

Moving Topic Maps to Mainstream – Integration of Topic Map Generation in the Users’ Working Environment

Karsten Böhm

(University of Leipzig, Germany
boehm@informatik.uni-leipzig.de)

Lutz Maicher

(University of Leipzig, Germany
maicher@informatik.uni-leipzig.de)

Hans Friedrich Witschel

(University of Leipzig, Germany
witschel@informatik.uni-leipzig.de)

Andrea Carradori

(University of Leipzig, Germany
mai02coq@studserv.uni-leipzig.de)

Abstract: Topic Maps are sophisticated indexes for dynamic, heterogeneous, structured, and unstructured information sources. In order to move Topic Maps towards the mainstream, the automatic generation of Topic Maps and its integration in the users working environment and processes must be improved. Our described approach for Topic Map generation is based on terminology extraction with relevance feedback, which improves our previous approaches especially for small corpora. The relation between the Subjects and Topics is the core of the Topic Map theory. We propose a methodology for the proper integration of this underlying theory in the generation of Topic Maps in order to obtain real interchangeable Topic Maps. The framework TOMATO is a scientific prototype which realises the described functionality and offers interfaces for integration in applications and web-based interfaces.

Keywords: Topic Map Generation, Topic Map Integration, Text Processing, Subject Identity

Category: H.3.1, H.3.3, H.5.3, H.3.5, I.2.7, I.7

1 Introduction

Topic Maps¹ are sophisticated indexes for dynamic collections of heterogeneous, structured and unstructured information sources.² Historically, Topic Maps are derived from the idea of back-of-the-book indexes. But they not only provide a simple accessing structure to information the way indexes do. In addition, the sophisticated Metadata they represent can be detached from the original information, interchanged, and reused. In this paper we will focus on Topic Maps as such indexes for dynamic text corpora. In contrast to the full text search like Google [see <http://www.google.de>]

¹ For an in-depth introduction to Topic Map related terminology we refer to the glossary of [TMDM, pp. 1-3]. All terminology used in this paper is capitalized.

² For a gentle introduction in the concept of Topic Maps we refer to [PEP00] and [<http://www.isotopicmaps.org>]. Therefore we renounce of a further introduction.

they supply the user with the significant subjects in a specific domain and their relations among each other. In contrast to category systems like the open directory project [see <http://dmoz.org>], Topic Maps are non-hierarchical graphs of conceptual nodes and typed n-ary associations. This means, that a “Subject of Discourse” is represented by a node of this graph, their relationships are represented by the typed associations.

A further advantage of Topic Maps is their inherent subjectivity because “they add semantics to data without modifying it. Moreover, one Topic Map may describe several information pools and several Topic Maps may apply to one single information pool“. [LEGR⁺02] Especially for personal or community knowledge management the inherent subjectivity is an advantage which must be exploited (see [THOM02]). This is discussed in detail in [HEYE⁺03]. Despite these advantages Topic Map based information access is still far from the mainstream. While most causative problems are discussed in [FREE02] we recognized three additional issues to solve: there is a pronounced need of means for *automatic generation interchangeable Topic Maps*, a means for collaborative refinement of these generated Topic Maps and means for easy integration of the generation process and of the Topic Map handling process in the working environment of knowledge workers. With our framework TOMATO we directly address these needs.

We have seen that there is a need for advanced automatic generation of Topic-Maps. In former contributions to the I-Know we have already presented our approach to Topic Map generation for large text corpora (see [BÖHM⁺02], [BIEM⁺03]). Because of the ongoing standardization process for the Topic Map family [see RATH03] and because of new developments in NLP (Natural Language Processing) this generation can be improved significantly. These enhancements and the underlying theory are presented in this paper.

In addition, we have seen that the easy integration of the generated Topic Maps in applications must be significantly simplified. TOMATO is a scientific prototype framework which gathers our technologies of Topic Map generation in one package and offers interfaces for one-stop integration in the users working environment and processes. The difference, compared to our previous solution, is the variety of TOMATO. Our previous solution was monolithic; this means the generation process was offered like a one-way street. The input was a set of pre-processed texts, the result a XML-file, representing the resulting Topic Map. TOMATO converts this one-way street into a highway with a variety of exits and interchanges. The integration of these techniques in desktop solutions like Office Suites or E-Mail applications is a future application era of TOMATO. In this paper we present GreenTOMATO, the first stable version of our scientific prototype TOMATO.

In the next chapter, we will give an overview of related research. We will continue by describing the relationship between Subjects and Topics as the core of the Topic Map theory. Then we will explain our terminology extraction process based on relevance feedback which provides Topic and Association candidates with good recall and precision.³ Afterwards, we will go on describing GreenTOMATO. We will then conclude with a discussion of our solution and outline further research issues.

