Managing TEX Resources with XML Topic Maps

Tomasz Przechlewski
Uniwersytet Gdański, Wydział Zarzadzania
81-824 Sopot
ul. Armii Krajowej 119/121
Poland
tomasz@gnu.univ.gda.pl

Abstract

For many years the Polish TEX Users Group newsletter has been published online on the GUST web site. The repository now contains valuable information on TEX, METAFONT, electronic documents, computer graphics and related subjects. However, access to the content is very poor: it is available as PS/PDF files with only a simple HTML page facilitating navigation. There is no integration with information resources from other sections of the site, nor with the resources from other LUG or CTAN sites.

Topic maps were initially developed for efficient preparation of indices, glossaries and thesauruses for electronic documents repositories, and are now codified as both the ISO standard (ISO/IEC 13250) and the XTM 1.0 standard. Their applications extend to the domain of electronic publishing. Topic maps and the similar RDF standard are considered to be the backbone of corporate knowledge management systems and/or the Semantic Web [3].

The paper contains an introduction to the Topic Maps standard and discusses selected problems of Topic Map construction. Finally the application of Topic Maps as an interface to the repository of TEX related resources is presented, as well as the successes and challenges encountered in the implementation.

1 Introduction

All the papers published for the last 10 years in the bulletin of the Polish TEX Users' Group (GUST, http://www.gust.org.pl/) are now available online from the GUST web site. The repository contains valuable information on TEX, METAFONT, electronic documents, computer graphics, typography and related subjects. However, access to the content itself is very poor: the papers are available as PS/PDF files with only a simple HTML interface facilitating navigation. There is no integration with other resources from that site. As CTAN and other LUGs' sites provide more resources it would obviously be valuable to integrate them too.

At first glance, the Topic Maps framework appears to be an attractive way to integrate vast amounts of dispersed TEX related resources. A primary goal of the proposed interface should be to support learning. If the project succeeds, we hope it will change slightly the opinion of TEX as a very difficult subject to become acquainted with.

The paper is organized as follows. The standard is introduced and selected problems of topic maps construction are discussed in the subsequent three sections. Then a short comparison of Topic Maps and RDF is presented. The application of Topic Maps as an interface to the GUST resource repository is described in the last two sections.

2 What is a Topic Map?

Topic Maps are an SGML/HYTIME based ISO standard defined in [1] (ISO/IEC 13250, often referred as HYTM). The standard was recently rewritten by an independent consortium, TopicMaps.org [19] and renamed to XML Topic Maps (XTM). XTM was developed in order to simplify the ISO specification and enable its usage for the Web through XML syntax. Also, the original linking scheme was replaced by XLINK/XPOINTER syntax. XTM was recently incorporated as an Annex to [1].

The standard enumerates the following possible applications of TMs [1]:[1]

- To qualify the content and/or data contained in information objects as topics, to enable navigational tools such as indexes, cross-references, citation systems, or glossaries.

[1] Examples of application of Topic Maps to real world problems can be found in [9, 21, 5, 18, 11, 12].

- To link topics together in such a way as to enable navigation between them.
- To filter an information set to create views adapted to specific users or purposes. For example, such filtering can aid in the management of multilingual documents, management of access modes depending on security criteria, delivery of partial views depending on user profiles and/or knowledge domains, etc.
- To add structure to unstructured information objects, or to facilitate the creation of topic-oriented user interfaces that provide the effect of merging unstructured information bases with structured ones.

In short, a *topic map* is a model of knowledge representation based on three key notions: *topics* which represent subjects, *occurrences* of topics which are links to related resources, and *associations* (relations) among topics.

A topic represents, within an application context, any clearly identified and unambigous subject or concept from the real world: a person, an idea, an object etc.

A topic is a instance of a topic type. Topic types can be structured as hierarchies organized by superclass-subclass relationships. The standard does not provide any predefined semantics to the classes. Finally, topic and topic type form a class-instance relationship.

