Topic maps – knowledge organisation seen from the perspective of computer scientists

Nils Pharo, Faculty of Journalism, Library and Information Science, Oslo University College

1. Introduction

Topic maps is a fairly new ISO certified standard (ISO 13250, 2002) for organising digital content. Its main implementation area has been the World Wide Web where it is used for structuring subject portals and similar document types. However, in theory a topic map can be used to model any real world-relationship, and the objects in the relationships do not have to be digitised.

The title of the paper signifies two things:

1. That the topic map standard is developed for facilitating knowledge organisation
2. That the topic map standard is developed by computer scientist rather than library and information scientists

I will start by giving a brief introduction to topic maps with examples showing how to create topics and relationships. Then I will discuss its usability in knowledge organisation of web resources. In the third part I will focus on some important challenges to successful implementation of topic maps. At the end of the paper I have added a tail suggesting how topic maps could be used to add value to bibliographic records.

2. The principles of topic map

The key concepts in topic maps are

- topics,
- associations, and
- occurrences.

Pepper's introductory text, the TAO of topic maps (Pepper, 2002) gives a good introduction to the basic idea behind the topic map technology. The point of departure for developing a topic map is to have a structured representation of the subject area in the form of an

ontology/taxonomy/thesaurus (Gilchrist, 2003) or something similar. The ontology will offer the contents of the topic map, where the topics will represent the entries in the major index. The ontology as such is not part of the topic map technology, the ontology can be characterised as the input to the topic map; it contains the topics and associations between topics. Garshol (2003) has compared the ontology-topic map relationship to the DTD-XML relationship, pointing out that the technology (e.g. XML and topic maps) is dependent on a specified content (e.g. DTD and ontology) if it is to be used in practice.

A *topic* is a resource within the computer that reifies some real-world subject. Topics can have *names*. They can also have *occurrences*, that is, information resources that are considered to be relevant in some way to their subject. Finally, topics can participate in relationships, called *associations*, in which they play roles as members.

The associations will tie together topics that the ontology constructor has found relate to one another in some respect.

One can say that the topic map technology offers a framework or architecture in which the user can organise representations of concepts and associative or other forms of links between these representations. A topic-page will typically link to objects (occurrences) which, in the form of digitised content (text, images, video etc), represent one or more aspects of the topic. There are no restrictions with respect to the structure or organisation of the topic, which may be hierarchical, network based or a combination of these.

It is usual to make a distinction between the meta level and the information level when understanding the role of topic maps. The relationships between the topics are implemented in the meta- or navigation-layer of the topic map whereas relationships between topics and resources representing the topics ties the content of the two layers together (see Figure 1)

