JIS

# Research in action

# Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all

**Lars Marius Garshol**

*Ontopia, Oslo, Norway*

Received 29 March 2004

**Abstract.**

**To be faced with a document collection and not to be able to find the information you know exists somewhere within it is a problem as old as the existence of document collections. Information architecture is the discipline dealing with the modern version of this problem: how to organize web sites so that users can actually find what they are looking for. Information architects have so far applied known and well-tried tools from library science to solve this problem, and now topic maps are sailing up as another potential tool for information architects. This raises the question of how topic maps compare with the traditional solutions, and that is the question this paper attempts to address. The paper argues that topic maps go beyond the traditional solutions in the sense that they provide a framework within which they can be represented as they are, but also extended in ways which significantly improve information retrieval.**

**Keywords**: information organization; information retrieval; subject searching; world wide web; web sites; comparative studies; topic maps; metadata; thesauri; taxonomies; ontologies; classification

*Correspondence to*: larsga@ontopia.net

## 1. Introduction

The task of an information architect is to create web sites where users can actually find the information they are looking for. As the ocean of information rises and leaves what we seek ever more deeply buried in what we do not seek, this discipline becomes ever more relevant. Information architecture involves many different aspects of web site creation and organization, but its principal tools are information organization techniques developed in other disciplines. Most of these techniques come from library science, such as thesauri, taxonomies, and faceted classification.

Topic maps are a relative newcomer to this area and bring with them the promise of better-organized web sites, compared to what is possible with existing techniques. However, it is not generally understood how topic maps relate to the traditional techniques, and what advantages and disadvantages they have, compared to these techniques. The aim of this paper is to help build a better understanding of these issues.

## 2. Objects and their metadata

Metadata is the foundation of all information retrieval, and so we start out by examining what metadata really is. A later section will consider the relationship between metadata and library science techniques.

### *2.1. What is metadata?*

It is generally assumed when organizing information that it consists of discrete pieces, though the terms used

for these vary. The pieces are sometimes referred to as 'documents', at other times as 'objects'. We will use the term *object* here for the entities being organized, as it does not seem appropriate to assume that they will all be documents in the traditional sense of the word.

*Metadata* is generally defined as 'data about data', which is of course a very broad definition. In computer science this is generally taken to mean information about a set of data in a particular representation, which typically means schema information, administrative information, and so on. However, in content management and information architecture, metadata generally means 'information about objects' ('objects' here used as defined above), that is, information about a document, an image, a reusable content module, and so on. Since it is the management of content we are primarily concerned with here, this is the definition we will use throughout this paper.

The best-known vocabulary for metadata is Dublin Core [1], which is a set of 13 properties that may be applied to information resources to describe them. These properties contain information such as 'title', 'creator', 'subject', 'description', 'publisher', 'date', 'language', etc. The Dublin Core specification defines the meaning of each property, but is silent on how to represent both the properties and their values, and is thus independent of any particular technology. Dublin Core was intended to aid 'resource discovery', that is, it was meant to support information retrieval. However, it is now commonly agreed that metadata is as useful for the management of content as it is for the discovery of it after publication, and so metadata in practice tends to be used for both purposes.

In general, metadata is best understood as 'any statement about an information resource', regardless of what it is being used for, which metadata vocabulary is being used, and how the metadata is represented. In this paper we will focus on the use of metadata as a finding aid, and ignore the other uses to which it may legitimately be put.

### *2.2. Metadata as a finding aid*

Obviously, searching for specific information in a large document corpus in the absence of *any* form of metadata (that is, information about the objects) is pretty much a hopeless task. The question is, what kind of information about the objects would help the user the most?

One common case is where the user has seen the object that is being sought once already, and so may remember specific details about it, such as words from the title, who wrote it, or when it was written. These clues can then be used to find the document by searching using the clues and trying out different searches until the right document shows up. Dublin Core metadata supports this form of retrieval quite well, since this is precisely the sort of information it contains.

This is a special case, however, and in the more general case the user wants information about a specific subject(1) and sits down in front of whatever user interface is offered to find the answer to the question 'what objects are about subject X?' The question then is, what help can the user interface give this user?

If we assume that the interface is based on Dublin Core metadata, it turns out that the answer is: not that much. Table 1 shows Dublin Core metadata for a paper presented at XML Europe 2003. Clearly this information does not help the reader much in trying to establish what the paper is actually about, though from the subject we can see that it has something to do with topic maps and Resource Description Framework (RDF). This highlights the problem: that standard metadata mostly provides administrative information, and that it says very little about the subject of an object. Of the Dublin Core metadata properties only a few address this question, and most of them do so indirectly:

- **title**: the title of the document usually offers good clues as to what the document is about, but it does not necessarily mention all names of all subjects the user is interested in, and it may also presuppose knowledge the user does not actually possess.
- **description**: this field is likely to describe what the document is about, but again may not facilitate search and discovery very effectively, for the same reasons that the 'title' field may fail to do so.
- **subject**: this field, which usually contains a set of keywords, is meant to convey precisely what the document is about. However, much depends on how extensive the set of keywords is, whether all related subjects are mentioned, and whether too many subjects are listed, leading the user to get too many hits.

Table 1

| | |
|---|---|
| Title | Curing the Web's Identity Crisis |
| Creator | Steve Pepper & Sylvia Schwab |
| Subject | RDF, topic maps, subject indicator |
| Publisher | IDEAlliance |
| Date | May 2003 |
| Language | English |
| Format | XML |

### 2.3. Subjects and precision

In addition to metadata not necessarily saying very much about the aboutness of an object a related problem is that making metadata describe the subject precisely may also be difficult. Let us imagine a user sitting down in front of an interface to all the papers presented at the IDEAlliance conferences, using a search interface based on Dublin Core, and looking for information on topic maps. The user is new to topic maps, and not yet interested in any specific aspect of topic maps, just the subject area in general.

