Topic Maps

Steve Pepper
Department of Linguistics, University of Oslo, Norway

Abstract
Topic Maps is an international standard technology for describing knowledge structures and using them to improve the findability of information. It is based on a formal model that subsumes those of traditional finding aids such as indexes, glossaries, and thesauri, and extends them to cater for the additional complexities of digital information. Topic Maps is increasingly used in enterprise information integration, knowledge management, e-learning, and digital libraries, and as the foundation for Web-based information delivery solutions. This entry provides a comprehensive treatment of the core concepts, as well as describing the background and current status of the standard and its relationship to traditional knowledge organization techniques.

INTRODUCTION

Topic Maps is a standard technology for describing knowledge structures and using them to improve the findability of information. It is based on a formal model that subsumes those of traditional finding aids, such as indexes, glossaries, and thesauri, and caters for the additional complexities of digital information. The model is defined in an ISO standard (ISO 13250), along with interchange syntaxes, a formal semantics, and a graphical notation. Ancillary standards define a Topic Maps Query Language (TMQL), a Topic Maps Constraint Language (TMCL), and mappings to other knowledge organization specifications, such as Dublin Core. Since its initial standardization in 2000, Topic Maps is finding increasing application as the foundation for Web sites and portals, as well as in knowledge management (KM), e-learning, and more.

This entry provides a comprehensive treatment of the core concepts, in addition to the background and current status of the standard, its relationship to traditional knowledge organization techniques, and examples of the kinds of applications for which it is being used.[1]

BACKGROUND AND HISTORY

Origin and Development

The concepts of Topic Maps originated in the context of the Davenport Group, during the development of the DocBook application, as an answer to the problem of how to automate the merging of (digital) back-of-book indexes. The key insight, due to Dr. Steven R. Newcomb, was that such indexes are in fact models of knowledge. Their digital encoding is based on their surface appearance, but they actually have an implicit, underlying structure. If that structure could be represented formally, it would be possible to automate the merging process.

An initial model was defined in 1992 using the concepts of the HyTime standard, of which Dr. Newcomb was coeditor, under the auspices of a project called the Conventions for the Application of HyTime (CApH), sponsored by the Graphics Communications Association Research Institute (GCARI). After a number of iterations, during which the model was gradually refined, the work was brought into ISO,[2] and went through further revision cycles before being adopted as ISO 13250:2000 under the editorship of Newcomb and coeditors Dr. Michel Biezunski and Martin Bryan.

The standard was originally defined in terms of an SGML DTD[3] but by the time it was published Extensible Markup Language (XML) was replacing SGML, especially on the Web. In order to create an XML-based version of the specification an ad hoc working group called TopicMaps.Org was formed by Newcomb and Biezunski and this resulted in the publication in March 2001 of *XML Topic Maps (XTM) 1.0*,[1] edited by Steve Pepper and Graham Moore.

In addition to defining a new, XML-compatible DTD for representing topic maps, the new specification removed the dependence on HyTime, clarified some of the

[1]This entry follows the established convention of using initial capitals ("Topic Maps") when referring to the standard itself or the technology in general, and lower case ("topic maps") when referring to the document-like artifacts created through the application of that technology.

[2]Responsibility was given to Working Group 3 of the subcommittee ISO/IEC JTC 1/SC 34.

[3]In fact, it was defined in terms of a "meta-DTD" according to the SGML Architecture facility offered by the draft HyTime standard. Although never finally standardized, the concepts inherent in meta-DTDs found their way into DITA, XML namespaces, and HTML "microformats."

Encyclopedia of Library and Information Sciences, Third Edition DOI: 10.1081/E-ELIS3-120044331

terminology, and simplified parts of the model. XTM gained immediate recognition and has to all intents and purposes replaced the original version. In 2003, the XTM DTD was folded back into the second edition of the ISO standard[2] and from that point on stewardship of the standard passed back to the ISO committee.

Current Status

In 2003 a road map was devised for the further development of the standard, to include a data model, reference model, query language, and constraint language. As of the present writing (late 2008), the Topic Maps family of ISO specifications is as follows[4]:

- ISO/IEC 13250: Information Technology—Topic Maps—
  - — Part 1: Overview and basic concepts[3]
  - — Part 2: Data model[4]
  - — Part 3: XML syntax[5]
  - — Part 4: Canonicalization[6]
  - — Part 5: Reference model[7]
  - — Part 6: Compact syntax[8]
  - — Part 7: Graphical notation[9]
- ISO/IEC 18048: Information Technology—TQML[10]
- ISO/IEC 19076: Information Technology—TMCL[11]
- ISO/IEC TR 29111: Information Technology—Topic Maps—Expressing Dublin Core Metadata using Topic Maps[12]

Work continues in SC 34/WG 3 and interested parties are encouraged to attend meetings and participate through their national standards bodies.[5]

CORE CONCEPTS

This section presents the core concepts of the Topic Maps paradigm. The back-of-book index is used throughout as a familiar point of reference for illustrative purposes. In line with long-established tradition in the Topic Maps community, this entry will use the domain of opera for its examples, based on the *Italian Opera Topic Map*.[13]

Subjects

The core concepts of Topic Maps are relatively few and for the most part easily grasped. They are often referred to as the "TAO of Topic Maps," after the eponymous paper which has served as a basic introduction since it was first published in 2000.[14] However, even more fundamental than the TAO is the emphasis on *subjects*.

In the Topic Maps world view, the most essential property of information is not where it resides (the document-centric view) or which application was used to create it (the application-centric view), but its *aboutness*, i.e., the subject (or subjects) that it is about. This subject-centric view lies at the very heart of the Topic Maps paradigm.

There are many reasons for insisting on the primacy of subjects: the starting point for most acts of information retrieval (e.g., searches on the Web) is one or more subjects; in the aggregation of information and knowledge, subjects usually constitute the most useful collation points; in fact, the whole purpose of creating information can be viewed as capturing knowledge *about certain subjects* in order that it can be shared and reused. While other forms of metadata such as author, publisher, and creation date are important for certain information management tasks, for the key end-user task of information retrieval, aboutness is most critical. Since the purpose of Topic Maps is to alleviate infoglut and improve the findability of information, subject-centricity is the central feature.

Subjects and topics

The concept of "subject" is defined in Ref. [4] as follows:

> 3.14 **subject**
>
> anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever

A subject, in other words, is literally anything about which the author of a topic map wishes to make assertions. Typical subjects in the domain of opera might be the composer Giacomo Puccini; his operas, including *Tosca* and *Madama Butterfly*; Lucca, the city where Puccini was born; and Teatro Costanzi, where *Tosca* was first performed. These are also the kind of subjects one would expect to find in the index of a book on opera.