Topic have three kinds of characteristics: *names* (none, one, or more), *occurrences*, and *roles* in associations. The links between topics and their related information (web page, picture, etc.) are defined by *occurrences*. The linked resources are usually located outside the map. XTM uses a simple link mechanism as defined in XLINK, similar to HTML hyperlinks.

As with topics, occurrences can be typed; occurrence types are often referred as *occurrence roles*. Occurrence types are also defined as topics. Using XML syntax, the definition of topic is quite simple:

```
<topic id="t-przechlewska-wanda">
  <instanceOf>
    <topicRef xlink:href="#person"/>
  </instanceOf>
    <baseName>
      <baseNameString>Plata-Przechlewska,
        Wanda</baseNameString>
  </baseName>
</topic>
```

Topic associations define relationships between topics. As associations are independent of the resources (i.e., data layer) they represent added-value information. This independency means that a concrete topic map can describe more than one information pool, and vice versa. Each association can have an *association type* which is also a topic. There are no constraints on how many topics can be related by one association. Topics can play specific roles in associations, described with *association role types* — which are also topics.

The concepts described above are shown in Fig. 1. Topics are represented as small ovals or circles in the upper half of the picture while the large oval at the bottom indicates data layer. Small objects of different shapes contained in the data layer represents resources of different types. The lines between the data layer and topics represents occurrences, while thick dashed ones between topics depicts associations.

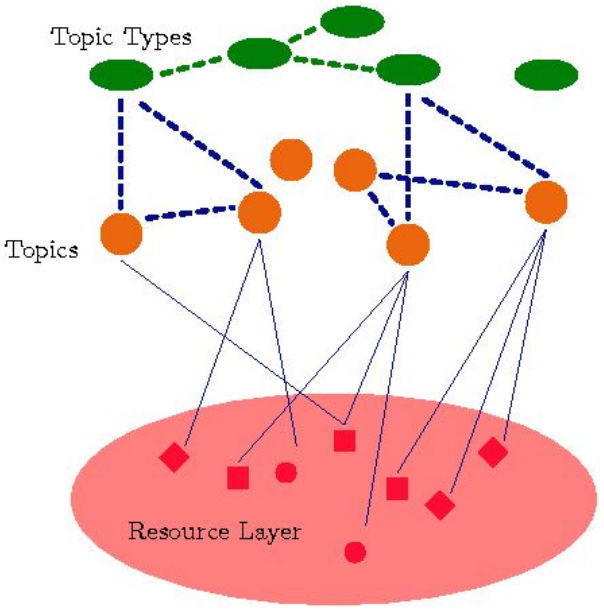

Figure 1: Topic map and resource layer.

Besides the above mentioned three fundamental concepts, the standard provides a notion of *scope*. All characteristics of topics are valid within certain bounds, called a *scope*, and determined in terms of other topics. Typically, scopes are used to model multilingual documents, access rights, different views, and so on.

Scopes can also be used to avoid name conflicts when a single name denotes more than one concept. An example of scope for the topic *latex* might be *computer application* or *rubber industry* depending on the subject of the topic. Only the topic characteristics can be scoped, not the topic itself.

3 Subject Identity and Map Merging

From the above short tour of TM concepts it should be clear that there is an exact one-to-one correspondence between subjects and topics. Thus, the identification of subjects is crucial to individual topic map applications and to interoperability between different topic maps.

The simplest and most popular way of identifying subjects is by identifyng them via some system of unique labels (usually URIs). A subject identifier is simply a URI unambiguously identifying the subject. If the subject identifier points to a resource (not required) the resource is called a subject indicator. The subject indicator should contain human-readable documentation describing the non-addressable subject [22].

As there are no restrictions to prevent every map author from defining their own subject identifiers and resource indicators, there is a possibility that semantic or syntactic overlap will occur. To overcome this, *published subject indicators* (PSIs) are proposed [17]. PSIs are stable and reliable indicators published by an institution or organization which desires to promote a specific standard. Anyone can publish PSIs and there is no registration authority. The adoption of PSIs can therefore be an open and spontaneous process [17, 6].[2]

Subject identity is of primary importance for topic map merging when there is a need to recognize which topics describe the same subject.