If the user now does a search for 'topic maps' in the keywords (that is, the 'subject' field), all papers which mention 'topic maps' as a keyword will be in the search result. One problem is that this will tend to be both introductory material as well as more specialized material, and the results will be a simple list of documents, probably showing title, author, and date.

The title, author, date, and description fields are very useful here, as they help the user choose between the results, but what of the quality of the results themselves? I actually tried this out[(2)] on the document corpus described in [2], and got the result shown in Table 2, in 'most recent first' order. (A metadata structure as simple as Dublin Core does not allow a 'most relevant' ranking.)

A glance at this list shows that most of these documents are not primarily about topic maps, but about subjects *related to* topic maps. However, if the authors had not listed 'topic maps' as a keyword, those searching for 'topic maps' would have been unable to find their papers at all.

Another problem with this corpus is that authors have been required to define their own keywords, which means that the choice of keywords can be quite eclectic. A random pick of the more unusual keywords from the corpus mentioned above finds ''xml' reserved', 'analyze evalapproach', 'pool', 'fnctional composition' (sic), 'semantic', 'Contrasting approaches: DTD, W3C schema, RELAX NG', 'Exposition: Recent history of MPEG-7', etc. Clearly these are not good keywords, for several reasons.

Different forms of the same keyword, or closely related keywords, is also a problem. In the corpus can be found 'topic navigation maps' (old name for topic maps), 'topic maps', 'XML topic maps' (a format for topic maps, often used as a synonym), 'XTM' (the acronym for the former), and so on. The problem here is that four keywords refer to two subjects, and that the two subjects are very closely related. None of this is actually captured, and the user simply has to search to find out, or know beforehand.

To conclude: the most useful metadata about a document is the keywords, since that is the only thing that explicitly describes what the document is about. The other metadata are useful in managing the documents and in helping the user decide which of their search hits they want to look more closely at.

The other conclusion to be drawn from this analysis is that having keywords be just a simple text field with no restrictions is not going to work very well.

## 3. Subject-based classification

Subject-based classification is any form of content classification that groups objects by the subjects they are about. This can take many forms, and is generally combined with other techniques in order to create a complete solution. In the example above, the use of

Table 2

| Title | Creator | Date |
|---|---|---|
| Tax Map: An integrated navigation tool for the IRS Call Center Research System | Michel Biezunski | December 2003 |
| Curing the Web's Identity Crisis: Subject Indicators for RDF | Steve Pepper & Sylvia Schwab | December 2003 |
| OKS 2.0: New Utensils for Topic Map Chefs | Pamela Gennusa | December 2003 |
| Semantic Web Servers - Engineering the Semantic Web | Graham Moore | December 2003 |
| Information Architecture with XML | Peter Brown | December 2003 |
| BookBuilder: Content Repurposing with Topic Maps | Nikita Ogievetsky & Robert Sperberg | December 2003 |
| The TAO of Topic Maps | Steve Pepper | December 2003 |

keywords to classify papers is a subject-based classification approach.

The relation between subject-based classification and metadata is that metadata properties or fields that directly describe what the objects are about by listing discrete subjects use a subject-based classification. This basic feature is common to all subject-based classifications, and as we will see the differences between the various techniques lie in what they say about the subjects, rather than in what they say about the objects.

This needs to be emphasized: there is a difference between describing the objects being classified and describing the subjects used to classify them. What we will discuss in this section are the different approaches to describing subjects. Metadata describes objects, and one of the ways in which it does that is by connecting objects to the subjects they are about. We will return to this idea below.

### 3.1. Controlled vocabularies

Controlled vocabulary is a rather broad term, but here we mean by it a closed list of named subjects, which can be used for classification. In library science this is sometimes known as an *indexing language*. The constituents of a controlled vocabulary are usually known as terms, where a *term* is a particular name for a particular concept. (This is pretty much the same as the common-sense notion of a keyword.)

It is common to distinguish between *term* and *concept* by saying that the former is the name of a concept, and that the same concept may have multiple names, and also that the same term may name multiple subjects. A controlled vocabulary consists of terms, and not directly of concepts, and in general each term will be disambiguated to refer to a single subject (that is, there will be no duplicate terms). Note that 'subject' as we have used the term so far is effectively equivalent to 'concept'.

The term vocabulary also means slightly different things in the term 'controlled vocabulary' from what it does in 'metadata vocabulary'. The former is, as we noted, a set of indexing terms, or subjects used for classification, while the second is a set of properties of objects.

The purpose of controlling vocabulary is to avoid authors defining meaningless terms, terms which are too broad, or terms which are too narrow, and to prevent different authors from misspelling and choosing slightly different forms of the same term. Thus we can prevent authors from using 'topic navigation maps' and 'topic map' by forcing them to choose 'topic maps'. We can also make it impossible to use 'fnctional composition' as an indexing term instead of the correct 'functional composition'.

The approach taken for IDEAlliance conference proceedings is what is known as an *uncontrolled vocabulary*, and this was abandoned recently in favour of a controlled vocabulary for precisely the reasons cited above. The simplest form of controlled vocabulary is simply a list of terms and nothing more. This is the approach currently used for the IDEAlliance conference proceedings, but more advanced schemes exist, as we will see below.

### 3.2. Taxonomies

The term taxonomy has been widely used and abused to the point that when something is referred to as a taxonomy it can be just about anything, though usually it will mean some sort of abstract structure. Taxonomies have their beginning with Carl von Linné[3], who developed a hierarchical classification system for life forms in the 18th century which is the basis for the modern zoological and botanical classification and naming system for species. In this paper we will use *taxonomy* to mean a subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy without doing anything further, though in real life you will find the term 'taxonomy' applied to more complex structures as well.

The benefit of this approach is that it allows related terms to be grouped together and categorized in ways that make it easier to find the correct term to use whether for searching or to describe an object. For example, this could help users and authors by making it clear that there are two closely related terms: 'topic maps' and 'XTM', and helping them choose the right one. (Or, at least, for the users, telling them that they should perhaps try both.)

