specify more than one topic name can be used to name topics within different *scopes* (about which more later), such as language, style, domain, geographical area, historical period, etc.

The second characteristic of a topic is that it can have one or more **occurrences**.

A topic occurrence is an occurrence (or set of occurrences) of a topic within one or more addressable information resources. It 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 a topic.

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.

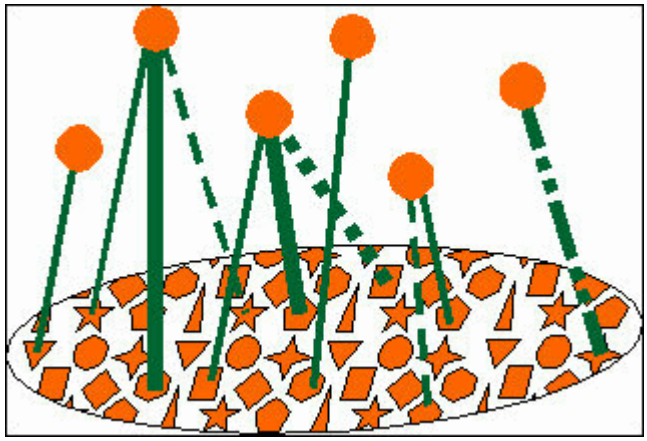
Fig. 4. Occurrences



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.

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**.

Fig. 5. Occurrence roles



As with topic types, occurrence roles are really topics, either explicitly or implicitly. If you define your occurrence roles as topics explicitly, you can use topic map facilities to say useful things about them (such as their names, and the relationships they partake in).

1.5. Topic associations

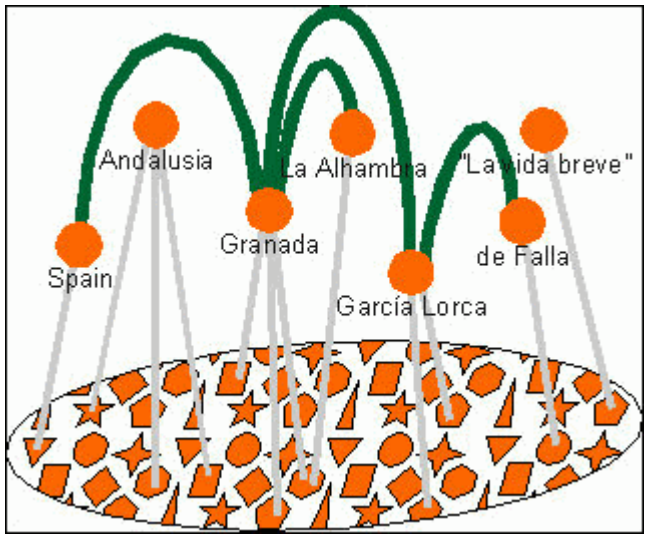
Up to now, all the constructs that have been discussed have had to do with topics as the basic organising principle for information. The concepts of “topic”, “topic type”, “name”, “occurrence” and “occurrence role” allow us to organise our information resources according to topic, and to create simple 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**.

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”

Fig. 6. Topic associations

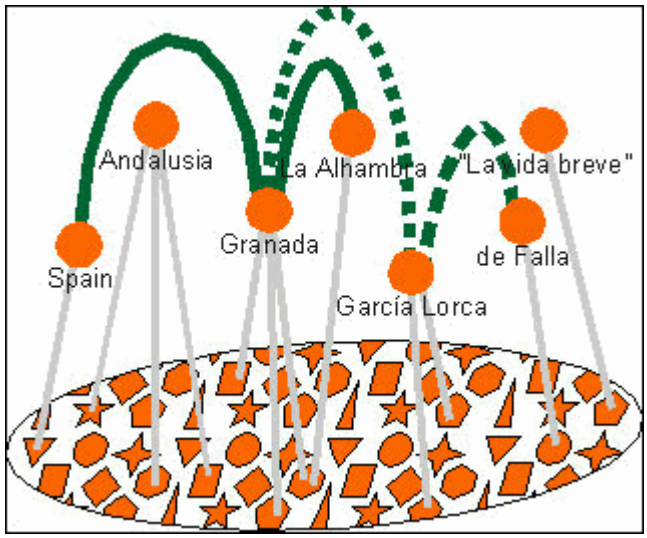


Just as topics can be grouped according to type (country, city, poet, etc.) and occurrences according to role (mention, article, commentary, etc.), so too can associations between topics be grouped according to their type. The **association type** for the relationships mentioned above might be “is in” (or geographical containment), “born in”, “written by”, “collaborated with”. As with most other constructs in the topic map standard, association types are themselves regarded as topics, whether or not they are explicitly declared to be so.