Two topics and their characteristics can be *merged* (aggregated) if the topics share the same name in the same scope (*name-based merging*), or if they refer to the same subject indicator (*subject-based merging*). Merging results in a single topic that has the union of all characteristics of merged topics. Merged topics play the roles in all the associations that the individual topics played before [22, 15].

4 Constraining, Querying, and Navigating the Map

The notion of a topic map template is used frequently in literature. As the name suggests, a *topic map template* is a sort of schema imposing constraints on topic map objects with TM syntax. The standard does not provide any means by which the designer of the TM template can put constraints onto the topic map itself. Standardisation of such constraints are currently in progress [14].

[2] For example, the XTM 1.0 specification contains a set of PSIs for core concepts, such as class, instance, etc., as well as for the identification of countries and languages [19].

Displaying lists of indexes which the user can navigate easily is the standard way of TM visualization. As this approach does not scale well for larger maps, augmenting navigation with some sort of searching facility is recommended. Other visualization techniques such as hyperbolic trees [15], cone trees, and hypergraph views (Fig. 2) can be used for visualization and navigation of topic maps. They display TMs as a graph, with the topics and occurrences as nodes and the associations as arcs. The drawback of such 'advanced' techniques is that users are usually unfamilar with them.

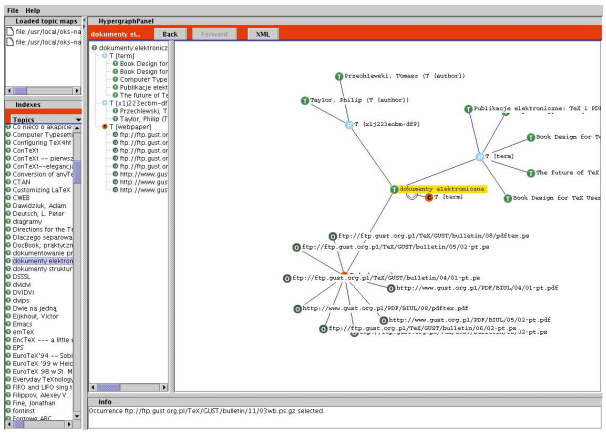

Figure 2: Hypergraph visualization with TMNav.

There are several proposed query languages for topic maps. None of them are part of the standard and there are inconsistencies in different TM engines. Two of the most prominent proposals are:

- TMQL (*Topic Maps Query Language*, [9]), with SQL-like syntax, provides both for querying and modifying topic maps (select, insert, delete, update).
- Tolog, inspired by logic programming language Prolog, supports requirements for TMQL with clearer and simpler syntax.

The introduction to the TM standard presented in this paper does not cover all the details of the technology. Interested readers can find an exhaustive description in [15], which contains detaileda introduction with numerous examples, and [16].

5 Topic Maps and RDF

The W3C promotes the Resource Description Framework (RDF) [10] as another framework for expressing metadata. RDF is a W3C standard envisioned to be a foundational layer of the Semantic Web.

The fundamental notion in the RDF data model is a *statement*, which is a triple composed of a

resource, *property*, and *value*. The RDF Schema (RDFS) [4] is a W3C working draft aimed at defining a description language for vocabularies in RDF. More expressive RDFS models have been proposed recently [23].

One key difference between RDF and topic maps is that topic maps are modelled on a concept-centric view of the world. For example, in RDF there are no 'predefined' properties, so to assign a name to a resource one has to use another standard (such as Dublin Core), something that is not necessary with topic maps. The notion of scope is also absent from RDF too.

The RDF and Topic Maps standards are similar in many respects [7]. Both offer simple yet powerful means of expressing concepts and relationships.