Figure 1 shows the placement of topic maps within a hypothetical taxonomical structure. As can be seen, this structure could easily help someone looking for information on topic maps or classifying a document to do with topic maps to pick the right terms to use.

Note that the taxonomy helps users by describing the *subjects*; from the point of view of metadata there is really no difference between a simple controlled vocabulary and a taxonomy. Metadata only relates objects to subjects, whereas here we have arranged the subjects in a hierarchy. So a taxonomy describes the subjects being used for classification, but is not itself metadata; it can be used in metadata, however. Figure 2 illustrates this.

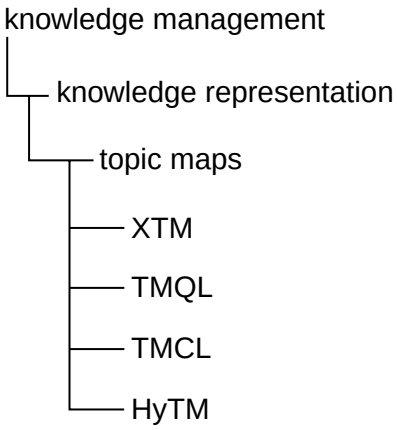


Fig. 1.

In Figure 2, the diagonal lines are metadata, while the horizontal and vertical lines that make up the taxonomy are part of the subject-based classification scheme. The distinction derives from the diagonal lines being statements about the paper, but the line between 'topic maps' and 'knowledge representation' is *not* a statement about the paper; it is a statement about 'topic maps'. One consequence of this is that if we have another paper about 'topic maps' we do not need to repeat that 'topic maps' belong under 'knowledge representation'.

As we said, the taxonomy provides more information about the concepts, and it does so to help the users. However, while the taxonomy does help the user, a number of important pieces of information about the concepts are not being captured here, such as:

- The fact that 'XML Topic Maps' is synonymous with 'XTM'.
- The difference between 'XTM' and 'topic maps'. (Many users use these interchangeably, but they do not mean the same thing.)
- The fact that 'topic navigation maps' is synonymous with 'topic maps', but should no longer be used.
- The relationship between topic maps and subject-based classification and topic maps and the semantic web.
- The relationship between XTM and XML and HyTM and SGML.
- The similarity between HyTM and XTM, and their difference from TMQL and TMCL, as well as the similarity between TMQL and XQuery.

All of these have consequences for the end user, since it means that they must search using precisely the right term, look in precisely the right places to find the terms, and so on. A taxonomy as we defined it here cannot handle these problems, though it should be noted that many systems referred to as taxonomies to some extent can, as they extend the basic model defined here.

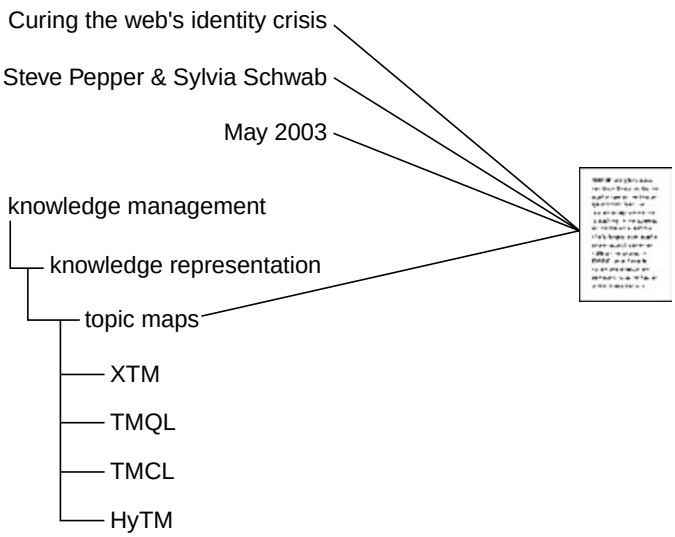


Fig. 2.

### *3.3. Thesauri*

Like the term 'taxonomy' the term 'thesaurus' has been used to describe all kinds of subject classification structures, though for thesauri there are two ISO standards describing their structure: ISO 2788 [3] describes monolingual thesauri, while ISO 5964 [4] is for multilingual thesauri. We will here discuss thesauri as they are defined in the ISO standards, while noting that in practice many users extend the structure somewhat, and in some cases the term is applied to structures differing substantially from what is described here.

Thesauri basically take taxonomies as described above and extend them to make them better able to describe the world by not only allowing subjects to be arranged in a hierarchy, but also allowing other statements to be made about the subjects. ISO 2788 provides the following properties for describing subjects:

**BT** short for 'broader term', refers to the term above this one in the hierarchy; that term must have a wider or less specific meaning. In practice some systems allow multiple BTs for one term, while others do not. (There exists an inverse property known as NT, for 'narrower term', which is implied by BT.) One could say that taxonomies as

described above are thesauri that only use the BT/NT properties to build a hierarchy, and don't make use of any of the properties described below, so it could be said that every thesaurus contains a taxonomy.

**SN** This is a string attached to the term explaining its meaning within the thesaurus. This can be useful in cases where the precise meaning of the term is not obvious from context. 'SN' stands for 'scope note'.

Since users often use 'XTM' when they mean 'topic maps', it would be useful to add a scope note to XTM saying something like 'The standard XML interchange format for topic maps. When discussing topic maps in general, and not just the format specifically, use the term 'topic maps'.'

**USE** Refers to another term that is to be preferred instead of this term; implies that the terms are synonymous. (There exists an inverse property known as UF.) For example, on 'topic navigation maps' we could put a 'USE' property referring to 'topic maps'. This would mean that we recognize the term 'topic navigation maps', but that 'topic maps' means the same thing, and we encourage the use of 'topic maps' instead. If we do this we would also have a 'UF' property on 'topic maps' referring to 'topic navigation maps', since this is implied by the 'USE' relationship.