³ For evaluation purposes methods and measures normally used in information retrieval one can adopt: precision, recall and F-value [see WITS04, pp. 77]. Precision describes which percentage of the found words (our Subject candidates) are domain-specific terms (real Subjects). Recall describes which percentage of the domain-specific terms (real Subjects) represented in the texts are found by the system. The F-value is the harmonic average of

2 Related Research

In [BÖHM⁺02] we have already shown the automatic generation of Topic Maps from collections of unstructured text-documents. The generation of Topic Maps described in [GRØN02] deals with the mapping of structured data (databases, RDF etc.) into Topic Maps and doesn't handle unstructured data like texts. Because Topic-Maps are advanced indexes which deal with hyperlinks and provide additional semantics, ideas from the following communities can be used for our research: indexing, (conceptual) hypertext and ontology engineering. Especially the use of text mining for generation purposes (indexes, hypertexts, and ontologies) is in the scope of our interests. But all these approaches do not cover the theoretic specialities of Topic Maps discussed in [chapter 3] because we are dealing with a novel research task.

By using our existing techniques in a practical environment for project based knowledge management [see MAIC⁺03B] the possibilities for enhancing the users' productivity were discussed. Still, the application of the shown Topic Map generation to a new application domain is cumbersome, inflexible, and almost always involves a lot of configuration work and domain specific programming. Some problems we are facing in the attempt of the automatic generation of Topic Maps are:

- *Usage of only one NLP method called reference corpus analysis.* The exclusive application of this method leads to satisfactorily results only for large text corpora. In addition, this approach requires manual parameterization. For our improvements [see chapter 4].
- *Ignorance of the underlying theory of Topic Map.* The ignorance of the Topic Map theory leads to not-interchangeable Topic Maps. Especially the relationship between Subjects and Topics as a foundation for Topic Map interchanges within communities must be considered. For our approach [see chapter 3].
- *Today the generation is a closed process.* For integration of Topic Map structured information access the integration of the methods in applications must be provided [see chapter 5]. We have to convert the one-way street approach of our previous solution into a highway with exits and interchanges.
- *No possibilities of human refinement of the resulting Topic Maps.* Because the automatic generation approach cannot be perfect, users should be able to refine the existing Topic Maps. If this is integrated, the resulting system could be used for collaborative approaches to light-ontology design [see BÖHM⁺02, HOLS⁺02].

In the past years, commercial and open source frameworks for handling Topic Maps have been developed. The Ontopia Knowledge Suite OKS [see <http://www.ontopia.net>] as a commercial solution and TM4J [see <http://tm4j.org>] as an open source solution are the most matured. These frameworks are based on the Topic Map standard family continuously improved by the ISO and the Topic Map community [RATH03]. They can be extended and customized by the users.

Precision and Recall. There are modifications of the F-value proposed, where the impact of Precision and Recall can be adjusted. We assume that for our purposes Precision is more important because TOMATO is based on relevance feedback. A bad Precision forces the users to evaluate a large number of irrelevant words.

On the other hand frameworks have been developed to make the text mining processes more flexible and more easily configurable. As an example for a general purpose framework we refer the interested user to GATE⁴, which is “an infrastructure for developing and deploying software components that process human language” [CUNN⁺03] to support a wide area of language engineering applications. The system supports the full lifecycle of language processing components, from corpus collection and annotation to system evaluation. WETA⁵ (see also [KERK03]) is an open source Workbench Environment for Text Analysis implemented in Java that builds uses the concept of agent based systems for their framework. The users of the software currently being developed will be able to implement their own functionality by adding specialized pre-processing filters and dedicated text mining algorithms.

Although these frameworks support the general language engineering processes which are most relevant for our approach, the proposed framework solution differs in that we focus especially on the creation, maintenance and integration of knowledge structures, namely Topic Maps, from textual sources. Thus, we are consciously narrowing down the application area in order to amplify the utility of our framework for the intended application in the domain of knowledge management.

3 The Theory of Topic Maps and the Usage of Subjects

Our previous approach to automatic Topic Map generation lacks the consideration of the Topic Maps’ theoretical fundament. This foundation is the usage of Subjects as central binding points in Topic Maps and Topic Map interchange scenarios. Because this theory is emerging at the moment, we did not include it in our previous approach two years ago. However, this causes improper Topic Maps and the impossibility of Topic Map interchange. In the following, the integration of Subjects in the automatic generation of Topic Maps will be discussed.