6 Building Topic Maps for the GUST Bibliographic Database

Similar to writing a good index for a book, creating a good topic map is carried out by combining manual labour with the help of some software applications. It is usually a two-stage task, beginning with the modelling phase of building the 'upper-part' of the map, i.e., the hierarchy of topic types and association types (the *schema* of the map) and then populating the map with instances of topic types, their associations and occurrences.

Approaches for developing a topic map out of a pool of information resources include [2]:

- using standard vocabularies and taxonomies (i.e., `www.dmoz.org`) as the initial source of topics types.
- generating TMs from the structured databases or documents with topic types and association types derived from the scheme of the database/document.
- extraction of topics and topic associations from pools of unstructured or loosely structured documents using NLP (Natural Language Processing) software combined with manual labour.

The first approach is concerned with the modelling phase of topic map creation, while the third one deals with populating the map.

Following the above guidelines, the structure of the BibTeX records was an obvious candidate to start with in modelling our map of GUST articles. It provides a basic initial set of topics including: *author*, *paper*, *keyword*, and the following association types: *author-paper*, *paper-keyword* and *author-keyword*. Abstracts (if present in BibTeX databases) can be considered as occurrences of the topic *paper*. The publication date and language can be used as scopes for easy navigation using them as constraints.

Other TAOs (topics, associations, and occurrences [16]) to consider are: author home pages (occurrence type), applications described in the paper (association type), papers referenced (association type). This information is absent from BibTeX files but, at least theoretically, can be automatically extracted from the source files of papers.

We started by analyzing the data at our disposal, i.e., TeX and BibTeX source files. Unfortunately, in the case of the GUST bulletin the BibTeX database was not maintained. This apparent oversight was rectified with simple Perl scripts and a few days of manual labour. The bibliographic database was created and saved in a XML-compatible file.[3]

TeX documents are often visually tagged and lack information oriented markup. The only elements marked up consistently and unambiguously in the case of the GUST bulletin are the paper titles and authors' names. Authors' home pages were rarely present, while email addresses were available but not particularly useful for our purposes. Neither abstracts nor keyword lists had been required and as a consequence were absent from the majority of the papers. Similarly, any consistent scheme of marking bibliographies (or attaching `.bib` files) was lacking, so there was no easy way to define the *related to* association between papers.

The benefit derived from keywords is much greater if they are applied consistently according to some fixed classification; otherwise, the set of keywords usually consists of many random terms which are nearly useless. Since we didn't want to define yet another 'standard' in this area, we would have liked to adopt an existing one. The following sources were considered: the TeX entry at `dmoz.org`, Graham Williams' catalogue.[4], collections of BibTeX files and `.tpm` files [20]

The accuracy of the TeX taxonomy subtree at `dmoz.org` was somewhat questionable, and we quickly rejected the idea of using it. Williams' catalogue of TeX resources does not include any information except the location of the resource in the structure of CTAN. As for BibTeX files, it appeared only MAPS and *TUGboat* were complete and up-to date[5] but only the latter contains keywords. Unfortunately, they don't comply with any consistent

[3] We reused the XML schema developed for the MAPS bulletin (`http://www.ntg.org.ln/maps/`).

[4] `http://www.ctan.org/tex-archive/help/Catalogue`

[5] Cahiers GUTenberg was not found, but the impressive portal of Association GUTenberg indicates appropriate metadata are maintained, but not published.

scheme. Due to the lack of any existing standard, the keywords were added manually on a common-sense basis, with the intention of being 'in sync' with the most frequent terms used in MAPS.[6]

Finally the following TAOs were defined (the language of the publication was considered to be the only scope):

- topic types: *author*, *paper*, and *keyword*;
- association types: *author-paper*, *paper-keyword*, and *author-keyword*;
- occurrence types: *papers* and *abstracts*.

The schema of the map was prepared manually and then the rest of the map was generated from the content of intermediate XML file with an XSLT stylesheet [8, 13]. The resulting map consists of 454 topics, 1029 associations, and 999 occurrences. A fragment of the map rendered in a web browser with Ontopia Omnigator (a no-cost but closed-source application, `http://www.ontopia.net/download/`) is shown in Fig. 3.