**TT** Is short for 'top term', and refers to the topmost ancestor of this term. The term at the other end of this property is the one that would be found by following the 'BT' property until you reach a term that has no 'BT'. This property is strictly speaking redundant, in the sense that it doesn't add any information, though it may be convenient.

**RT** Short for 'related term', refers to a term that is related to this term, without being a synonym of it or a broader/narrower term. For 'topic maps' we could use this to indicate that 'subject-based classification' and 'ontologies' are terms related to 'topic maps'.

In short, thesauri provide a much richer vocabulary for describing the terms than taxonomies do, and so are much more powerful tools. As can be seen, using a thesaurus instead of a taxonomy would solve several practical problems in classifying objects and also in searching for them.

### *3.4. Faceted classification*

The term *faceted classification* has (you saw this coming, right?) been used to mean many different things. It was originally proposed by S.R. Ranganathan in the 1930s [5] and works by identifying a number of *facets* into which the terms are divided. The facets can be thought of as different axes along which documents can be classified, and each facet contains a number of terms. How the terms within each facet are described varies, though in general a thesaurus-like structure is used, and usually a term is only allowed to belong to a single facet [6].

In faceted classification the idea is to classify documents by picking one term from each facet to describe the document along all the different axes. This would then describe the document from many different perspectives. Ranganathan's original proposal (also known as Colon Classification) consisted of five facets:

- **Personality:** This facet was intended for the primary subject of the document, and is considered the main facet.
- **Matter:** The material or substance the document deals with.
- **Energy:** The processes or activities the document describes.
- **Space:** The locations described by the document.
- **Time:** The time period described by the document.

The classification of a book on Norwegian rural architecture in the 17th century might run something like this:

- Personality: architecture.
- Matter: wood.
- Energy: design.
- Space: Norway.
- Time: 17th century.

Faceted classification may seem very different from a thesaurus, but in fact it could be seen as simply a very disciplined way to construct a thesaurus as well as to use it for classification purposes. This is not to say that faceted classification is not useful, only that it is less different from thesauri than it may seem at first glance.

There exists an XML interchange syntax for faceted classification, known as XFML (eXchangeable Faceted Metadata Language), which was inspired by XTM, and has some features in common with it. XFML does not require the use of any specific set of facets, nor any specific set of terms within each facet, but does use a thesaurus-like structure for the terms within each facet. For more information, see [7].

It should be noted that there exists a generalized view of faceted classification wherein each facet is generalized to the point where it becomes a general property, and where the notion of a document is generalized to mean any kind of object. In this view of faceted classification there is little difference between

faceted classification and ontologies as they are described below.

### 3.5. Ontologies

The term ontology has, need we say it, been applied in many different ways, but the core meaning within computer science is a model for describing the world that consists of a set of types, properties, and relationship types. Exactly what is provided around this varies, but these are the essentials of an ontology. There is also generally an expectation that there be a close resemblance between the real world and the features of the model in an ontology.

In this section we have essentially been discussing languages for describing the subjects used in subject-based classification, steadily progressing towards more powerful means of description. Ontologies represent the culmination of this progression, in the sense that all of the above are fixed-vocabulary languages for subject description, while ontologies have open vocabularies.

In a taxonomy the means for subject description consist of essentially one relationship: the broader/narrower relationship used to build the hierarchy. The set of terms being described is of course open, but the language used to describe them is closed, since it consists only of a single relationship.

Thesauri extend this with the RT and UF/USE relationships, and the SN property, which allow them to better describe the terms. (TT, being redundant, is ignored here.) Again the language is closed, since this is the entire vocabulary available for describing the terms. In fact, thesauri could in theory be considered an ontology where there is only one type, called 'term', one property, called 'scope note', and three relationships (BT/NT, USE/UF, and RT). In practice thesauri are not considered ontologies because their descriptive power is far too weak, precisely because of this limited vocabulary.

Faceted classification does not really extend this language, but provides a consistent and useful discipline for applying it. That is, faceted classification does not introduce any new properties or relationships, though it could be said to have a new type: facet. It could be described as simply requiring the creator of the indexing language to create a set of facets and then fill each with a thesaurus that does not overlap with the others.

With ontologies the creator of the subject description language is allowed to define the language at will. Ontologies in computer science came out of artificial intelligence, and have generally been closely associated with logical inferencing and similar techniques, but have recently begun to be applied to information retrieval. A number of technologies and tools exist in this area, but we will focus on topic maps in this paper, since they were created to be an ontology framework for information retrieval.

### 3.6. Other subject-based techniques

A number of other terms and techniques are commonly applied today within information architecture, without fitting neatly into the classification scheme used in this section. One such term is *categories*, which are often used to group objects on websites. Categories are just terms in a subject-based classification, that is, a controlled vocabulary. The categories can be a plain list, or they can be arranged in a taxonomy. That is really all there is to it. (This works the other way as well; every taxonomy or thesaurus consists of categories into which objects are grouped.)

Another technique that is much used is *synonym rings*, which connect together a set of terms as being equivalent for search purposes. (That is, if you search for 'topic navigation maps' you should also find 'topic maps', for example.) Essentially synonym rings express a synonym relationship between a set of terms, and so are similar to the UF/USE relationship of thesauri, except that there is no indication of one term being preferred above the others. Synonym rings are a rather special-purpose construct, but their function, to indicate that a set of terms are synonymous (that is, refer to the same concept), is very worthwhile indeed. The term 'synonym ring' alludes to the fact that within a synonym ring every term is synonymous with every other term in the same ring; mathematicians know this as an *equivalence class*.

An *authority file* is similar to a synonym ring, the only difference being that it consists of UF/USE relationships instead of synonym relationships. So in an authority file one term in each synonym ring is indicated as being the preferred term for that subject.

What this means is that a thesaurus includes not just a taxonomy, but also an authority file. It does not include a synonym ring, however, though it could be used to support searching in the same way that a synonym ring is used.