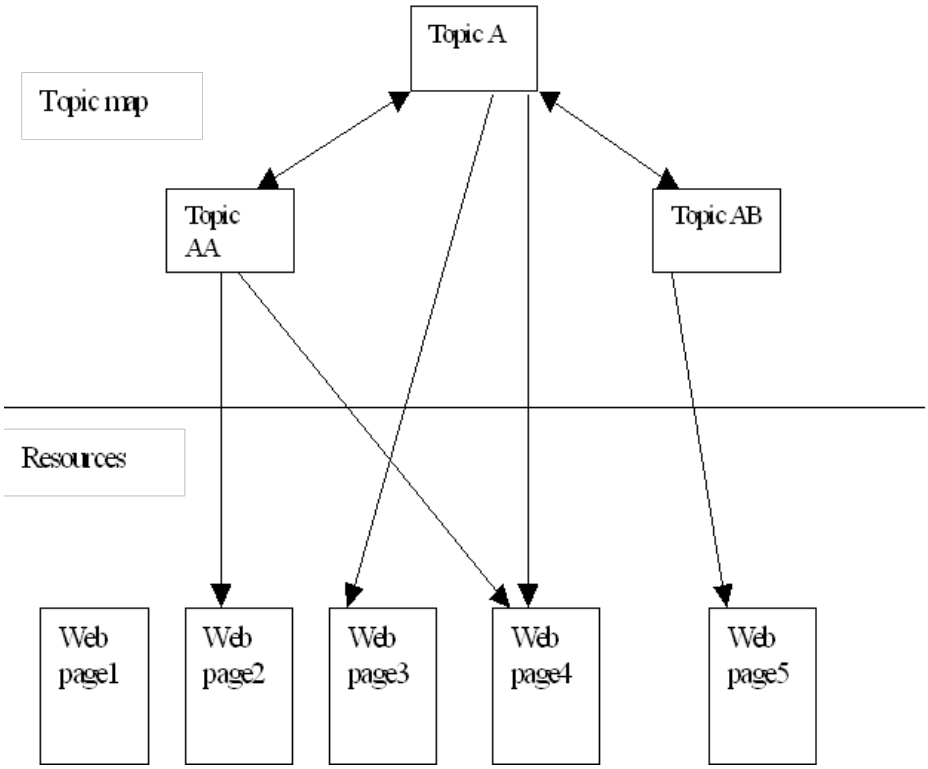

Figure 1 Topic maps' two layers

Currently there are several subject portals that are developed using topic maps (TM), most of these portals are Norwegian, probably due to the fact that there is a very strong TM-environment in Norway (participants in the ISO-standard committee, commercial firms selling TM-solutions etc). The example shown in Figure 2 is the portal "forskning.no", which present new ideas in science for young adults.

Technically speaking a topic map is expressed as a document type in SGML or XML (XML topic maps, 2001). Today most topic maps are in the XML-format XTM (XML Topic Maps). Below I present some excerpts from an XTM-file which maps my relationships to my employer, which is Oslo University College. In theory a topic map can be made for hand and presented using XSLT, as I did for this example. In the real world topic maps are typically presented as HTML generated by a topic map engine. The XTM-file would be made using some kind of editor coupled with a database containing the ontology, all of which may be part of the engine.

Example 1 shows the topic representing me – Nils – including related topic types, “nils” is an instance of both of the topic types “employee” and “teacher”. Also it contains an occurrence, in the form of an internal resource (using the <resourceData>-element).

Example 1 Topic map entry for topic "nils"

```
<topic id="nils">
 <instanceOf>
  <topicRef xlink:href="#employee"/>
 </instanceOf>
  <instanceOf>
   <topicRef xlink:href="#teacher"/>
 </instanceOf>
 <baseName>
  <baseNameString>Nils Pharo</baseNameString>
 </baseName>
 <occurrence>
  <instanceOf>
   <topicRef xlink:href="#description"/>
  </instanceOf>
  <resourceData>Nils has worked at Oslo UC since 1997</resourceData>
 </occurrence>
</topic>
```

In Example 2, below, you will se the topic implementation of Oslo University College (“ouc”). This also has an occurrence, this time being external and of the type “website”, and addressed using the <resourceRef>-element. In addition we here see an example of the use of

Example 2 Topic map entry for topic "Oslo University College"

```
<topic id="ouc">
 <instanceOf>
  <topicRef xlink:href="#institution"/>
 </instanceOf>
 <subjectIdentity>
  <subjectIndicatorRef xlink:href="http://home.hio.no/~nilsp/psi/hio.psi"/>
 </subjectIdentity>
 <baseName>
  <baseNameString>Oslo University College</baseNameString>
 </baseName>
 <occurrence>
  <instanceOf>
   <topicRef xlink:href="#website"/>
  </instanceOf>
  <resourceRef xlink:href="http://www.hio.no/"/>
 </occurrence>
</topic>
```

subject indicators (using the <subjectIdentiy>-element), which is a very central feature of topic maps and which will be discussed later.

The third example (Example 3) exemplifies how a relationship or association is defined, and also that associations themselves can be typed. We see that the association “nils-ouc-association” is an instance of the topic type “employment” with the topic “nils” and “ouc” as

members playing different roles in the association, the roles are specified using the topic types employee and employer respectively.

```
<association id=" nils-ouc-association">
 <instanceOf>
  <topicRef xlink:href="#employment"/>
 </instanceOf>
 <member>
  <roleSpec><topicRef xlink:href="#employee"/></roleSpec>
  <topicRef xlink:href="#nils"/>
 </member>
 <member>
  <roleSpec><topicRef xlink:href="#employer"/></roleSpec>
  <topicRef xlink:href="#hio"/>
 </member>
</association>
```

Example 3 Topic map entry for association between "nils" and "Oslo University College"

These examples emphasise the most important functions of a topic map:

1. topics have names
2. topics are knit together using associations
3. a topic may be categorised by an unlimited number of topic types
4. topic types are topics on a higher level of abstraction
5. association types are association on a higher level of abstraction
6. occurrences may be external or internal to the topic map
7. topics can be disambiguated using subject indicators

In Section 3 we shall see how these features of topic maps can be used to structure a web site.