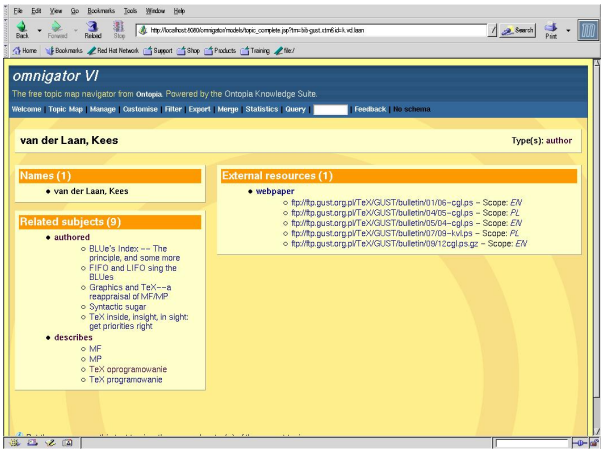

Figure 3: A fragment of GUST topic map rendered with Omnigator.

Omnigator shows the map as a list of links arranged in panels. Initially only a list of subjects (index of topic types) is displayed. When a link for a topic is clicked on, the topic is displayed with all the information about its characteristics (names, associations, occurrences). In Fig. 3, an example page for author *Kees van der Laan* is shown. The right panel contains all the relevant resources while the lower left has all the related topics, i.e., papers written by Kees and other subjects covered. The user can easily browse both papers authored by him and switch to pages on some other interesting subject. The panel with resources contains information on the resource type allowing fast access to the required data.

Similar functionality can be obtained with the freely available TM4Nav or even by using a simple XSLT stylesheet [13].

7 Integrating Other TEX Resources

So far there is nothing in TMs which cannot be obtained using other technologies. The same or better functionality can be achieved with any database management system (DMS). But integrating TEX resources on a global scale needs flexibility, which traditional RDBMS-based DMS applications lack. For example, topic maps can be extended easily through merging separate maps into one, while DMS-based extensions usually require some prior agreement between the parties (e.g., LUGs), schema redefinitions, and more.

To verify this flexibility in practice, we extended the GUST map with the MAPS and TUB BIBTEX databases. For easy interoperability in a multi-language environment, the upper half of the map was transferred to a separate file. With the use of scope, the design of multi-language topic types was easy, for example:

```
<topic id="english">
 <subjectIdentity>
   <subjectIndicatorRef
    xlink:href="http://www.topicmaps.org/\
      xtm/1.0/language.xtm#en"/>
   </subjectIdentity>
  <baseName>
    <baseNameString>EN</baseNameString>
  </baseName>
</topic>
 ...
<topic id="author">
 <baseName><scope>
   <topicRef xlink:href="#english"/></scope>
  <baseNameString>author</baseNameString>
 </baseName>
 <baseName><scope>
    <topicRef xlink:href="#polish"/> </scope>
   <baseNameString>autor</baseNameString>
  </baseName>
</topic>
```

Other topic types were designed similarly. Scopes for other languages can easily be added.

The 'lower part' of the map was generated from (cleaned) BIBTEX records with `bibtex2xml.py`

[6] There are 814 bibliographic entries in MAPS base and 895 different keywords. The most popular keywords in MAPS BIBTEX file are: LATEX – 51, NTG – 42, plain TEX – 37, PostScript – 28, ConTEXt, TEX-NL, METAFONT, SGML, and so forth. There are small number of inconsistent cases (special commands vs. specials, or configuration vs. configuring) and fine-grained keywords (Portland, Poland, Bachotek, USSR!).