## 4. Topic maps

Topic maps originated in work on the merging of electronic indexes and so are very much a subject-based

classification technique. In fact, topic maps are organized around *topics*, and each topic is used to represent some real-world thing. In the terminology we used above, topics represent concepts, the same way terms in an indexing language refer to concepts. In topic maps the concepts are called *subjects*, and the standard emphatically states that a subject can be 'anything whatsoever'. We will return to the consequences of this later on.

In topic maps, three constructs are provided for describing the subjects represented by the topics: names, occurrences, and associations. These describe the names, properties, and relationships of subjects, respectively, and we will cover the uses of each of these in more depth in the following sections. We will not give an in-depth introduction to topic maps, as that is beyond the scope of this paper. Those wanting more information should read the classic introduction [8].

### *4.1. The names of subjects*

In topic maps, a topic can be given any number of names. Giving more than one name to a topic effectively means that all the names refer to the same subject; that is, that the names are all synonyms. This means that every topic is really a synonym ring, though it can be a ring with just one member.

It is allowed in topic maps for different topics to have the same name(4), something that taxonomies and thesauri do not allow. In practice this is something that occurs all the time, and so it is important to support it. Many classification systems avoid duplicate names by including some disambiguation in the name itself. For example, Paris the city and Paris the hero of Greek mythology may have been differentiated as 'Paris (France)' and 'Paris (Greek myth.)'. In topic maps this is not necessary, and the types, occurrences, and associations of the topics will generally distinguish them anyway.

The power of topic maps with regard to names does not stop here, however. A name can be given a *scope*, which is a set of topics representing the context in which the name is appropriate. A name that has an empty scope is considered to have unlimited validity. This enables the creator of the topic map to define a kind of language for describing names.

For example, one could create a topic for topic maps and give it the name 'topic maps' (with empty scope), but also the name 'topic navigation maps', in the scope 'obsolete'. This would have the same effect as the USE/UF construct in thesauri, or as an authority file, but is more powerful, since one can say *why* the term 'topic navigation maps' should not be used. (The reason being, of course, that it is obsolete.)

This approach to names enables new usage areas that do not exist in the same way with traditional tools. For example, different terminologies are often used within a single organization, whether because of regional differences, or because of differences in corporate cultures within the organization. Topic maps can support this by scoping the names with topics representing the corporate cultures or areas where the terms are used. This allows the topic map to say that 'corporate culture A calls this region "APAC", while culture B calls it "Asia-Pacific"'.

Users can then in their profile (which may be permanent, or just for a single visit to the site) state their preferred terminology, and the site can then display the correct names for each topic for them. In cases where a topic has no name in the scope 'corporate culture A', the unscoped name will be chosen instead.

This can also be used to support multilingual information, since topics representing languages can also be used as scopes. Thus we could give the topic for topic maps (English) the name 'emne kart' (Norwegian) and have the system display the names in the user's preferred language.

Since scopes consist of topics, it is up to the creator of the topic map to define their own language for describing the names of subjects. Topic maps provide all the power of traditional techniques in this regard, but go far beyond it, in that they remove several of the restrictions, and also allow the relationship between the terms and the subjects to be described in full, instead of expressed in terms of specific desired behaviour ('prefer this term', or 'treat these terms as equivalent').

### *4.2. Types*

In topic maps topics can be typed, which provides considerable power for describing the world from which the topics are taken. This is a capability that is missing from traditional classification techniques. Using this, one could create types and assign them to topics, and thus say that 'topic maps' is-a 'technology', 'XTM' is-an 'interchange format', 'Norwegian' is-a 'language', 'HyTM' is-an 'interchange format', 'TMCL' is-a 'constraint language', and 'TMQL' is-a 'query language', and so on. (This covers a number of the statements we could not express in our discussion on taxonomies – Section 3.2.)

This may sound like a very simple capability, and so it is, but it is also very powerful, and provides the

foundation for several important capabilities that we will return to. The most immediate benefit, however, is that in traditional techniques the set of terms is in a sense flat, since there is no way to distinguish different *kinds* of terms. Languages are bundled in with people, places, technologies, and organizations, with no way to tell them apart.[5]

Once explicit types have been provided it is possible to let the user perform searches such as 'find "paris", but show only "places"', or to show lists of all cities, separate from other kinds of subjects. Without types there is no way to do this, since the necessary information will be missing.

Again, since the types are themselves topics the creator of the topic map can choose which types to use, which is another reason why we say that topic maps have an open vocabulary. (It may seem confusing that types are not among the constructs listed at the beginning of this section. The reason is that type assignments are considered a relationship, that is, an association, a category to which we will return below.)

### *4.3. Occurrences*

Occurrences relate topics to the information they are relevant to. They effectively perform the same function as the page numbers in a back-of-book index: they indicate where information about the subject can be found in the book. And so it is with occurrences, they connect the topics to information resources that contain information about them. This is effectively the inverse of the 'subject' property in Dublin Core, which connects an object to the subjects it is about. Occurrences in topic maps do the same thing, but go from the subject (represented by a topic) to the object (or information resource).

In topic maps the representation of the resource/subject relationship is more structured, however. Occurrences have types, which allow us to distinguish between different kinds of relationships to the information resource. One can distinguish a biography from a portrait, a description from a tutorial, a video clip from a specification, and so on. Scope can also be applied to an occurrence, for example to distinguish material suitable for novices from that suitable for intermediate learners.

Occurrences use Uniform Resource Identifiers (URIs) to identify the information resource being connected to the topic, which means that any kind of information resource anywhere can be connected to the topic. However, the occurrence does not need to use a URI; the information can also be given as a string stored directly in the topic map. This is useful for attaching simple properties to topics, such as date of birth, phone number, description, and so on.