The Topic Maps data model (TMDM)[4] defines a model for representing subjects such as these, and assertions about them, within an information system.[6] In order to represent a subject a proxy is required, and this is called a "topic." A topic is thus the (symbolic) representation of a (nonsymbolic) referent, i.e., the subject. In the words of the standard:

> 3.18 **topic**
>
> symbol used within a topic map to represent one, and only one, subject, in order to allow statements to be made about the subject

[4]Parts 2 and 3 were published in 2007. Most of the remainder exist as stable drafts and are expected to be published in 2009.

[5]For further information, see the SC 34 Web site at http://www.jtc1sc34.org and the WG 3 Web site at http://www.isotopicmaps.org.

[6]The Topic Maps Reference Model (TMRM),[7] which offers a more abstract and low-level model, is described below.

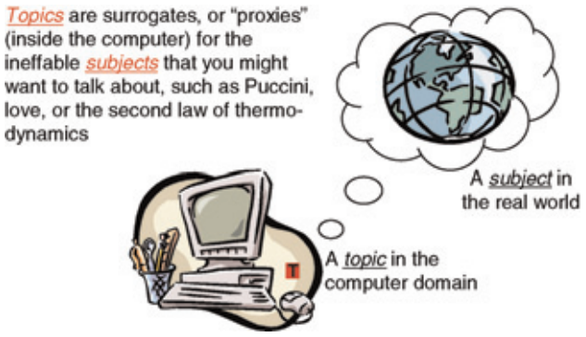

Fig. 1 Subjects and topics.

A topic map covering the domain of opera will thus include topics that represent subjects such as Puccini, *Tosca*, and Lucca.

The collocation objective

Topic Maps employs the nearly synonymous words "topic" and "subject" as very precise terms for the two components of a representation relationship: symbol and referent (Fig. 1).

There is a one-to-one relationship between topics and subjects in the sense that each topic, by definition, represents one and only one subject.[7] The goal of any Topic Maps application is to ensure that each subject is represented by one, and only one, topic. Once this goal (referred to as the "collocation objective") is achieved, everything that is known (within a given system) about a particular subject becomes accessible from a single point, via the one and only topic representing that subject, and thus the problem of information retrieval is in principle solved.

The means by which the collocation objective is achieved within a system relies heavily on the Topic Maps concept of "identity," which can also be applied across systems, and thus provides the foundation for "global knowledge interchange"—the potentially unrestricted collation of knowledge from disparate sources.[8]

Identity

The fundamental issue in achieving the collocation objective is to know when two topics represent the same subject, a situation that can often arise when merging topic maps. When this is the case, the two topics in question must be merged into a single topic that has the union of the characteristics (i.e., names, occurrences, and associations) of the original topics. Once the two topics have been merged, order will have been restored and collocation will once more obtain.

Since the merging of topic maps is intended to be an automated process, the key question is, how can a machine "know" when two topics represent the same subject? The solution adopted in the TMDM is to use explicit, globally unique identifiers. If two topics share an identifier they are deemed to represent the same subject and merging occurs. Identifiers usually take the form of Uniform Resource Identifiers (URIs)[9] and these come in two flavors: "subject identifiers" and "subject locators."[10]

Subject locators, as their name suggests, are URIs that identify subjects via their location. They can therefore only be used with subjects that have a specific location that can be expressed using a URI—in other words, network-addressable information resources, such as documents (in the broadest sense), newsfeeds, and other data that can be retrieved via Web-based queries. Subjects like this can be identified *directly* via their network addresses.

Most subjects, like Puccini, *Tosca*, and Lucca, do not fall into this category. They reside outside any computer system and are thus in some sense "ineffable"; there is a chasm between the "outside world" in which they exist and the computer system in which they are represented, and this chasm can ultimately only be bridged by human intellect. This is achieved in the Topic Maps paradigm through *indirect* identification using subject identifiers.

A subject identifier is (also) a URI, but it does not address the subject directly; rather it addresses it indirectly, via a "subject descriptor,"[11] which the TMDM defines as an "information resource that is referred to from a topic map in an attempt to unambiguously identify the subject represented by a topic to a human being."

Fig. 2 shows how an ineffable (or "non-addressable") subject, such as Puccini, can be identified indirectly via a subject descriptor, whose address (http://psi.ontopedia.net/Puccini) functions as a subject identifier.

The distinction between direct and indirect identification, which is supported in both the TMDM and the interchanges syntaxes, provides a solution to what has been widely referred to as the "identity crisis" that has caused concern in the Semantic Web community over a number of years.[16] At the same time, the duality of subject identifier and subject descriptor neatly reflects the

[7]Because of this, the terms "topic" and "subject" are often used interchangeably in informal discourse about Topic Maps.

[8]The term "global knowledge interchange" is due to Steve Newcomb.[15]

[9]The TMDM actually allows any kind of "locator notation" to be used, but the interchange syntax (XTM) only supports Internationalized Resource Identifiers (IRIs), an extended form of URIs which permits international characters.

[10]A third kind of identifier, item identifiers, are only of interest to implementors of Topic Maps systems. Subject locators were called subject addresses in the first version of the standard.

[11]Subject descriptor is sometimes also termed "subject indicator."

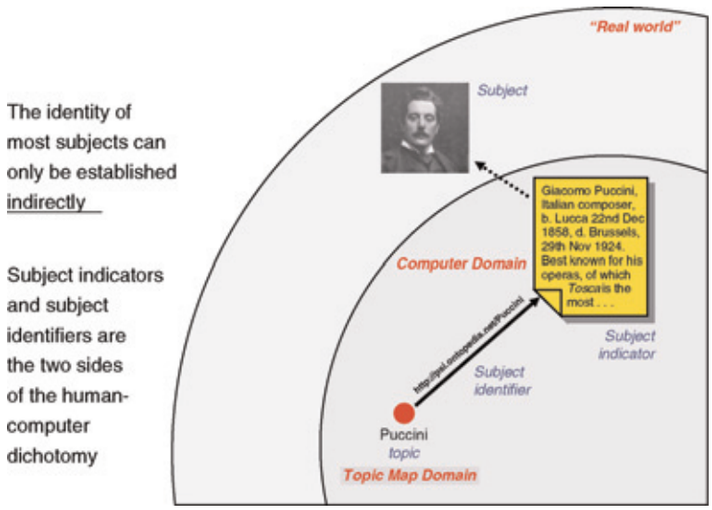

Fig. 2 Subject identifiers.

dichotomy of the human/computer relationship. Subject identifiers are for use by computers in ascertaining whether or not two topics represent the same subject: if two topics have an identifier in common, they are simply merged; computers have no need to resolve the URI to ascertain exactly what the subject is. Humans, on the other hand, need to be able to do this, because it is they who are ultimately responsible for assigning identifiers; the subject descriptor provides them with an "indication" (through a description, definition, illustration, or whatever) of the subject in question.