(`http://bibtexml.sf.net`) and than transformed to MAPS XML with an XSLT stylesheet. Keywords were added to TUB entries using a very crude procedure.[7]

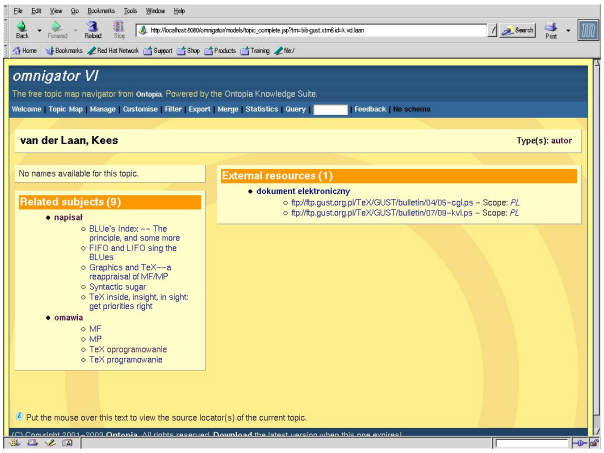

Figure 4: Topic map fragment from Fig. 3 scoped to Polish language.

Simple name-based merging of all three maps results in over 25,000 TAOs (≈ 1000 authors, more than 2000 papers). Some of the subjects were represented with multiple topics. As an example the Grand Wizard was represented as the following four distinct topics: 'Knuth, Don', 'Knuth, Donald', 'Knuth, Donald E.', 'Knuth., Donald E.'.[8]

As identity-based merging is regarded as more robust, some identifiers have to be devised first. Establishing a PSI for every TEX author seemed overly ambitious. Instead, a dummy subject identifier was chosen, such as: `http://tug.org/authors#initials-surname`. This can still produce multiple topics for the same subject, but now we can eliminate unwanted duplicates by defining an additional map consisting solely of topics like the following [18]:

```
<topic id="de-knuth">
 <subjectIdentity>
  <subjectIndicatorRef
    xlink="http://tug.org/authors#d-knuth"/>
  <subjectIndicatorRef
   xlink="http://tug.org/authors#d-e-knuth"/>
 </subjectIdentity>
</topic>
```

Merging this map with the 'base' map(s) will result in a map free of unwanted duplicated topics with all variant names preserved.

[7] Acronyms, such as LATEX, METAFONT, or XML, present in the title were used as keywords.

[8] First name variants, abbreviations and middle names cause problems in many more cases.

For further extensions, we plan to incorporate CTAN resources. For that purpose, Williams' catalogue and/or the TPM files from TEX Live project can be used. As the catalogue contains author names, it would be for example possible to enrich the map with the *author-application* association. Further enrichment will result if we can link applications with documents describing them. However, some robust classification scheme of TEX resources should be devised first.

8 Topic Map Tools

As with any other XML-based technology, topic maps can be developed with any text editor and processed with many XML tools. However, for larger-scale projects specialized software is needed. There are a few tools supporting topic map technology, developed both commercially and as Open Source projects. We have considered both Ontopia Omnigator (mentioned in the previous section) and TM4J (free software).

TM4J (`http://tm4j.org`) is a suite of Java packages which provide interfaces and default implementations for the import, manipulation and export of XML Topic Maps. Features of the TM4J engine include an object model which supports XTM specification with the ability to store topic map in an object-oriented or relational database, and an implementation of the tolog query language.

Based on TM4J a few projects are in progress: TMNav for intuitive navigation and editing of topic maps, and TMBrowse for publishing maps as set of HTML pages (similarly to Omnigator).

These projects are in early stages and our experience with TMBrowse indicates that current version frequently crashes with bigger maps and is significantly slower than Omnigator. There were problems with `tolog` queries as well.

As all these projects are actively maintained progress may be expected in the near future.

9 Summary

Topic maps are an interesting new technology which can be used to describe the relation between TEX resources. The main problem is topic map visualization. Available tools are in many cases unstable and non-scalable, but we can expect improvement.

The system presented here can certainly be improved. It is planned to extend it with the content of Williams' catalogue. The maps developed in the project are available from `http://gnu.univ.gda.pl/~tomasz/tm/`. At the same address, the interested reader can find links to many resources on topic maps.