Again, the occurrence types are topics, and so the creator of the topic map is free to define occurrence types at will. To compare with traditional techniques, they generally do not provide an open vocabulary for properties and occurrences, and although in some cases they can be extended to support one that support tends to be weak. Having very specific properties like 'phone number', 'date of birth' etc. on terms can be very awkward when terms cannot be typed, since most specific properties only apply to a few types and make no sense when attached to other types. (What is the date of birth of a city? Or the phone number of a technology?)

### *4.4. Associations*

Associations represent relationships between subjects, and like occurrences they can be typed. This allows any kind of relationship to be expressed. The relationships in traditional classification schemes have very little semantic content, whereas in topic maps one generally tries to make the typing of associations as specific as possible.

Associations are the final construct we need in order to be able to fully represent the set of statements given at the end of our discussion on taxonomies (section 3.2), and Figure 3 shows the result.

In this figure, types are indicated with different shapes, while we have left out the alternative names. The types of the associations can be seen from the labels on the lines.

Compared to traditional classification schemes this is very different. First of all, we no longer have a hierarchy but a network of subjects. Secondly, the relationships between the subjects are clearly defined instead of being generic. From the point of view of searching, this is very powerful, since it allows us to do queries like 'show me all technologies used with topic maps', or 'show me every interchange format based on SGML', and so on. There are also other uses, as we will see.

## 5. Comparison

A summary of the relationship between topic maps and traditional classification schemes might be that topic maps are not so much an extension of the traditional schemes as on a higher level. That is, thesauri extend taxonomies, by adding more built-in relationships and

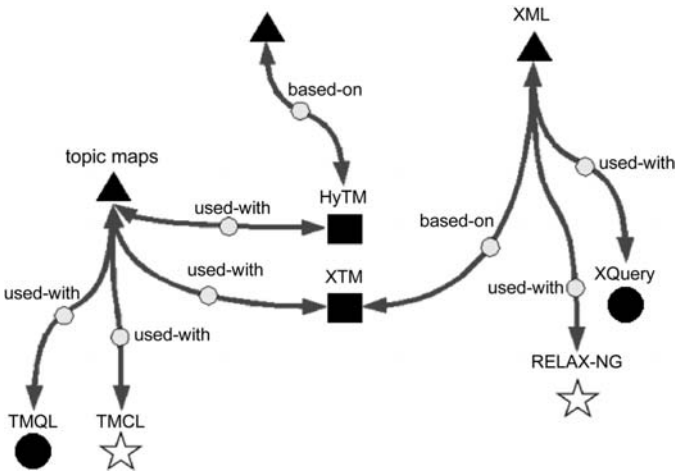


Fig. 3.

properties. Topic maps do not add to a fixed vocabulary, but provide a more flexible model with an open vocabulary.

A consequence of this is that topic maps can actually represent taxonomies, thesauri, faceted classification, synonym rings, and authority files, simply by using the fixed vocabularies of these classifications as a topic map vocabulary. We'll show how this works in the following section.

### *5.1. Traditional classifications in topic maps*

Kal Ahmed has created a standard topic map ontology that can be used to represent thesauri in topic maps [9]. There are, essentially, two ways to do this, one that models terms as subjects (one topic per term) and one that models terms as names (one name per term; one subject per concept). In the former approach we would be creating a topic map where the subject of each topic would be a term, which is different from the latter approach where the subject of each topic would be a concept. We will only consider the latter approach here.

To create a topic map from a thesaurus using this approach, follow these principles:

- For each term that has no USE relationship, create a topic of type 'term', and make the term the name of the topic.
- For each term that has a USE relationship, make the term a name (in scope 'non-preferred term') of the topic for the preferred term.
- A scope note is an occurrence of type 'scope note'.
- The RT relationship is represented as an association of type 'related term'.
- The BT and NT relationships are represented as an association of type 'broader/narrower' (with the roles specifying which topic is broader and which is narrower).

That is all. From the resulting topic map you can easily reproduce the original thesaurus, and the resulting structure really is a thesaurus in the same way that the original data was. Doing this is also quite easy; converting a thesaurus in a simple text format to a topic map can usually be done in a couple of hours. Note that this also covers taxonomies, since they just have names and the BT/NT relationships, so nothing more is needed to support them. Any extensions that have been made to the thesaurus or taxonomy model can easily be represented in a similar way.

For faceted classification the situation is very similar. A standard topic map ontology for faceted classification based on XFML exists, as does an XSLT style sheet for converting XFML documents to topic maps. For more information, see Garshol [10]. The XFML ontology extends the thesaurus ontology, by adding a new type 'facet' and a new association type 'belongs-to-facet'.

Creating a topic map from a faceted classification system is thus done as follows:

- For each facet, create a topic of type 'facet' and give it the name of the facet.
- For each top-level term within each facet, create a topic of type 'term', with the term as the name, and associate it with the facet it belongs to using the 'belongs-to-facet' association type.
- For each term below the top level, create a topic of type 'term', with the term as the name, and associate it with its parent using the 'broader/narrower' association type.

You have now represented your faceted classification as a topic map.

### *5.2. Merging metadata and classification*

The central message of this paper so far has been that metadata is statements about objects (for example, documents) while thesauri and similar techniques provide statements about subjects used in classification. Subject-based classification, of course, is the use of subjects in metadata.

Topic maps effectively unify these two approaches. Topic maps are organized around topics, which represent subjects. That is, in a topic map you find topics. Every topic you find represents a subject out in the real world that it is a symbol or stand-in for in the topic map. The definition of subject is essentially 'anything

whatsoever'. What this means is that from the point of view of a topic map, *objects are just a special kind of subject*.

What this means is that the whole machinery we have created to describe the concepts in a subject-based classification is also available to describe the objects being classified. Thus, we can create a topic type for objects ('document', perhaps), and express the metadata using names, occurrences, and associations. The topic types let us keep track of what is a document and what is a concept, but we no longer need different technologies for metadata and classification.