According to the proposed Topic Maps Data Model [TMDM], a Topic Map is a set of Topics and Associations. A Topic can represent *any* Subject of discourse. This can be existing “things” like documents or cars, but also abstract “things” like thoughts or notions. Associations represent relationships between these Subjects. Occurrences connect the Subjects to pertinent information resources [TMDM, pp. V]. In theory, each Topic Characteristic (Basename, Occurrence and Association Role belonging to a certain topic) is a statement about the Subject represented by the topic, not about this Topic. We have to bear in mind that a Topic Characteristic is only a statement about a certain Topic, if this certain Topic is the Subject of the according Topic. This special method is called reification and used, for example, to assign statements about an Association to a Topic Map. In this case the association is the Subject of the Topic which collects all information about this association.

The main concept of Topic Maps is that “every topic represents one, and only one, subject” within a Topic Map [TMDM, ch. 5.4]. If more than one Topic represent the same Subject within a Topic Map these Topics must be merged. This method of semantic integration ensures that all information concerning a special Subject is bound to one central Topic. This central Topic allows access to all the appropriated information concerning the represented Topic.

⁴ The GATE-Framework is available under the GNU Public License at <http://gate.ac.uk>

⁵ WETA is available at <http://www.weta-group.net>

While this concept is especially interesting for purposes of Topic Map exchange, the problem of addressing Subjects occurs. Topics have to describe their subject in order to support Topic Map Engines in their decision of whether two Topics represent the same Subject. The Subject of a Topic is referred by a URI. In this case the known problem of what URIs really do identify occurs [see BERN02]. The question is, whether, for example, an email address identifies the regarding account or the person which has this account. While this problem isn't solved in RDF, Topic Maps provide the distinction between Subject Locators and Subject Indicators (or rather Identifiers) [PEPP⁺03 provides a good starting point for that discussion]. A Subject Locator is a URI of an addressable information resource, which *is* the subject, i. e. a document. In contrast, a Subject Indicator is an information resource which should describe unambiguously the (not-addressable) Subject of a Topic to a human being. A Subject Identifier is the URI of the Subject Indicator.

For addressable Subjects this mechanism works well within a given Topic Map and in interchange scenarios. But for non-addressable Subjects naming and interpretation problems emerge. The first problem occurs if the authors of two different Topic Maps used different Subject Indicators to describe exactly the same Subject. The second problem arises because abstract Subjects can't be discriminated sharply [KENT03 provides a good introduction in this discussion]. The usage of identical Subject Indicators by different Topic Map authors only *indicates* Subject uniqueness. The authors' range of interpretation can lead to the usage of the same Subject Indicator for different Subjects. To solve these problems Published Subject Indicators (PSI) are proposed [for more details we refer to OASIS03]. A PSI is a Subject Indicator which is common sense to all its users. Established taxonomies, ontologies, catalogues, and vocabularies provide good (domain specific) PSI candidates.

A solution which wants to establish a reliable subject handling has to fulfil the following tasks:

- *Task 1: Extracting new subjects and their relations from a new text in a given domain.* We need an algorithm which detects unknown subjects and their relations in new texts. These subjects must be stored (in some way) for later reuse of subject extracting and occurrences detection.
- *Task 2: Detecting the occurrence of known subjects in a given text.* If some subjects are already known these must be detected in a given text. The sentences in which the Subject occurs are candidates for the according Subject Indicator.
- *Task 3: Interchange the information about the subjects between two distributed Topic Maps.* If a Topic Map is interchanged the information about the represented subjects should be integrated in this Topic Map. This also allows the receiver of the Topic Maps to fulfil both of the tasks above for the "received" subjects .

4 Terminology Extraction

In the following we will describe, how task 1 (Extraction of new Topics) is realised in our solution. The realisation of task 2 and task 3 is still in its infancy so that it will not be part of this paper. Our approach is based on the assumption that for each domain the used terminology provides good candidates for Subjects. A terminology is the

system of (mental) concepts and their representation of a domain, which consists of all common domain-specific terms [WÜST91, p. V]. The advantage of such terms that are often technical ones is that the relation between concept and representation is usually unique and exact [BLAN97]. This means, that normally technical terms are not ambiguous and do not have synonyms [MAYN99], especially if the terminology is standardized.

We have to bear in mind that technical terms can be either single words or phrases, which normally consist of nouns and adjectives [see for detail WITS04]. In addition the recognition of named entities will be fruitful for Topic Map generation. We assume that each named entity used in domain specific texts will be a relevant expert, product or enterprise.

In our previous approach for generation of Topic Maps we solely used a statistical method for extracting domain specific terms as good Topic candidates [see for detail BÖHM⁺02]. This technique, which we will call reference corpus analysis, extracts all words as good Topic candidates, the relative frequency of which is significantly higher in the given specialised texts than in a given reference corpus. This reference corpus consists of a huge well balanced amount of common-language texts and thus helps to estimate how often words *should* occur in everyday language. For extracting Association candidates, the same method holds for calculating collocations of Topic candidates. On the one hand, the quality of the results of this statistic method increases with the size of the given corpus. This means the smaller the used corpus is, the more the precision declines.