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 is worth noting that topic types can be regarded as a special 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 (indeed, for having a special construct for names at all): The semantics are so general and universal that it is useful to standardise them in order to maximise interoperability between systems that support topic maps.

Fig. 7. Association types



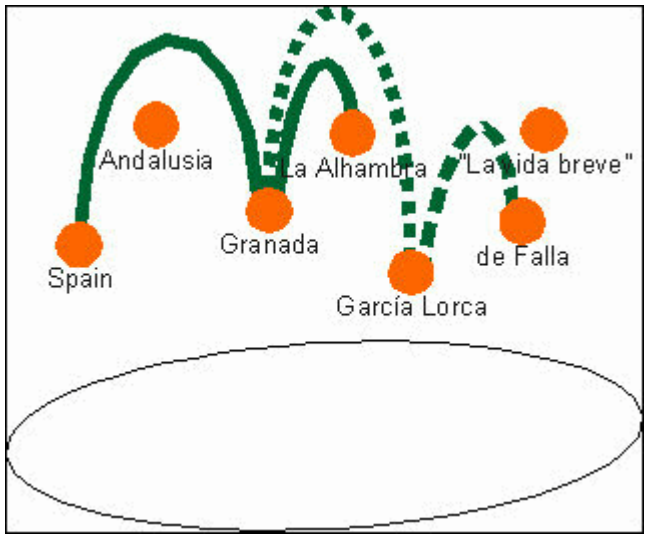
It should also be pointed out that while both topic associations and normal cross references are hyperlinks, they are very different creatures: In a cross reference, the anchors (or end points) of the hyperlink occur *within the information resources* (although the link itself might be outside them); 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, 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 (however this requires the additional concepts of *scope* and *public subject* to which we will return later).

Fig. 8. Topic maps as independent resources



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 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 class/instance and part/whole (meronymy/holonymy) relationships, 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 automatically 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.[5]

1.6. 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

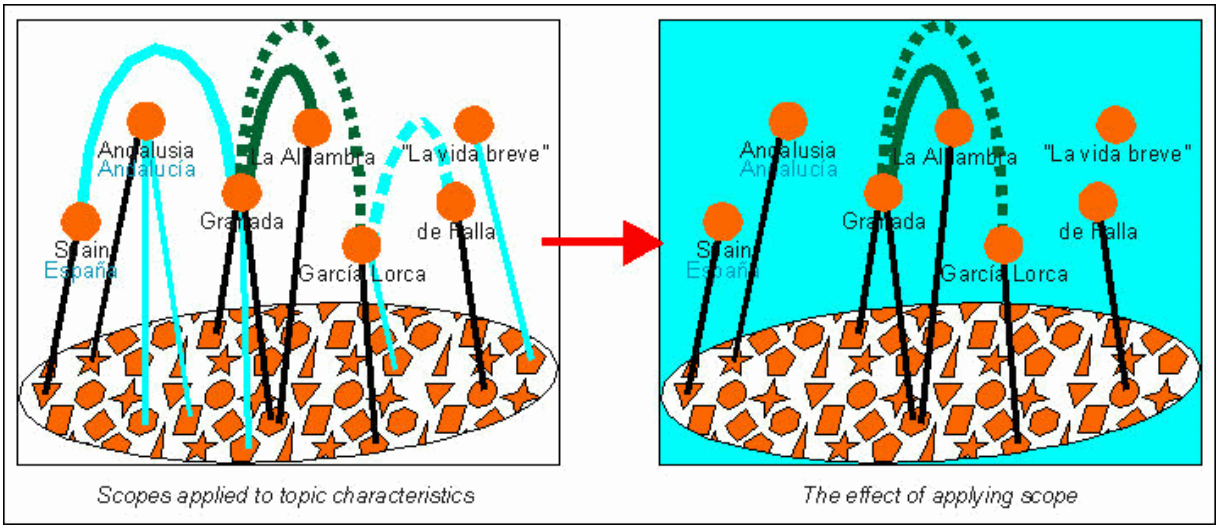
known as **topic characteristics**.