2. Knowledge organisation of Web sites using topic maps

The example used in this section is collected from the Web site “forskning.no” which is aimed at presenting new research for “young adults” in newspaper format. The site is in Norwegian[1], but necessary translations are performed for explaining the content.

First, a topic map needs content in the form of topics and associations characterising some phenomenon to have any meaning (i.e. the ontology). In the case of “forskning.no” the

[1] This illustrates one possible weakness of topic maps; the lack of international application

ontology represents “science”, which includes the natural sciences, social sciences, and the humanities.

Figure 2 shows the homepage of the Web site. Note that the web site’s content is presented in HTML created by input from their topic map engine (called ZTM- Zope Topic Maps), the topic map can be imported and exported in XTM-format.

On the left hand side we see a navigation bar leading to the major categories (e.g. kultur (culture) and samfunn (society)) of the web site as well as the ontology. The main part of the homepage contains titles and ingresses to new articles on different kinds of subjects (or topics). These articles are examples of occurrences.

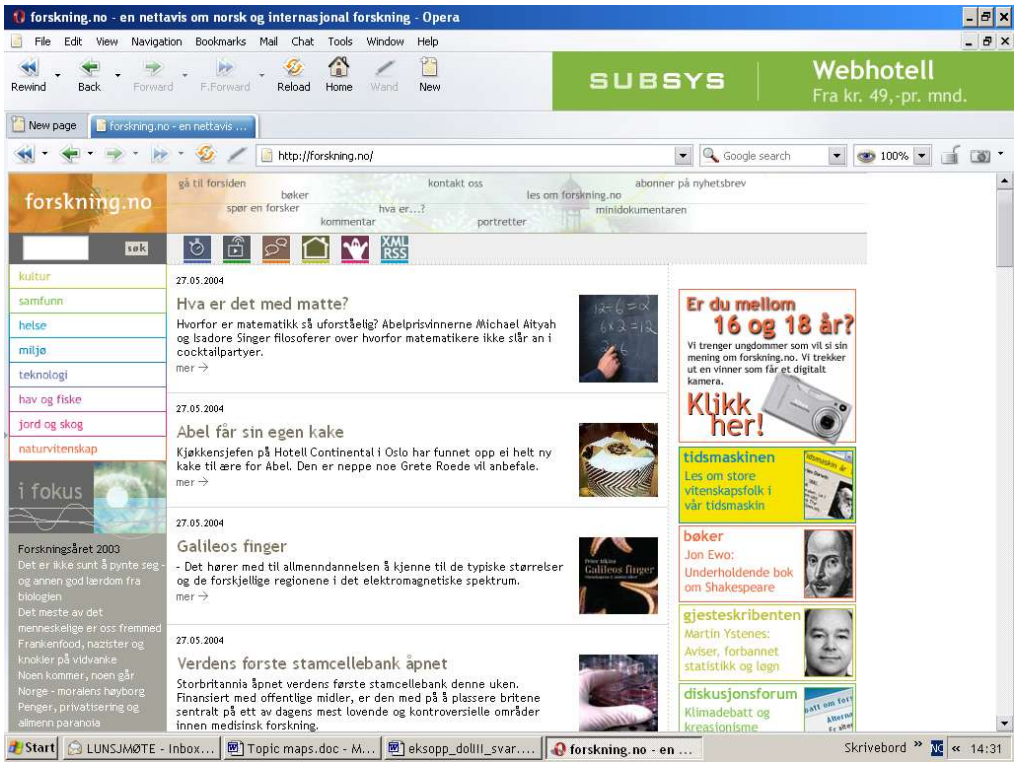

Figure 2 The home page of the web site forskning.no

In Figure 3 we show the page representing “teknologi” (technology), which is one of the major categories (i.e. topics) of the web site.