The concepts of indirect identification using subject identifiers and subject descriptors has been extended into a whole paradigm, known as "published subjects" that seeks to provide an open, distributed solution to the problem of defining and discovering globally unique subject identifiers.[17]

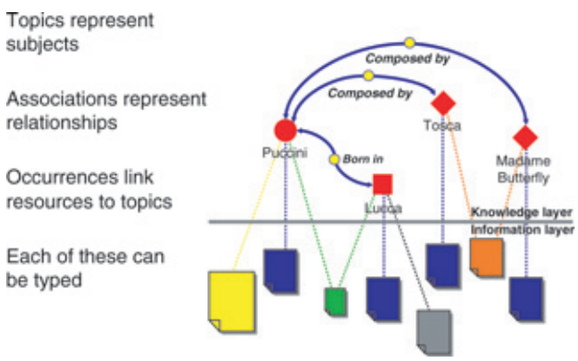

Fig. 3 The TAO model.

The TAO of Topic Maps

Having covered the essentials of subjects and how they are represented by topics, we now turn to assertions about subjects.

A topic map is a representation of a set of assertions about one or more subjects. The TMDM defines three kinds of assertion—or "statement"—that can be made about a subject: "names," "occurrences," and "associations" (Fig. 3). Associations represent general relationships between subjects; occurrences represent a particular form of "aboutness" relationship in which the participants are an information resource and a subject that the resource is "about"; and names represent relationships between subjects and the labels used by humans to refer to them. The following section describes each of these in more detail.

Associations and roles

Associations express relationships among subjects, for example between Puccini and *Tosca*, or between Puccini and Lucca. They correspond to *See also* entries in a back-of-book index and serve a similar navigational purpose. But there is an important difference in that associations permit the nature of each relationship to be made explicit using "association types": thus, the relationship between Puccini and *Tosca* can be stated to be of type "composed by," whereas that between Puccini and Lucca is of type "born in." The ability to distinguish association types adds precision, provides more information to the user, and improves findability. Association types are themselves subjects and are therefore represented as topics.

A single association may involve any number of subjects. Binary associations (involving two subjects, as in the examples given so far) are by far the most common, and also the easiest to process; they correspond somewhat to transitive verbal constructs in natural language, such as "*Puccini* composed *Tosca*." Ternary associations (involving three subjects), while less common, can be useful for relationships that in English might be expressed using a ditransitive verb or an additional thematic role ("*Tosca* kills *Scarpia* with a *knife*"). Associations of higher arity are infrequent; the relationships they might be used to represent (such as that between an opera and multiple librettists) are usually better expressed using multiple binary associations. Unary associations (involving a single subject) are also possible; they often correspond to intransitive verbs and can be used to represent "relationships" that in other modeling paradigms might be expressed using Boolean properties (e.g., "*Turandot* was unfinished").

Describing a relationship in natural language can give a false impression of directionality: there is no difference, in terms of the basic relationship being described, between the statements "*Tosca* was composed by Puccini" and "Puccini composed *Tosca*."[12] This situation is reflected in Topic Maps by the absence of any formal notion of direction in associations. The order in which topics are specified has no significance; instead, the nature of the subject's involvement in the relationship is expressed through the concept of "association role." Each participant in an association is deemed to play a "role" of a certain type. In order to clarify the respective roles of the participants in the example above (and avoid the possible interpretation of Puccini having been composed by *Tosca* instead of vice versa), it must be explicitly stated that Puccini plays the role of (say) "composer," and that *Tosca* plays the role of (say) "work." In more precise Topic Maps terminology, "composer" and "work" are "association role *types*" and, like association types, they are represented by topics (Fig. 4).[13]

Association types (and role types) are defined by the topic map author according to the requirements of the information, the knowledge it embodies, and the application for which it is intended. There is thus no limit to the kinds of relationship that may be expressed in Topic Maps. However, because of their ubiquity (and utility) in knowledge modeling, two association types are given special status. These are "type-instance" and "supertype–subtype" which carry particular semantics

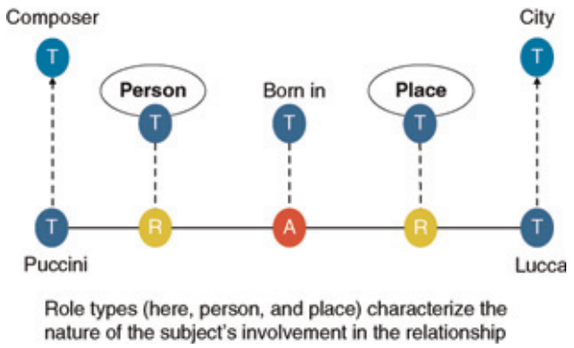

Fig. 4 Anatomy of an association.

defined in terms of membership of sets of topics called "topic types."[14]

Topic types constitute a built-in mechanism for classifying topics and for building a certain kind of hierarchy.[15] Topic types for the example topics mentioned so far might be "composer" (for Puccini), "opera" (*Tosca*), and "city" (Lucca). These might in turn be defined as subtypes of "person," "musical work," and "place," respectively.

It is important not to conflate the concepts of topic type and association role type. The former classifies a subject in terms of its "essential being," whereas the latter merely describes the role it plays in a particular relationship. Sometimes, however, the same topic is used as both a topic type and an association role type. This is because the essential being of a subject is often defined in terms of some kind of relationship: it is precisely because Puccini composed *Tosca* (and many other operas) that he is today regarded first and foremost as a composer; he both *plays the role of* "composer" (in the relationship with *Tosca*, etc.) and at the same time *is a* "composer." On the other hand, the roles he plays in his relationships with his teacher Ponchielli, his wife Elvira, and his birthplace Lucca, are of different types, namely "pupil," "husband," and "person" (Fig. 5).[16]

[12]The difference is rather one of *focus*.

[13]Exactly what constitutes the roles in this relationship (as opposed to the role types) is less immediately intuitive; one approach is to think in terms of Puccini *qua* composer of *Tosca* (as opposed to Puccini *qua* composer of *Madama Butterfly*—or indeed *qua* husband of Elvira or *qua* pupil of Ponchielli).

[14]Their special status consists in the following: 1) the standard defines identifiers for these association types (and their corresponding role types) in the TMDM; 2) special syntax is available for type-instance in XTM and Compact Topic Maps Syntax (CTM); and 3) they invoke the application of inheritance mechanisms in certain forms of processing in the query and constraint languages.

[15]That is, type hierarchies—*not* subject hierarchies of the kind used in subject classification systems, where the subjects do not necessarily (or usually) constitute sets or classes, and the transitivity of type hierarchies does not hold. These can be represented in Topic Maps (as discussed below) but not by using the predefined association types.

[16]Note that there is almost always an implicit supertype–subtype relationship between the role type and the type of the role playing topic, but there is no general rule for which is which.