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”, you all know what I am talking about (at least those of you who are present at the GCA's XML Europe '99 conference). Or do you? How do you know that I'm 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 you are assuming a scope set by the context of what I have said up to now and where I have said it.

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.

Fig. 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.

1.7. 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, Denmark, Poland and Germany were to merge their topic maps, there would be a need to be able to assert that the topics “Spania”, “Spanien” and “Hiszpania” 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 an electronic resource which unambiguously identifies the subject in question. 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]

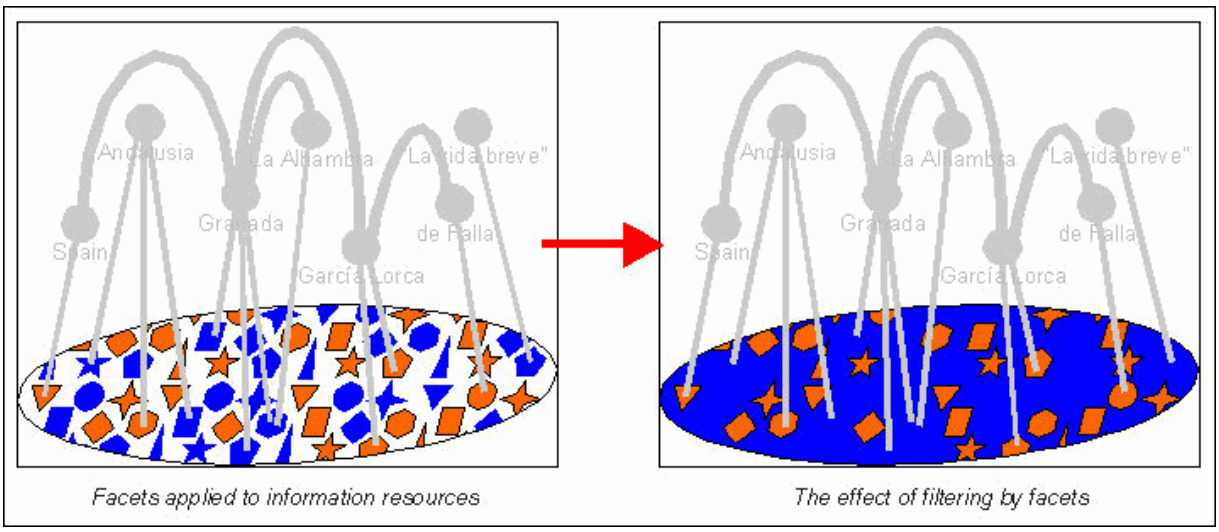
1.8. Facets

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”.

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*.

Fig. 10. Applying facets for filtering



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). In fact, at the time of writing it is being considered whether to actually split facets out as a separate architecture within the standard. Nevertheless, despite being orthogonal, facets provide a useful mechanism that

complements and significantly extends the power of topic maps.

1.9. Conclusion

The new topic map standard provides a standardised way of modelling the structure of the knowledge contained in information resources in such a way as to enable new means of navigation and ultimately also new means of organisation of that information.

The applicability of topic maps extends to all spheres of information management, not least commercial reference works and legal publishing.

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

2. 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?

Paradoxically, 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 paper, 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 *I* managed to spot;[7] 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 minimum. Both of these impose new requirements on the editorial system used to manage those assets.

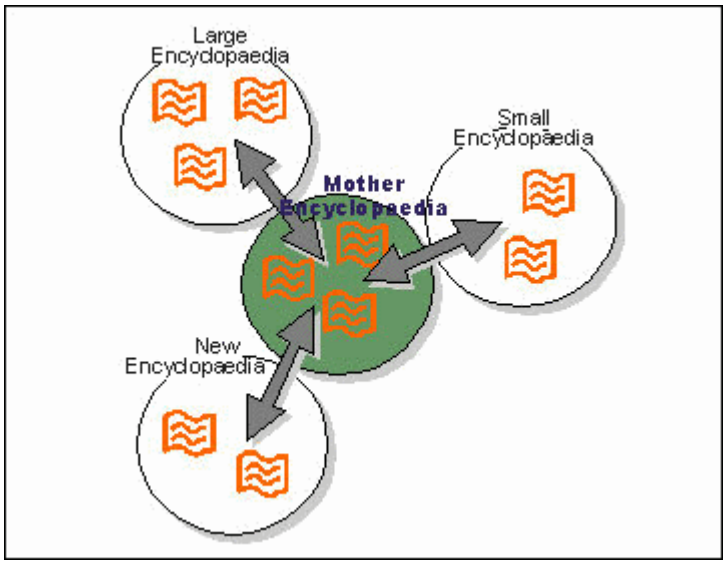
Before looking at how topic maps can help solve these problems, here is some background drawn from the ideas of two leading European publishers of reference works, PWN and KF.