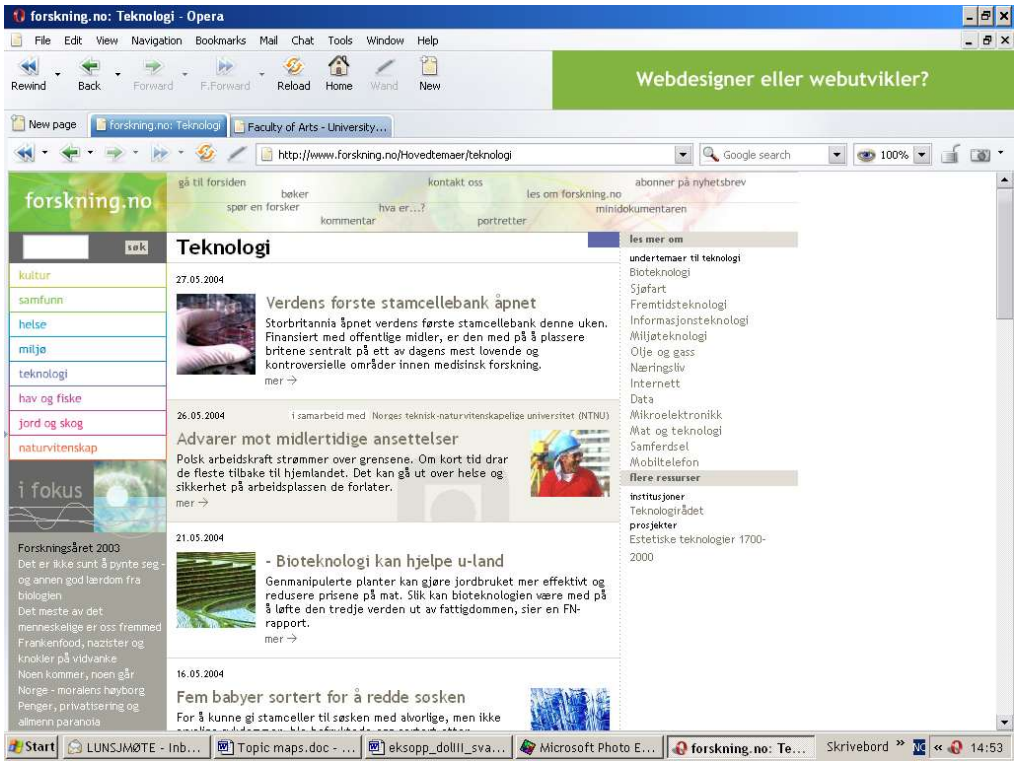

Figure 3 The main page of the topic "teknologi" (technology) on forskning.no

We see that the same major categories are represented in the left hand “global” navigation bar. In the main page there are references to articles (occurrences) that deal with different topics related to technology, we may thus say that technology also is a “*topic type*”. On the right hand side we see a list of subtopics to technology, e.g. “bioteknologi” (bio-technology), “sjøfart” (maritime science), and fremtidsteknologi (future technology).

The third forskning.no-page (Figure 4) reproduces (part of) an article the world’s first stem cell bank. The article represents an *occurrence* of something related to stem cell research, but also to topics such as “fremtidens medisin” (future medicine), “forskningspolitikk” (research politics), “etikk” (ethics), bioteknologi (bio-technology), and genteknologi (gene technology), which are all listed on the left hand side navigation bar as “temaer” (themes).