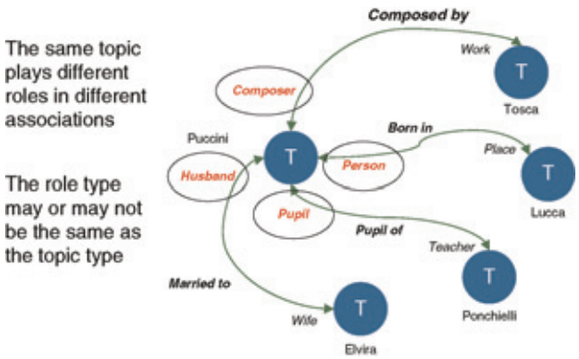

Fig. 5 Role type and topic type.

Occurrences

Occurrences represent a particular kind of relationship—that of "aboutness" between an information resource and a subject. The resource in question may be very small, such as a string representing a date, in which case it is normally included in the topic map and known as an "internal occurrence." Or else it may be stored externally, because of its size, notation, provenance, or whatever, and referenced via a locator—normally a URL—which corresponds to a page number in a back-of-index (itself a locator for some piece of information relevant to the subject in question). In the latest version of the Topic Maps standard, support is included for specifying the data type (i.e., notation or format) of an occurrence.

Like associations (and association roles), occurrences may be classified by type; typical "occurrence types" in the domain of opera might be "biography" and "date of birth" (for topics of type "composer"); "synopsis" and "premiere date" (for topics of type "opera"); and "map" (for topics of type "city"). Occurrence types provide more information to the user about the nature of the related resource (in contrast to an unqualified list of page numbers in a typical back-of-book index).

Finally, occurrence types—like association types and role types—may be freely defined (again, as topics) to suit the needs of users and applications; there are no predefined types.

Names

Because of their importance in providing labels by means of which humans can refer to and discourse about subjects, names are given privileged status in the Topic Maps model. They may be regarded as a special type of occurrence that carries naming semantics. Like occurrences, names may be typed (and once again, name types are represented as topics), but unlike occurrences they are restricted to a single data type, "string," in order to ensure that they can always be processed by any system.

It is also possible to specify "variants" of a name, and this facility is widely used to represent sort keys, alternative orthographies, transliterations, misspellings, and so on.

In order to cater for synonyms, the model allows multiple names for the same subject. Distinctions such as language, preferred status, and context of use may be indicated using name types or the concept of "scope" (discussed below).

There is no specific facility in the model for handling homonyms because there is no need: there are no restrictions on what names a subject can have, nor on whether two (or more) subjects have the same name, so homonymy is not an issue as far as the model is concerned. It is left to the application to provide disambiguation based on contextual information available in the topic map: the first line of disambiguation is usually the topic type ("Tosca the opera" as opposed to "Tosca the character"); when this does not suffice, some other association can be used, as with "Puccini's *Bohème*" (i.e., the *La Bohème composed by* Puccini) as opposed to "Catalani's *Bohème*," or "the Paris *located in* France" as opposed to "the Paris *located in* Texas."

Fig. 3 illustrates the basic TAO model and depicts an information layer (below), containing information resources of any and every shape and form, and a knowledge layer (above). The latter consists of (typed) topics and associations, and is connected to the information layer via (typed) occurrences. Identity, scope, and reification (described below) are not depicted.

Ontologies

The term "typing topic" is used informally to refer to any topic that is used (or intended to be used) to type some other construct, whether it be a topic, association, association role, occurrence, or name. Taken together, the set of typing topics in a topic map constitutes a description of the kind of "things" (i.e., subjects) and the kind of relationships that exist in the domain in question (e.g., opera). Viewed in this way, the typing topics can be said to represent the "ontology" of the topic map and the topic map to contain its own ontology.[17]

Constraints on the types (or classes) that constitute an ontology are defined using TMCL, discussed below.

Additional Concepts

The basic model of (typed) topics that can have names and occurrences and play roles in associations with other topics is very simple and yet powerful enough to express

[17]There is no single, broadly accepted definition of the term "ontology" in the information sciences, nor is it defined in the Topic Maps standard, but the informal usage adopted here corresponds very closely to that of John Sowa who defines ontology as a "classification of the types and subtypes of concepts and relations necessary to describe everything in the application domain."[18]

a broad range of knowledge structures—in fact, to make any kind of assertion about any kind of subject—and many current applications of Topic Maps use no more than this core functionality. Additional concepts that provide further expressivity are "scope" (for qualifying assertions in terms of their contextual validity) and "reification" (for making assertions about assertions).

Scope

Knowledge that is aggregated from different sources through the merging of topic maps is likely to contain contradictory statements or have varying relevance in different contexts. For this reason, Topic Maps includes a built-in mechanism called "scope" for expressing the contextual validity of any assertion. Scope is expressed as a set of topics which qualify a statement (i.e., a name, occurrence, or association) and indicate the context in which the assertion represented by the statement may be considered valid. If no scope is explicitly specified, the scope is said to be "unconstrained."

The interpretation of what it means to be "valid," and precisely how the scoping topics "indicate" context, is left to the application and as a result the interoperability of scope is limited. However, usage conventions emerging among users of Topic Maps are resulting in increased interoperability. These include specifying the historical applicability or natural language of names and expressing the provenance of occurrences and associations.

Examples of the use of scope in the *Italian Opera Topic Map* include:

- English names of certain operas.
- Provenance of conflicting assertions (e.g., whether the opera *Isabeau* was first performed at the Teatro Colon or the Teatro Coliseo).
- Source of multiple synopses of the same opera (e.g., from Arizona Opera and Opera News).
- Context in which a character has a certain voice type (e.g., Musetta, a soprano in Puccini's *La Bohème* and a mezzo in Leoncavallo's).

Reification

Scope can be regarded as a special case (asserting the *contextual validity* of an assertion) of a more general capability (asserting *anything* about an assertion). Assertions represent relationships between subjects, but a relationship can be regarded as a subject in its own right, about which one might want to make further assertions. One way to do this is to represent the relationship as a topic from the outset; another is to "reify" the name, occurrence, or association in question.

Reification (literally "thingification") in the Topic Maps sense of the term is defined as "making a topic represent the subject of another topic map construct," in other words, turning a name, occurrence, association role (or even the topic map itself) into a topic in order to make assertions about the thing it represents.[18] In the *Italian Opera Topic Map* the relationship between *Tosca* and Rome (where the opera takes place), is reified as a "The setting of Tosca in Rome" in order to provide an appropriate subject for classifying Susan Nicassio's book *Tosca's Rome*.

The most widespread use of reification is to reify the topic map itself, thereby creating a topic that can be used for specifying metadata about the topic map (for example, using Dublin Core as discussed below).

THE TOPIC MAPS FAMILY OF STANDARDS

The model described in the preceding sections is specified in detail in the TMDM, which is Part 2 of the ISO standard.[4] This section provides a brief overview of the remaining parts (excluding Part 1, which is merely a nonnormative introduction), along with short descriptions of two related standards, ISO 18048 (TQML) and ISO 19756 (TMCL), and a technical report, ISO TR 29111 (Expressing Dublin Core Metadata in Topic Maps).