2.1. 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 at this conference one year ago. 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.*[8]

Fig. 11. PWN's "Mother Encyclopaedia"



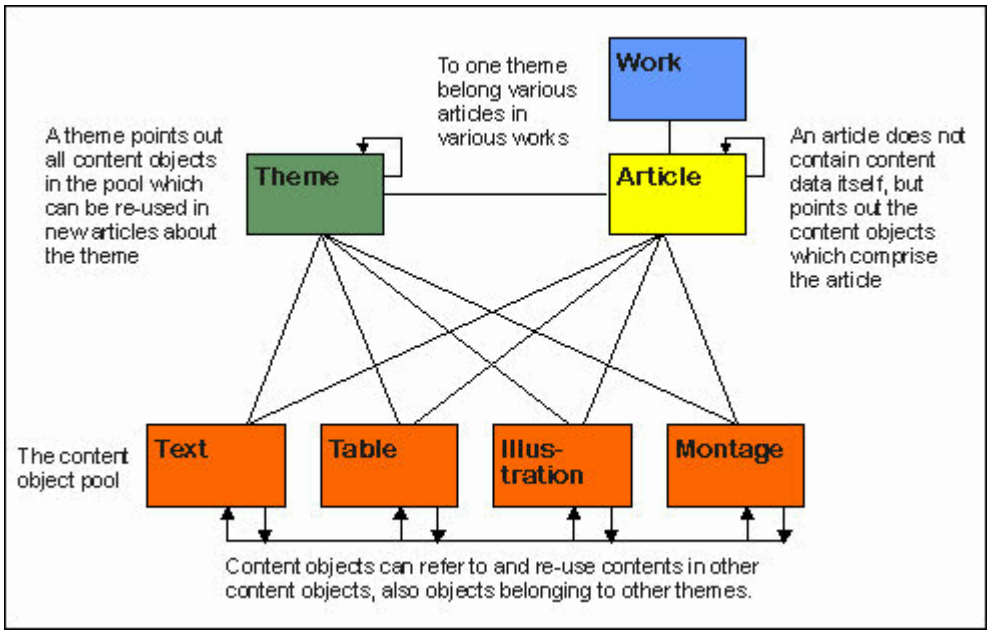
The benefits of this approach were identified as being: *identification* (i.e. establishing the connections between articles about the same subject in different works, and *ease of maintenance* (core data relating to a particular subject is stored in one place and therefore easier to update and reuse).

2.2. 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 organised 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.*

Fig. 12. KF's "Theme model"



> *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 Kunnskapsforlagets 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.*[9]

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.

2.3. 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 organising 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 organised 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:

- alphabetical list of headwords
- cross references within the text
- "see also" references

- 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, prioritised 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, etc. etc.