And once we have everything in a single representation we can start to cross the boundaries by for example describing the authors of a paper further, and connecting them with the terms from the subject-based classification. Effectively what has happened is that the straightjacket has been removed, and we can now say anything we want to. The technology is no longer the limiting factor; instead the limits are set by our imagination and by how much information we are able to maintain within the economic constraints of our projects.

An example of how this can be done is shown in Figure 4.

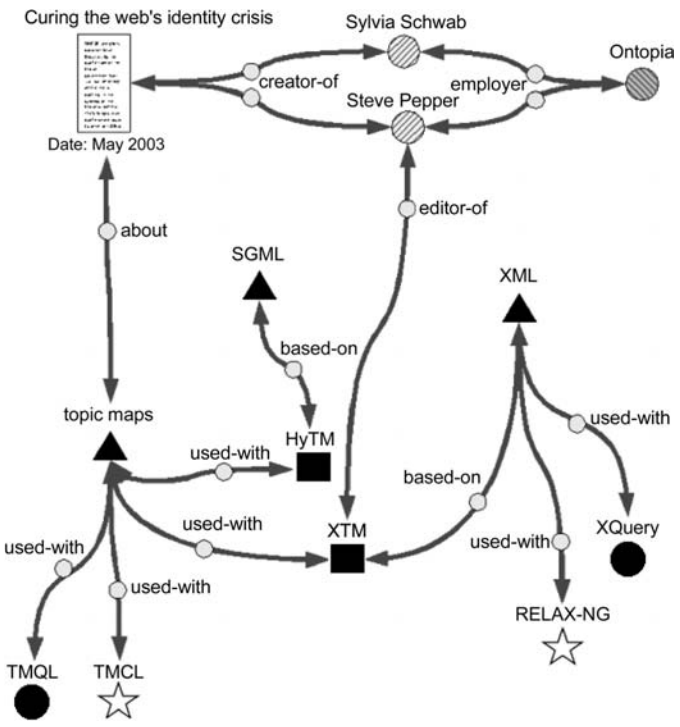


Fig. 4.

Clearly, this is not doable with traditional classification systems.

### 5.3. *Benefits and costs*

The main benefit of this approach is that it leads to more richly descriptive classification and metadata systems than the traditional approaches. This again means that more precise searches are possible (see below for more on this), and that navigation systems driven by this information can be much richer and more flexible than a simple tree-based navigation.

In fact, with a richly descriptive classification system like a topic map the information in the classification system becomes a valuable resource in its own right. In many cases one will find that information that was previously provided in the classified documents, such as the structure of the organization, information about project participation and ownership, etc., is duplicated in the classification system. The documents were necessary, because a hierarchical classification system would not be able to express all the information one wants to capture about the organizational units and projects. With topic maps this limitation goes away, and so the documents can also go away, and their function be taken over by the navigational structure of the portal. This can effectively turn the navigational structure into a knowledge management application, while allowing it to be used for classifying content.

Quite often when using topic maps in this way one finds that having captured information in a richly structured form enables the creation of entirely new information products from the same source. This is similar to how moving documents from word processing formats to XML enables multi-channel publishing from the same structured source, but on a higher level.

Another benefit of using a standard like topic maps is that a larger set of tools from various vendors become available. These tools will provide things like programming APIs, query languages, schema languages (see below), graphic visualization, portal integration, content management, workflow, natural language querying, and so on and so forth.

Of course, topic maps, like any other technology, have downsides. Experience with topic maps is still in its infancy, so there may be disadvantages and risks that are currently not known. One of the more obvious is that topic maps are a relatively new technology, and so topic map expertise may be harder to find than expertise in traditional classification.

Also, while topic maps allow a richer structure and allow the users to define it themselves, this also means

that the effort required to build a classification system may become greater, since there is more to build. To some extent the existence of predefined ontologies and the information design patterns provided by topic maps alleviate this, but the problem is still real, especially for newcomers to the technology.

### 5.4. Searching

Topic maps offer unprecedented power when it comes to searching, in the sense that they offer very good support for full-text searching, very good support for complex queries, and also provide an excellent basis for natural language querying.

Full-text searching a document corpus typically returns the documents that mentioned the term being searched for, but a topic map-based system can do far better. Instead of returning documents, the system can return the topics that best match, together with additional information, and this provides a starting point for jumping into the topic map and browsing around to find the answer to the specific question.

To take a simple example, if we were to do a full-text search for 'XSLT' in the conference proceedings for the IDEAlliance conferences, there is of course a huge number of hits, but at the very top comes the topic 'XSLT', which represents the XSLT standard. From there one can find the specification, papers about XSLT, which standards organization produced XSLT, tools implementing XSLT, tools using XSLT, etc.

However, because of the extra structure in topic maps, we do not have to stop here. We can ask for 'papers about topic map query languages', for example, or 'papers about topic map query languages given by someone based in Japan', or 'papers about query languages', or 'papers about RDF by someone who's written papers on topic maps'.

Through the use of a topic map query language these queries can get arbitrarily complex. The main problem is that the user has to know both the query language and the ontology of the topic map, and most users will be unwilling to do either. The solution to this is to build a friendly forms-based user interface where the user can set up the query by selecting terms in drop-down lists or suchlike.

There is, as mentioned above, a third possibility, which is natural language querying. Some experimentation on this has been done using topic maps, and it indicates that quite powerful search capabilities can be created by matching natural language against names in the topic map, then looking at the matched topics to see whether they are topic types, association types, or individuals, and finally putting together a query in a query language from this. It is too early to be able to conclude that this will work well for real end users, but the results so far seem very promising indeed.

### 5.5. Schemas

In topic maps, where users can define the classification vocabulary themselves, there is substantial risk that those creating the classification may use the vocabulary incorrectly. For example, they may use associations in ways that do not make sense, attach occurrences to topics of the wrong type, type topics inconsistently, or leave out required information.

For relational databases and XML the traditional solution to this problem is to define a *schema* that formally defines the allowed structure of the data set. This is also the solution used with topic maps, although at the moment there is no standard schema language for topic maps. A standard, called TMCL (Topic Map Constraint Language), is under development, and non-standard languages are available already.