Syntaxes and Notations

Part 3 of ISO 13250[5] specifies an XML-based syntax called XTM whose purpose is to enable topic maps to be interchanged between systems. This is the core interchange syntax; it is defined in terms of a mapping to the TMDM and all conforming Topic Maps systems are expected to support it.[19] However, because it is rather verbose, XTM is not generally considered suitable for hand-editing, and for that reason other, nonstandard syntaxes were devised (Fig. 6). The most widely used of these are LTM[19] and AsTMa,[20] both of which offer compact, text-based notations. A standard text-based notation called CTM[8] has been defined more recently; a visual notation called Graphical Topic Maps Notation (GTM)[9] provides a common way of visualizing topic maps and their ontologies, and a canonicalization syntax (CXTM)[6] supports conformance testing of Topic Maps systems.

The Reference Model

Part 5 of ISO 13250[7] defines a low-level model called the TMRM which is more abstract and has fewer ontological commitments than the TMDM. Its purpose is to serve as a minimal, conceptual foundation for subject-centric data models such as the TMDM, and to supply ontologically

[18]This is equivalent to nominalization in natural language, as in "Reagan met Gorbachev in Reykjavík. *The meeting* took place in October 1986."
[19]The original HyTime-based syntax (HyTM) is no longer part of the standard and most tools do not support it.

```
<topic id="la-boheme">
  <instanceOf><topicRef xlink:href="#opera"/></instanceOf>
  <baseName>
    <baseNameString>La Bohème</baseNameString>
    <variant>
      <parameters>
        <subjectIndicatorRef
          xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm#sort"/>
      </parameters>
      <variantName><resourceData>Boheme, La</resourceData></variantName>
    </variant>
  </baseName>
  <occurrence>
    <instanceOf><topicRef xlink:href="#homepage"/></instanceOf>
    <resourceRef
       xlink:href="http://www.opera.it/Opere/La-Boheme/La-Boheme.html"/>
  </occurrence>
  <occurrence>
    <instanceOf><topicRef xlink:href="#premiere-date"/></instanceOf>
    <resourceData>1896-02-01</resourceData>
  </occurrence>
</topic>
```

Fig. 6 XTM 1.0 syntax.

neutral terminology for their disclosure. It defines what is required to enable mapping between different subject-centric data models in order to meet the overall goal of the Topic Maps standards, that each subject has a single location for all the information about it.

The TMRM is defined in terms of proxies consisting of properties which themselves are key/value pairs; the only ontological commitments are two types of relationship: *sub* (subclass of) and *isa* (instance of). In addition to requirements on constraint languages and merging operations, the TMRM includes a path expression language and a mapping to the TMDM, and thereby provides a formal semantics for the latter. It also provides a formal foundation for the related standards TQML and TMCL.

The Query Language

Topic Maps Query Language[10] provides a standard way of accessing Topic Maps-based information, just as Structured Query Language (SQL) does for relational data and XQuery does for XML data. The initial version provides access functionality only; later versions are expected to also cover updates. The language currently offers three very powerful and isomorphic ways to express queries:

- SELECT expressions, with FROM, WHERE, ORDER BY, UNIQUE, OFFSET, and LIMIT clauses.
- FLWR expressions, with FOR, WHERE, ORDER BY, and RETURN clauses.
- Path expressions, similar in flavor to Xpath, which can be used alone or in combination with the foregoing.

All three types of expression can return results in tabular form; FLWR expressions can additionally generate XML and Topic Maps output.

The Constraint Language

Topic Maps Constraint Language[11] defines a language for expressing ontology-based constraints that a class of topic maps is expected to follow, thus ensuring greater consistency and more predictable results for both users and applications.

Typical constraints on the *Italian Opera Topic Map* might be as follows:

- All topics of type "composer" must play the role of "composer" in at least one "composed-by" association.
- Every "composed-by" relationship must involve exactly two topics of types "composer" and "opera," playing the roles "composer" and "work," respectively.

TMCL defines constraint types and an interpretation of instances of those types. The interpretation indicates what it means for an instance of a given constraint type to be evaluated in the context of a TMDM instance, with the result of the evaluation being either true or false. Constraint types are defined in terms of the TMDM and their formal interpretation in terms of TQML.

KNOWLEDGE ORGANIZATION

One of the most important aspects of Topic Maps is the ability of the model to represent virtually any kind of knowledge structure or data model. It is this capability—coupled with the ability to merge arbitrary topic maps—that underlies its value proposition: improved information management and enhanced findability through connecting disparate systems and collating information and knowledge from different sources.

Given that the original impetus for the development of the Topic Maps paradigm came from the requirement to be able to merge back-of-book indexes, it is only to be expected that the model should handle every feature of indexes. But the model was also constructed in such a way that it extends to every other known form of knowledge organization, including thesauri, bibliographic records, glossaries, and subject classification systems[21]—and can also subsume hierarchical, relational, and associative data models in general. This section explains how such structures are mapped to the Topic Maps model.

Indexes

Most back-of-book indexes consist of a set of *entries*, arranged alphabetically by *main heading* (sometimes called an "access point" or "subject heading") and containing zero or more *subentries* (or "subheadings"), *reference locators*, and *cross-references*.[22]

In Topic Maps terms, every *entry* (except those that only consist of a *See* cross-reference, discussed below) corresponds to a topic and its *main heading* corresponds to a topic name. *Subentries* are named topics that are related to the topic of the main entry by an association of type "subentry of" (with the role types "main entry" and "subentry"). *Reference locators* equate to the locators of (external) occurrences, which would normally take the form of URIs rather than page numbers. *See also* references correspond to generic associations (of type "see also," or simply "related to") between topics that represent the respective entries. Finally, *See* references indicate alternate labels for the topics represented by the referenced entries and thus correspond to additional names (of type "alternate name") for the topics representing those entries.

In some indexes the typeface used for main headings may vary (usually italic in addition to roman, sometimes also boldface). This is a simple form of classification by topic type. Similarly, the use of a different typeface for a reference locator (i.e., page number) is a simple form of classification by occurrence type. Finally, some books contain multiple indexes (e.g., people, places, subjects), which again corresponds to the use of topic types.

Glossaries

A glossary is a list of terms in a particular domain of knowledge along with their definitions. Like indexes, glossaries consist of a set of alphabetically arranged entries, each of which is the equivalent of a main heading. Subentries are not usual, but *See* and *See also* cross-references may occur. Most importantly, each glossary entry contains a definition instead of reference locators.

Using Topic Maps a glossary can be modeled exactly like an index, except that instead of multiple external occurrences, the topic corresponding to each glossary entry will have a single internal occurrence of type "definition." Thus multiple indexes and glossaries can all be represented within a single topic map.

Thesauri

A thesaurus consists of a set of *terms* and *scope notes*, organized through three kinds of semantic relationship: *equivalence*, *hierarchy*, and *association*. Terms are the names of concepts and thus correspond to topic names; scope notes consist of information pertinent to the topic and therefore correspond to occurrences of type "scope note."