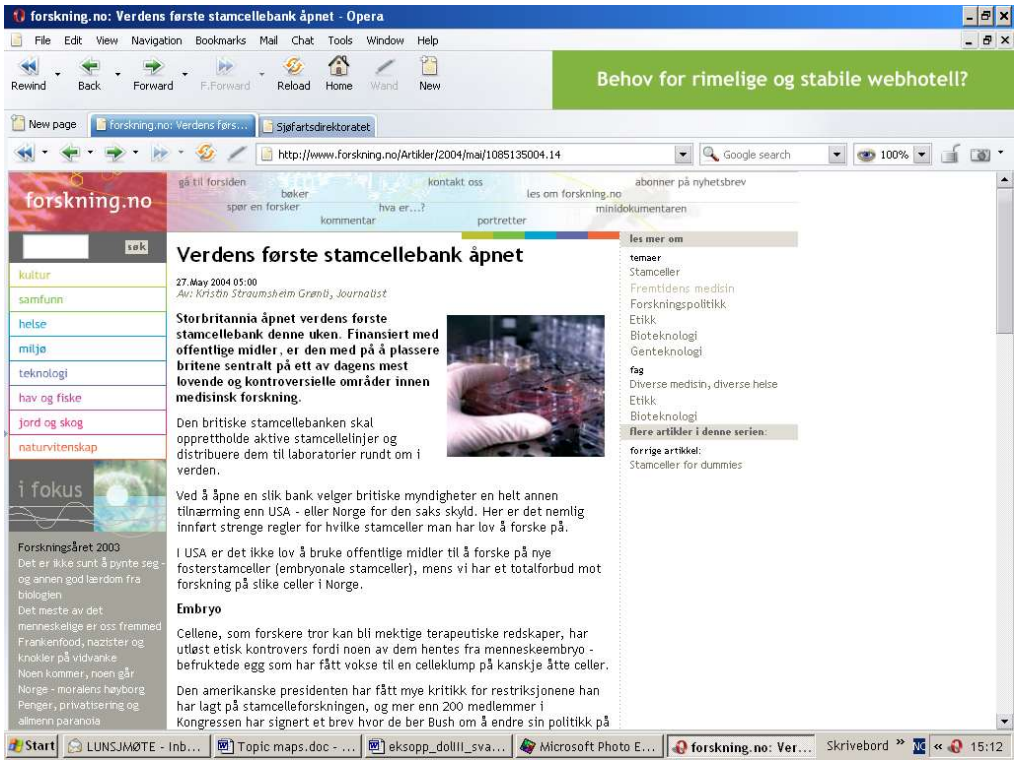

Figure 4 Article on forskning.no concerning stem cell research

Of course it is not necessary to use topic maps to create this kind of site, in fact there are probably thousands of examples of this kind of sites that do not use topic maps. Topic maps, however, is very powerful in the sense that it provides the administrator of the web site with an organisational tool to keep track of topics, associations and occurrences, i.e. the ontology and pages representing the topics. In the case of forskning.no most occurrences are created by the organisation, but there are also links to outside resources. In the latter case control, of course, is much more limited. In the next section we shall look at some of the challenges using topic maps.

Topic maps – from a knowledge organisational perspective

The idea of topic maps is that it should bring the documents to the subject descriptions (topics) rather than embedding the subject description in the documents, which is typically the case in metadata initiatives like RDF and different ways of implementing Dublin Core. Garshol (2004) emphasises that topic maps differ from traditional library and information science (LIS)-initiated systems for knowledge organisation in the sense that "topic maps are not so much an extension of the traditional systems as on a higher level. That is, thesauri extend taxonomies, by adding more built-in relationships and properties. Topic maps do not

add to a fixed vocabulary, but provide a more flexible model with an open vocabulary." He concludes, and I agree with him, that topic maps can be used as a technology to represent taxonomies, thesauri, faceted classification systems etc.

In my opinion it is particularly interesting to analyse the idea of topic maps in light of thesaurus construction and facetted classification. The topic map-environment is particularly clear in stating that topic maps open up for non-hierarchical knowledge organisation systems. This, however, necessitates the construction of non-hierarchical ontologies. In practice most, if not all, web sites organised using topic maps will consists of one or more hierarchies with cross-hierarchy references. Rather than dismissing hierarchies as the structural format of the ontology it is more important to discuss how to implement the non-hierarchical references.

A thesaurus is an example of a hierarchical ontology that permits explicit references between items (terms) that are in a non-hierarchic (associative) relationship to one another. There are, however, following the ISO thesaurus standard (ISO 2788, 1986) strong recommendations advocating the kind of associations that should be realised in the thesaurus (e.g. the discipline or field of study and the objects or phenomena studied).

If a thesaurus-like ontology should be used for structuring web sites it is necessary to reconsider the strong constraints set by the thesaurus rules. On the other hand we should be very sceptical towards not having any rules for which associations to implement in a topic map. A large topic map may quickly become over-complex and unpredictable in use if there are no restrictions on the implementation of associative relations. This is one of the reasons that a constraint language (TMCL – Topic Map Constraint Language) current is being developed for defining schemas and constraints on topic map models.

A very interesting, and potentially powerful, idea is the implementation of *subject identity*. As pointed out by Pepper (2002) the subject (which is the topic map terminology's equivalent to a thesaurus' concept, that which the topic is about) may itself be an "addressable information resource [and] its identity may be established directly through its address" (URI). In most cases, however, the subject is not a digital resource; in such cases one may use Published Subject Indicators, or PSIs to define the identity of the resource.

A PSI is, according to a working group (OASIS, 2003) a digital resource providing

> "some kind of compelling and unambiguous *indication* of the identity of a subject to humans. It may be a textual definition, description or name; it may be a visual, audio or other representation of the subject; or it may be some combination of these. A subject indicator is distinct from the subject that it indicates".

The subject indicator is addressed by a subject *identifier*. There are no formal rules as to the format of PSIs, the OASIS working group states that it should be human-interpretable and contain metadata about itself. The working group invites everyone to start making PSIs.