References

[1] ISO/IEC. Topic Maps, ISO/IEC 13250, 2002. http://www.y12.doe.gov/sgml/.

[2] Kal Ahmed, Danny Ayers, Mark Birbeck, and Jay Cousins. *Professional XML Meta Data*. Wrox Press, 2001.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):35–43, 2001. http://www.sciam.com/2001/0501issue/0501berners-lee.html.

[4] Dan Brickley and R. V. Guha. RDF vocabulary description language 1.0: RDF schema, 2002. http://www.w3.org/TR/rdf-schema/.

[5] Anna Carlstedt and Mats Nordborg. An evaluation of topic maps. Master's thesis, Göteborg University, 2002. http://www.cling.gu.se/~cl8matsn/exjobb.html.

[6] Paolo Ciancarini, Marco Pirruccio, and Fabio Vitali. Metadata on the Web. On the integration of RDF and topic maps. In *Extreme Markup Languages*, 2003. http://www.mulberrytech.com/Extreme/Proceedings/xslfo-pdf/2003/Presutti01/EML2003Presutti01.pdf.

[7] Lars M. Garshol. Living with Topic Maps and RDF. http://www.ontopia.net/topicmaps/materials/tmrdf.html, 2002.

[8] Michael Kay. *XSLT Programmer's Reference 2nd Edition*. Wrox Press, 2001.

[9] Rafał Ksieżyk. Trying not to get lost with a topic map. In *XML Europe Conference*, 1999. http://www.infoloom.com/gcaconfs/WEB/granada99/ksi.HTM.

[10] Ora Lassila and Ralph R. Swick. Resource description framework (RDF). Model and syntax specification, 1999. http://www.w3.org/TR/REC-rdf-syntax/.

[11] Xia Lin and Jian Qin. Building a topic map repository, 2000. http://www.knowledgetechnologies.net/proceedings/presentations/lin/xialin.pdf.

[12] Ashish Mahabal, S. George Djorgovski, Robert Brunner, and Roy Williams. Topic maps as a virtual observatory tool, 2002. http://arxiv.org/abs/astro-ph/0110184.

[13] Sal Mangano. *XSLT Cookbook*. O'Reilly, 2002.

[14] Mary Nishikawa and Graham Moore. Topic map constraint language requirements, 2002. http://www.isotopicmaps.org/tmcl/.

[15] Jack Park and Sam Hunting, editors. *XML Topic Maps. Creating and using Topic Maps for the Web*. Addison-Wesley, 2002.

[16] Steve Pepper. The TAO of topic maps, 2000. http://www.ontopia.net/topicmaps/materials/tao.html.

[17] Steve Pepper. Published subject: Introduction and basic requirements, 2003. http://www.oasis-open.org/committees/.

[18] Steve Pepper and Marius L. Garshol. Lessons on applying topic maps. http://www.ontopia.net/topicmaps/materials/xmlconf.html, 2002.

[19] Steve Pepper and Graham Moore. XML topic maps (XTM) 1.0, 2000. http://www.topicmaps.org/xtm/1.0/.

[20] Fabrice Popineau. Directions for the TeXlive systems. In *EuroTEX 2001, The Good, the Bad and the Ugly, Kerkrade*, pages 152–161. NTG, 2001. http://www.ntg.nl/maps/pdf/26_20.pdf.

[21] Tomasz Przechlewski. Wykorzystanie map pojeć w zarzadzaniu repozytoriami dokumentów elektronicznych. In *Materiały Konferencji: MSK 2003*, 2003.

[22] Hans Holger Rath. Semantic resource exploitation with topic maps. In *Proceedings of the GLDV-Spring Meeting 2001*, 2001. http://www.uni-giessen.de/fb09/ascl/gldv2001/.

[23] Michael K. Smith, Chris Welty, and Deborah L. McGuinness. OWL web ontology language guide, 2003. http://www.w3.org/TR/owl-guide/.