The equivalence relationship handles synonyms and thus corresponds to topics with multiple names; a USE relation would indicate the default (preferred) topic name, while a USED FOR relation would indicate a name of type "non-preferred" or "alternate."

Hierarchy relationships are created by pointing to *broader terms* (BT) and *narrower terms* (NT). Since these are the inverse of each other, they correspond to associations of a single type ("broader/narrower") which link topics representing broader concepts (e.g., "opera") to topics representing narrower concepts (e.g., "aria").[20]

Finally, associative relationships (indicated using *related terms*, RT) equate to *See also* references in back-of-book indexes: they are (nonhierarchical) relationships of no specific type and thus correspond to associations of type "related to."

Since indexes, glossaries, and thesauri can all be represented by the same model, they can be combined in a single topic map, thus removing potential redundancy. In fact, if the topics that represent index and glossary entries are arranged hierarchically, using associations of type "broader/narrower," the result *is* a thesaurus—or possibly multiple thesauri, depending on the number of hierarchies involved. Individual documents or information resources are classified against such a thesaurus by making them occurrences of one or more thesaurus topics.

Taxonomies and subject classification systems are similar to thesauri; they are often somewhat simpler, and they may be more constrained (for example, some taxonomies allow only a strict-type hierarchy based on the generic supertype/subtype relationship), but their representation as topic maps and their practical application is essentially the same.

Faceted Classification

A faceted classification system can be viewed as a set of classification subjects that are grouped along axes called

[20]In some thesauri other, more semantically explicit kinds of hierarchical relationship are used (e.g., generic and partitive). These can be represented using more specific association types. Note that the BT/NT relationship, despite being hierarchical, is *not* the same as the supertype/subtype relationship defined in the Topic Maps standard.

facets. In a topic map this is represented as a set of topics, each of which is related via an association to a topic that represents a facet, and some of which may be organized hierarchically via additional associations.

Bibliographic Records

Bibliographic records are descriptions of information entities, which may be abstract (works) or concrete (documents).[23] They employ bibliographic languages to describe the attributes of those entities, ranging from the properties of *works* (e.g., author, title, subject), to the properties of *expressions* (translator, editor, language), *manifestations* (publisher, ISBN, format), and *items* (location, call number, condition). Attributes are described as property/value pairs in which the value can be a string, date, integer, or some other data type; or else the name of another entity (such as the author or publisher).

Bibliographic records based on languages of this kind can be represented as topic maps by creating topics to represent each information entity and using statements (names, occurrences, or associations, as appropriate) to represent the attributes. Those attributes with naming semantics (e.g., title) become topic names; attributes that name other entities (e.g., author, publisher), or that represent terms in an authority file (e.g., a predefined set of media types) become associations and give rise to topics that represent the entities or terms in question; other attributes are represented as either internal or (occasionally) external occurrences.

One widely used bibliographic language is the Dublin Core Metadata Set[24] which defines a basic set of 15 abstract "elements," along with a number of "other elements and element refinements," a set of encoding schemes, and a (media) type vocabulary. An ISO Technical Report[12] describes how to represent Dublin Core descriptions using Topic Maps.[21] A similar approach can be used with MARC-21.

In the FRBR model, relationships between works and expressions are represented by an association, as are relationships between expressions and manifestations, and between manifestations and items. Entities of all kinds (including groups 2 and 3) are represented as topics.[26]

Other Data Models

In addition to the knowledge models described above, most forms of structured data can be represented as topic maps through some kind of schema-dependent mapping. This section looks briefly at how this works with XML (hierarchical), RDBMS (relational), and Resource Description Framework (RDF) (associative) data, respectively.

[21]See also Ref. [25].

~~The~~ XML allows the structure of information to be represented in a hierarchy of elements and subelements, with some annotation of elements (via attributes), including linking across (and beyond) the element hierarchy. The semantics of the element and attribute types are vocabulary dependent, but they can be extracted and represented as a topic map using techniques described in Refs. [27,28].

In the entity–relation (ER) model, entities equate to topic types and relations to association types, while attributes correspond to names, occurrences, and identifiers (Fig. 7). When mapping relational data to a topic map, each row in an entity table (for example, "Organizations") gives rise to a topic of the corresponding type (in this case, "organization"); each column can result in an identifier, name, occurrence, or association, depending on whether the data is an ID, a string with naming semantics, some other kind of string data, or a foreign key reference, respectively. Each row in a relation table (for example, "Employment") results in an association of the given type (here, "employed by") whose role players are found via the foreign key references.

The associative (graph-based) model of the RDF is much closer to the Topic Maps model than XML or ER, and this means that RDF data can be used in Topic Maps systems with almost no mapping, as demonstrated in Refs. [29,30]. It has also been demonstrated that data conforming to EXPRESS data models[22] can be transformed to Topic Maps with relative ease, although this work has not yet been published.

AREAS OF APPLICATION

As the preceding section has shown, virtually any form of structured data, information, or knowledge can be represented as a topic map. This is not to suggest that all information should be maintained directly in Topic Maps form. A more common scenario is to use Topic Maps as a meta-model for integrating portions of data that originate from different systems (some of which may be Topic Maps-based). The procedure is to generate multiple topic maps from disparate sources and then merge them to provide a unified view of the information gestalt.

It is the flexibility of the Topic Maps model and the robustness of the identity paradigm on which merging is based that provide the clue to identifying application scenarios in which the technology can excel. Another quality of Topic Maps is its intuitiveness: experience shows that the TAO model is very easy for users to grasp, presumably because it derives from artifacts, such as indexes, that humans have used for centuries to locate and manage information. The associative nature of the model is also a

[22]The EXPRESS modeling language is part of the product data standard STEP (ISO 10303).

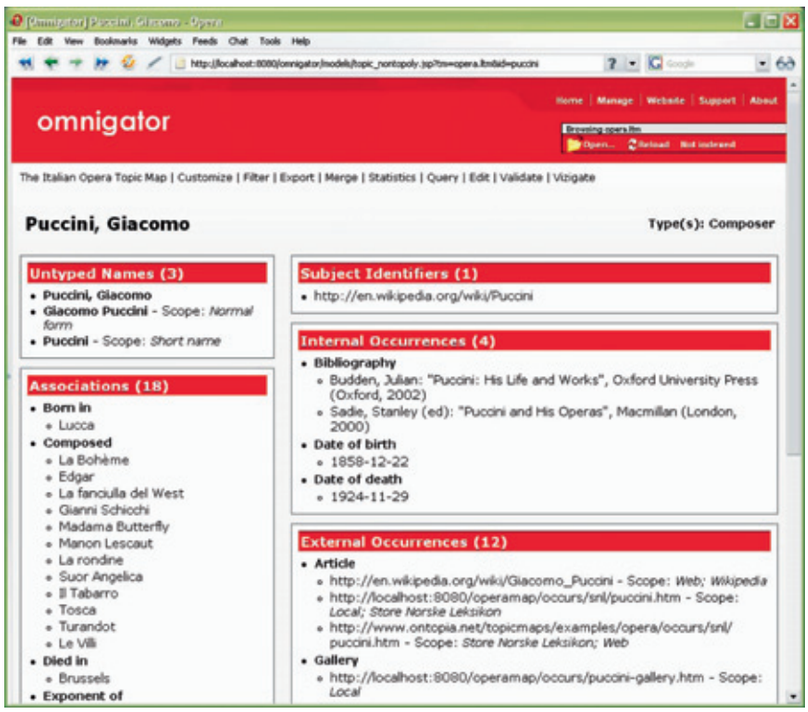

Fig. 7 A topic page in the Omnigator.

factor in this, reflecting as it does how people think, learn, and store information in their own "semantic memories."