“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 (km2)	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"
>
```

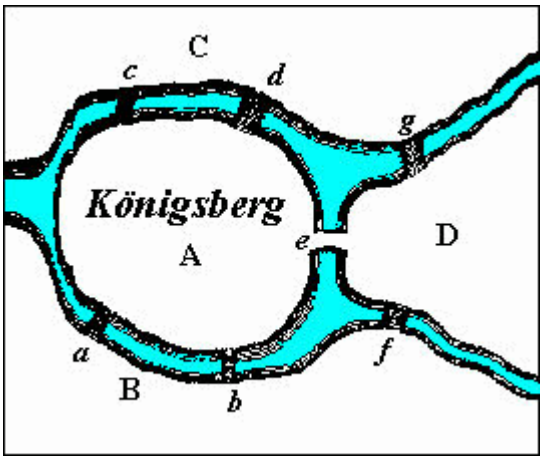
3. Euler, Topic Maps, and Revolution

So what does all this have to do with Euler – or revolution, for that matter?

Leonhard Euler, born in Switzerland in 1707, was one of the greatest mathematicians of all time – and also one of the most prolific. He spent most of his life serving at the court of Catherine the Great in St. Petersburg, and at the Academy of Sciences of Frederick the Great in Berlin. He contributed to every mathematical field that existed at the time, and created several new ones. Euler could turn his mind to any practical problem – provided it involved mathematics. One such was the famous *Bridges of Königsberg* problem.

The city of Königsberg (now the Russian enclave of Kaliningrad) lies astride the River Pregel (or Pregolya) at a point where it separates into two branches and forms an island. In Euler's day the river was crossed by seven bridges and it was a favourite Sunday afternoon pastime for the city's inhabitants to stroll around the town crossing the bridges.

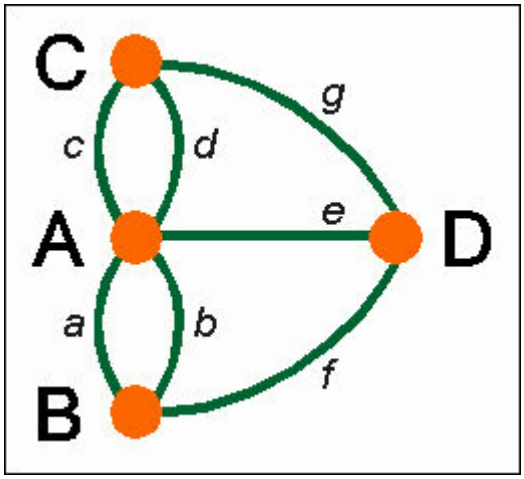
Fig. 13. The seven bridges of Königsberg



The question inevitably arose as to whether it was possible to find a route which would allow all seven bridges to be crossed in the course of an afternoon stroll without recrossing any of them. Since their attempts had always failed, most people believed that the task was impossible, but it was not until 1736 that the problem was treated from a mathematical point of view.

As you have guessed, Euler solved the problem – in fact by proving that no such route was possible. He recognised that the exact form of the land masses, river and bridges was irrelevant to the problem, and he reduced it to the following diagram showing four nodes (representing the land masses), connected by seven arcs (representing the bridges):

Fig. 14. Nodes and arcs representing Königsberg



From this simplification of the problem he proceeded to work out a number of principles that are relevant to any system of interconnected nodes, and thus gave rise to the branch of mathematics that is today known as “graph theory”.

The connection, of course, is that topic maps are also graphs and that the methods that are brought to bear by mathematicians and computer scientists to solve such diverse problems as the “shortest path problem”, the “Chinese postman problem”, the “travelling salesman problem”, and a host of others, can fruitfully be applied to extending our understanding of what topics maps are and how they can most

profitably be applied.

But that isn't the only connection: One of Euler's colleagues while at the Berlin Academy was the mathematician Jean Le Rond d'Alembert, and Euler once had a memorable encounter with the philosopher Denis Diderot in St.Petersburg during which he produced his famous algebraic “proof” of the existence of God.[10] d'Alembert and Diderot were well acquainted, of course, since they were the joint editors of *Encyclopédie*, generally regarded as the first modern encyclopaedia and an important factor in the development of ideas leading up to the French revolution – all of which somehow brings us back to our point of departure...

Life is full of connections; knowledge and creativity thrive on them, encyclopaedias, dictionaries and other reference works need to be able to exploit them – and topic maps are the tool that enable them to do so, efficiently and reliably. That's why they are part of the revolution.

[1]Ksiezyk, Rafal: “Plato, SGML and Revolution” in *Proceedings of the SGML/XML Europe '98 Conference*, Graphics Communications Association, Alexandria 1998.

[2]Private communication to the author.

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

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

[5]The current version of the topic map standard has “built in” support for expressing transitivity only for class/instance relationships (through the topic typing mechanism). It was felt that more experience needed to be gained actually applying topic maps before any attempt be made to formally standardise a taxonomy of basic relationship types. This would be a very fruitful area for further research.

[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.

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

[8]Ksiezyk, op.cit.

[9]Henriksen, Petter: *Kunnskapsforlagets publishing system – PUS: Specification of requirements* (1997).

[10]Unfortunately the proof cannot be reproduced in these proceedings because the DTD in use does not support formulae!