It is not necessary to use PSIs for defining topics, but for the successful merging of two or more topic maps they are particularly useful. The idea of merging topic maps is very intriguing. It makes it possible to combine topic maps which cover similar subject areas from different perspectives. In such cases PSIs are used to state whether two topics treat the same subject, in turn making it possible to create a topic with facets from both topic maps.

One particular challenge in developing PSIs is the question of authority; what should be the preferred PSI for each topic. Any subject that sparks controversy, like e.g. racism, abortion, positivism etc, would doubtlessly generate numerous PSIs.

This would, however, be very much in the nature of the Web. Its traditionally anarchistic structure and lack of centralised control is what has led to the Web's very heterogeneous selection of resources. Also, it is very much in line of Berners-Lee's thoughts about the Semantic Web (Berners-Lee, Hendler & Lassila, 2001) stating that the challenge of the Semantic Web is "to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web."

Pepper, at a talk on a seminar on Web portals in Oslo, argued, along similar lines, that a Darwinian evolution would, eventually, yield the "best" definitions in the form of PSIs – given enough time.

This discussion has emphasised an important distinction between computer science (CS) and library and information science (LIS) in their view on knowledge organisation. When dealing

with knowledge organisation LIS people tend to be dogmatic in contrast to the pragmatism of CS people. This is firmly grounded in historical reasons, knowledge organisation has been handled by LIS for centuries. That, however, should not hinder us in trying out "fresh" ideas related to our field from other disciplines. The topic map technology presents an opportunity to implement well-known knowledge organisational principles and to try out principles that are less hierarchical.

A tail on bibliographic records and topic maps

In this paper I have discussed topic maps as a tool for structuring Web sites. Here some thoughts will follow concerning how one might apply topic maps on a bibliographic database.

Given that the documents described by the bibliographic records are indexed by terms from a controlled vocabulary, e.g. a thesaurus, it would be reasonable to implement that thesaurus as a topic map. Each main term represents a topic and the hierarchical and associative relationships in the thesaurus are treated as associations. In addition associations could be made between topics that are used as index terms for the same documents. These associations might demand certain strength in order to be generated, e.g. 3 documents should be indexed by the same terms in order to activate an association between two topics.

Given such a scenario we give the searchers the opportunity to "surf" the topic layer which would consists of internal occurrences of bibliographic records or pointers to such records. But in addition a query to the database could use the information inherent in the topic map to rank documents based on their associative as well as hierarchic relationships. These relationships could be treated as paths with lengths. E.g, a query using terms representing topics X and Y could make the system rank documents in the following order:

1. documents on topics X and Y
2. documents on topic X
3. documents on topic Y
4. documents on topics with hierarchical paths of length 1 from X and Y
5. documents on topics with associative paths of length 1 from X and Y
6. documents on topics with hierarchical paths of length 1 from X
7. etc.

The relationships might of course also generate other kinds of interesting data to use by ranking algorithms, e.g. usage statistics on path use etc.

References

Berners-Lee T., Hendler J., & Lassila O. (2001). The Semantic Web. *Scientific American*, 284 (5), pp. 34-43.

Garshol, L.M. (2003). *Living with topic maps and RDF* [electronic version]. Retrieved June 2, 2004, from http://www.ontopia.net/topicmaps/materials/tmrdf.html

Garshol, L.M. (2004). *Metadata? Thesauri? Taxonomies? Topic maps! : making sense of it all* [electronic version]. Retrieved June 2, 2004, from http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html

Gilchrist, Alan 2003: Thesauri, taxonomies and ontologies : an etymological note. *Journal of documentation*. 59(1), 7-18.

ISO 2788 (1986). *ISO/TC46 Documentation: guidelines for the establishment and development of monolingual thesauri*. 2nd ed. Geneve: ISO.

ISO 13250 (2002). *ISO/IEC 13250 Topic Maps: information technology document description and processing languages* [electronic version]. Retrieved June 2, 2004, from http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

OASIS (2003). Published *subjects: introduction and basic requirements* [electronic version]. Retrieved June 2, 2004, from http://www.oasis-open.org/committees/download.php/3050/pubsubj-pt1-1.02-cs.pdf

Pepper, S. (2002). *The TAO of topic maps : finding the way on the age of infoglut* [electronic version]. Retrieved June 2, 2004, from http://www.ontopia.net/topicmaps/materials/tao.html

XML Topic Maps (XTM) 1.0 [electronic version] (2001). Retrieved June 2, 2004, from http://www.topicmaps.org/xtm/1.0/