Most applications of Topic Maps fall into four broad categories: enterprise information integration (EII), KM, e-learning, and Web publishing.

From the EII perspective, Topic Maps offers an out-of-the-box "meta-model" for integrating information, and a powerful identity mechanism for enabling subject-based merging. A topic map can provide an aggregation layer on top of existing information systems, or function as a hub for transferring data between systems, or both. Either way, Topic Maps removes the need for costly point-to-point system integration.

From the KM perspective, Topic Maps provides—arguably for the first time—the ability to really capture and manage some degree of *human knowledge*—not just information—in a standards-based form, enabling it to be shared and reused across departments, organizations, and systems. The topic/association layer constitutes a distributable and queryable knowledge base, which can be tightly coupled with content management systems via the occurrence axis.

In the domain of e-learning, Topic Maps has particular relevance, not just because of its strengths as a content management technology, but because it bridges the gap between information and knowledge.[31] In Norway, successful experiments have been conducted in which school students are encouraged to create topic maps to record what they have learned, and the National School Curriculum itself now has its definitive expression in the form of a topic map.[32]

One application whose potential is as yet largely untapped is the use of Topic Maps as the foundation for digital libraries. At its most basic, a digital library can be defined as digital content that is organized along bibliographic principles. It requires technology of a new kind that is attuned to the needs of digital information while at the same time supporting bibliographic principles and practices. Topic Maps is just that. An early example of the application of Topic Maps to this purpose is the award-winning New Zealand Electronic Text Centre (Fig. 8).[33]

In all of these applications, and many others besides, the World Wide Web provides the most important channel for disseminating information, via Web sites, portals, and intranets. Despite in some ways predating the Web, Topic Maps has proven to be well-suited to this form of publishing: the TAO model offers a readymade "information architecture" in which each page is devoted to a single well-defined topic and contains information collated from the names and occurrences of that topic, with intuitive navigation paths between pages generated from associations between topics. An increasing number of second and third generation Web sites are currently reorganizing their content into topic maps, and a product such as the *Omnigator*,[34] which is a Web-based browser for topic maps (Fig. 7), is essentially a tool for creating "instant Web sites": when it loads a topic map, the result is a bare-bones, out-of-the-box Web site.

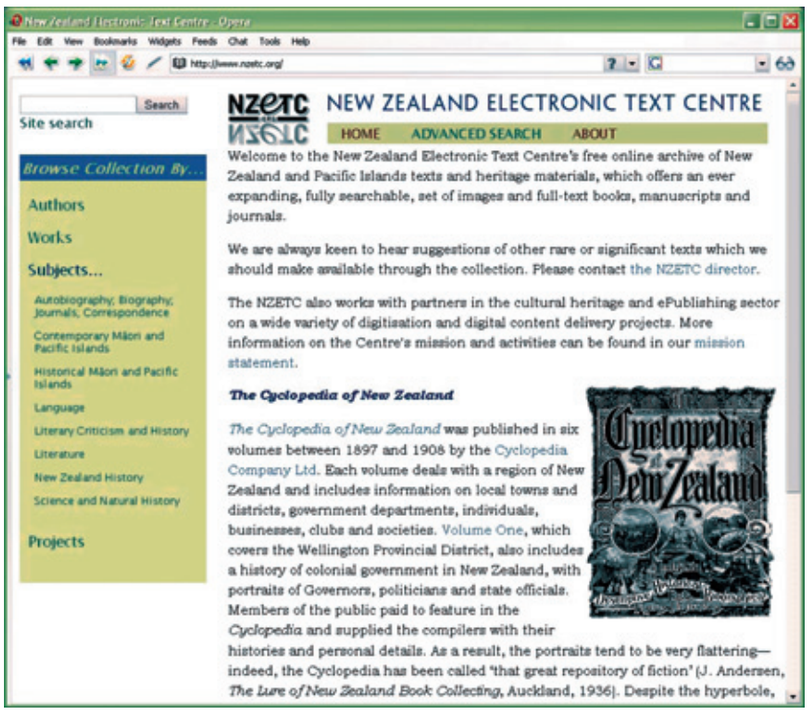

Fig. 8 A Topic Maps-based digital library.
Source: www.nzetc.org.[33]

CONCLUSION

Topic Maps is a paradigm-shifting technology. Through its emphasis on the centrality of subjects, rather than documents or applications, it presages a radically new way of using computers to manage information and knowledge, dubbed subject-centric computing.

The key strengths of Topic Maps are a flexible model that can represent any kind of data structure; a robust model of identity; the ability to merge arbitrary topic maps; a well-defined syntax for exchanging topic maps between systems; status as an ISO-approved international standard; and a vibrant and committed user community.[23]

Topic Maps is particularly well-suited for integrating information from disparate sources, bridging the gap between information and knowledge, and acting as an information architecture for Web-based information delivery.

REFERENCES

1. Pepper, S.; Moore, G., Eds. *XML Topic Maps (XTM) 1.0*; TopicMaps.Org 2001, http://www.topicmaps.org/xtm/1.0/ (accessed November 15, 2008).
2. *ISO/IEC 13250:2003 Information Technology—SGML Applications—Topic Maps,* 2nd Ed.; International Organization for Standardization: Geneva, Switzerland, 2003.
3. *ISO/IEC 13250 Information Technology—Topic Maps—Part 1: Overview and Basic Concepts*, http://www.jtc1sc34.org/repository/0877.zip (accessed November 15, 2008).
4. *ISO/IEC 13250:2006 Information Technology—Topic Maps—Part 2: Data Model*, http://www.isotopicmaps.org/sam/sam-model/ (accessed November 15, 2008).
5. *ISO/IEC 13250:2007 Information Technology—Topic Maps—Part 3: XML Syntax*, http://www.isotopicmaps.org/sam/sam-xtm/ (accessed November 15, 2008).
6. *ISO/IEC 13250 Information Technology—Topic Maps—Part 4: Canonical Syntax*, http://www.isotopicmaps.org/cxtm/ (accessed November 15, 2008).
7. *ISO/IEC 13250 Information Technology—Topic Maps—Part 5: Reference Model*, http://www.isotopicmaps.org/tmrm/ (accessed November 15, 2008).
8. *ISO/IEC 13250 Information Technology—Topic Maps—Part 6: Compact Syntax*, http://www.isotopicmaps.org/ctm/ (accessed November 15, 2008).
9. *ISO/IEC 13250 Information Technology—Topic Maps—Part 7: Graphical Notation*, http://www.isotopicmaps.org/gtm/ (accessed November 15, 2008).
10. *ISO/IEC CD 18048: Information Technology—Topic Maps—Query Language (TQML)*; International Organization for Standardization: Geneva, Switzerland, 2007, http://www.isotopicmaps.org/tmql/ (accessed November 15, 2008).
11. *ISO/IEC CD 18048: Information Technology—Topic Maps—Constraint Language (TMCL)*; International