Using a schema language it is possible to say things like 'only persons may be employed, and they must be employed by an organization', 'only persons and organizations can have phone numbers', 'query languages must be used with a technology', and so on. Thus the typing capabilities of topic maps give us a starting point from which to describe the classification rules, and the schema language gives us a way to express the rules that can be enforced by software.

### 5.6. Identity and merging

Topic maps have a mechanism for formally declaring the identity of a topic's subject. What this means is that in topic maps there is a way to state what the subject of a topic is, in such a way that if another topic is found (for example in another topic map) that represents the same subject, one can know with certainty that it does represent the same subject.

The identity mechanism in topic maps is based on URIs, which can be used in two different ways. If subject of the topic is an information resource the subject can be identified very easily: simply by referring to the resource using a URI. Two topics that refer to the same resource in this way must inevitably represent the same resource.

The problem is more difficult when the subject is not an information resource, but instead an abstract concept like 'Norway', 'Ontopia', or 'Lars Marius

Garshol'. None of these subjects have a URI that can be used to refer to them, so the same solution cannot be used in these cases. The solution is instead to create a resource defining the subject and then referring to the defining resource.

This might lead to confusion, in that it might be unclear whether the subject is the resource referred to, or the concept described by the source. Topic maps solve this problem by distinguishing between the two kinds of references, so that it is always clear what is being meant. So if two topics are found to be referring to the same resource saying 'my subject is what is defined by this resource' clearly both have the same subject.

The ability to know when two topics represent the same thing provides great power, which can be exploited in many different ways:

- Topic maps from different sources which have some overlapping data can be merged together automatically (provided they have used the same identifiers).
- Topic map modules containing commonly used information can be maintained separately and reused in different contexts whenever they are needed.
- Common vocabularies can be developed as sets of defined topics and reused in different organizations. The information maintained by different users can then be merged or interchanged as desired.
- Information from different sources can be brought together into a single topic map and integrated into a meaningful whole.
- Different web sites can be integrated with one another since they will know what on one web site corresponds to what is on the other.

This list is hardly comprehensive, since the thinking about how to exploit these possibilities is still quite young.

Nothing like this facility exists in traditional subject-based classification, though XFML provides the same facility. However, although XFML does make it possible to tell when two terms from different XFML maps represent the same concept, there is only limited help in that. If the different maps give the same concept different names, or different parents, or place them in difference facets, then XFML cannot represent the resulting merged concept. To some extent this is a problem inherent in how hierarchical classification systems model the world, which topic maps do not share.

## 6. Conclusion

What this paper has tried to show is that topic maps provide a common reference model that can be used to explain how to understand many common techniques from library science and information architecture. It has also showed how these techniques can be implemented using topic maps, and how topic maps can go far beyond the possibilities provided by traditional techniques.

By using topic maps to represent metadata and subject-based classification it is possible to reuse existing classifications and classification techniques, while at the same time describing the world more precisely where desired.

## Acknowledgements

Many thanks to Marte Brenne for references to useful definitions of some of the library science terms used in this paper. Many thanks to Murray Altheim for expanding my view of what faceted classification is and for providing useful references. Thanks also to Isabella Kubosch, Magnus Halle, Runar Eggen, and Tord Høivik for useful references. Many thanks to Pam Gennusa, Sylvia Schwab and Steve Pepper for reading and commenting on early versions of the paper. Many thanks to Tor Arne Dahl for useful corrections and questions.

## Endnotes

(1) The relation between the word 'subject' as used here and as defined in topic maps is important, and we will return to it later. Meanwhile, take 'subject' to mean 'any concept in which the user may potentially be interested'.

(2) The actual query, for those who would like to know, was: discussed-in (topic-maps: keyword, $PAPER: paper), presented-at (PAPER: paper, $CONFERENCE: conference), when-event ($CONFERENCE: event, $DATE: time) order by $DATE desc?

(3) Originally Carl Linnæus. The traditional English misspelling of the name is Carl Linnaeus.

(4) You may have heard of the 'topic naming constraint', which is a rule in topic maps that only allows duplicate names in certain limited cases. This rule has now been removed from topic maps, so the statement is true.

(5) In faceted classification the division into facets can to some extent provide this, though it need not do so, and even when it does the types tend not to be very granular – that is, they tend to not be very precise.

## References


[1] S. Weibel et al., *RFC 2413: Dublin Core Metadata for Resource Discovery*, The Internet Society (1998). Available at: www.ietf.org/rfc/rfc2413.txt (accessed 26 May 2004).

[2] L.M. Garshol and S. Pepper, The XML Papers. In: *XML 2002 Conference Proceedings* (IDEAlliance, 2002). Available at: www.ontopia.net/topicmaps/materials/xmlconf.html (accessed 26 May 2004).

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

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

[5] M. Steckel, *Ranganathan for IAs*. In: *Boxes and Arrows* (online magazine), 7 October 2002. Available at: www.boxesandarrows.com/archives/ranganathan_for_ias.php (accessed 19 April 2004).

[6] E. Svenonius, *The Intellectual Foundation of Information Organization* (MIT Press, Cambridge, MA, 2000).

[7] P. van Dijck, *Introduction to XFML*, XML.com, 22 January 2003. Available at: www.xml.com/pub/a/2003/01/22/xfml.html (accessed 19 April 2004).

[8] S. Pepper, *The TAO of Topic Maps*, Ontopia (2002). Available at: www.ontopia.net/topicmaps/materials/tao.html (accessed 19 April 2004).

[9] K. Ahmed, *Topic Map Design Patterns For Information Architecture*, XML 2003. Available at: www.techquila.com/tmsinia.html (accessed 19 April 2004).

[10] L.M. Garshol, *FML Ontology and Converter* (forthcoming). Available at: http://psi.ontopia.net/xfml