[23]Witness the annual user's conference in Oslo, the annual academic conference (Topic Maps Research and Applications, TMRA) in Leipzig, the Web site http://www.topicmaps.com, various mailing lists, etc.

Organization for Standardization: Geneva, Switzerland, 2007, http://www.isotopicmaps.org/tmcl/ (accessed November 15, 2008).
12. *ISO/IEC TR 29111 Information Technology—Topic Maps—Expressing Dublin Core Metadata Using Topic Maps*; International Organization for Standardization: Geneva, Switzerland, 2007, http://www.jtc1sc34.org/repository/0884.htm (accessed November 15, 2008).
13. Pepper, S. *The Italian Opera Topic Map*; http://www.ontopedia.net/omnigator/ (accessed November 15, 2008).
14. Pepper, S. *The TAO of Topic Maps*; Ontopia, 2002, http://www.ontopia.net/topicmaps/materials/tao.html (accessed November 15, 2008).
15. Newcomb, S.R. A perspective on the quest for global knowledge interchange. In *XML Topic Maps: Creating and Using Topic Maps for the Web*; Park, J., Hunting, S., Eds.; Addison-Wesley: Boston, MA, 2003, http://www.pearsonhighered.com/samplechapter/0201749602.pdf (accessed November 15, 2008).
16. Pepper, S.; Schwab, S. *Curing the Web's Identity Crisis: Subject Indicators for RDF*; Ontopia, 2003, http://www.ontopia.net/topicmaps/materials/identitycrisis.html (accessed November 15, 2008).
17. Pepper, S. *The Case for Published Subjects*; Ontopia, 2006, http://www.ontopia.net/topicmaps/materials/The_Case_for_Published_Subjects.pdf (accessed November 15, 2008).
18. Sowa, J.F. *Knowledge Representation: Logical, Philosophical and Computational Foundations*; Brooks/Cole: Pacific Grove, CA, 2000.
19. Garshol, L.M. *The Linear Topic Map Notation v 1.3*; Ontopia, 2006, http://www.ontopia.net/download/ltm.html (accessed November 15, 2008).
20. Barta, R.; Heuer, L. *AsTMa = 2.0 Language Definition*; http://astma.it.bond.edu.au/astma = -spec-2.0r1.0.dbk (accessed November 15, 2008).
21. Garshol, L.M. *Metadata? Thesauri? Taxonomies? Topic Maps!* Ontopia, 2004, http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html (accessed November 15, 2008).
22. Mulvany, N.C. *Indexing Books*; University of Chicago Press: Chicago, IL, 1994.
23. Svenonius, E. *The Intellectual Foundations of Information Organization*; MIT Press: Cambridge, MA, 2000.
24. DCMI Usage Board, *DCMI Metadata Terms*, http://dublincore.org/documents/dcmi-terms/ (accessed November 15, 2008).
25. Pepper, S. Expressing dublin core in topic maps. In *Scaling Topic Maps*; http://www.springerlink.com/content/94u5130537r38172/ (accessed November 15, 2008).
26. Oh, S.G. *MARC, FRBR and RDA: Topic Maps Perspective*, Topic Maps 2008, http://www.topicmaps.com/tm2008/oh.ppt (accessed November 15, 2008).
27. Pepper, S.; Garshol, L.M. *The XML Papers*; Ontopia, 2002, http://www.ontopia.net/topicmaps/materials/xmlconf.html (accessed November 15, 2008).
28. Garshol, L.M.; Bogachev, D. TM/XML—Topic maps fragments in XML. In *Charting the Topic Maps Research and Applications Landscape*, http://www.springerlink.com/content/m376708254802517/ (accessed November 15, 2008). See also http://www.ontopia.net/topicmaps/tmxml.html (accessed November 15, 2008).
29. Pepper, S.; Vitali, F.; Garshol, L.M.; Gessa, N.; Presutti, V., Eds. *A Survey of RDF/Topic Maps Interoperability Proposals*; W3C Working Group Note: February 10, 2006, http://www.w3.org/TR/rdftm-survey (accessed November 15, 2008).
30. Pepper, S.; Presutti, V.; Garshol, L.M.; Vitali, F., Eds. *Guidelines for RDF/Topic Maps Interoperability*; W3C Editor's Draft: June 30, 2006, http://www.w3.org/2001/sw/BestPractices/RDFTM/guidelines-20060630.html (accessed November 15, 2008).
31. Lavik, S.; Meløy, J.R.; Nordeng, T.W. BrainBank learning—building personal topic maps as a strategy for learning. In Proceedings of XML 2004, Washington, DC, http://www.idealliance.org/proceedings/xml04/papers/21/brainbank.pdf (accessed November 15, 2008).
32. Lavik, S.; Nordeng, T.W.; Meløy, J.R.; Hoel, T. Remote topic maps in learning. In *Leveraging the Semantics of Topic Maps*, http://www.springerlink.com/content/g56x15p381uh32k9/ (accessed November 15, 2008).
33. *New Zealand Electronic Text Centre*, http://www.nzetc.org (accessed November 15, 2008).
34. Ontopia. *Omnigator*; Ontopia, 2001–2006, http://www.ontopia.net/omnigator (accessed November 15, 2008).
35. Maicher, L.; Garshol, L.M.; Eds. *Scaling Topic Maps*; In Third International Conference on Topic Maps Research and Applications, TMRA 2007, Leipzig, Germany, October 2007; Springer-Verlag: Berlin, Heidelberg, October 2008.
36. Maicher, L.; Park, J.; Eds. *Charting the Topic Maps Research and Applications Landscape*; In First International Workshop on Topic Maps Research and Applications, TMRA 2005, Leipzig, Germany, October 2005; Springer-Verlag: Berlin, Heidelberg, 2006.
37. Maicher, L.; Sigel, A.; Garshol, L.M.; Eds. *Leveraging the Semantics of Topic Maps*; In Second International Conference on Topic Maps Research and Applications, TMRA 2006, Leipzig, Germany, October 2006; Springer-Verlag: Berlin, Heidelberg, 2007.