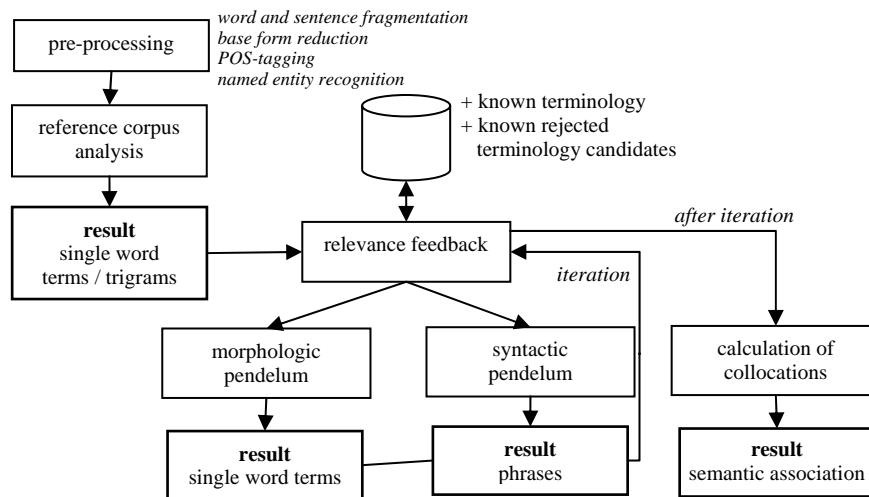


Figure 1 The implemented text processing algorithm

To improve the quality, especially for small and domain specific corpora, additional techniques are available in the field of Natural Language Processing (NLP). One example is the automatic detection of named entities, as described in [BIEM⁺03].

On the other hand, the syntactic pendulum allows the extraction of phrases (many technical terms consist of more than one word) and the morphological pendulum

increases the recall of the system by finding more single word terms than the reference corpus analysis can do [see WITS04].

Another important aspect is relevance feedback: it is often necessary to adapt extraction results to specific user needs. Including knowledge about a user's choice of terms in previous iteration steps can greatly improve precision in the next steps.

Figure 1 sketches the outline of our text processing algorithm briefly. A morphological pendulum searches new term candidates which contain trigrams of characters that were considered domain-specific by the system and the user one iteration step before. One example is the trigram "cyt" in a medical corpus which occurs in many interesting technical terms such as "thrombocytopenia" or "cytomegalovirus".

The syntactic pendulum searches for phrases which have a defined structure (i.e. correspond to a certain part-of-speech pattern) and consist of terms which were considered as domain specific by the system and the user one iteration step before. For a complete discussion of this algorithm and its limitations (i.e. problems of the automatic choice of parameters) we recommend a thorough examination of [WITS04].

Among the used technologies, one can use collocations of higher order (co-hyponyms) as typing candidates for associations or similarity measures for semantic features [see HEYE'01]. In addition, language detection for Topic Map generation of multilingual corpora will be useful.

5 The Framework (Green-)TOMATO

To eliminate the known shortcomings of our Topic Map generation approach we are developing the framework TOMATO. This framework has two goals: on one hand to improve the generation of Topic Maps by using the text processing algorithm described above and on the other hand to ease the integration of Topic Maps in the users' working environment and processes. TOMATO is a scientific prototype framework. Our first stable version is GreenTOMATO. Because TOMATO is an *application* of the existing Topic Map frameworks, we decided to build it on top of TM4J, the most matured and agile open source solution for Topic Map handling.

TOMATO is designed as a set of operations for the generation and integration of Topic Maps. To eliminate the monolithic approach, these operations are configurable and applicable in a great variety of ways, and they are usable via simple and flexible interfaces. TOMATO handles different corpora in parallel. Each corpus is encapsulated in a processing environment, which will be called Tomatlet. TOMATO is the container for the Tomatlets. In the following, we will describe GreenTOMATO, which includes the text processing algorithm described above, but which lacks the sophisticated Subject Identity handling demanded above. Figure 2 sketches the architecture of GreenTOMATO. In the following, the classes of operations are described shortly.

01 Generation of the Topic Maps

GreenTOMATO handles the terminology extraction process described above via a web-based interface for new corpora or additional texts in existing corpora. Topic Maps will be generated automatically from the databases delivered by this process. This process of generation is fragmented into atomic operations which can be combined in a flexible way. In addition, the structure of the produced Topic Map can

be customized. The result of the generation process is stored in a TM4J data model and a relational database backend.

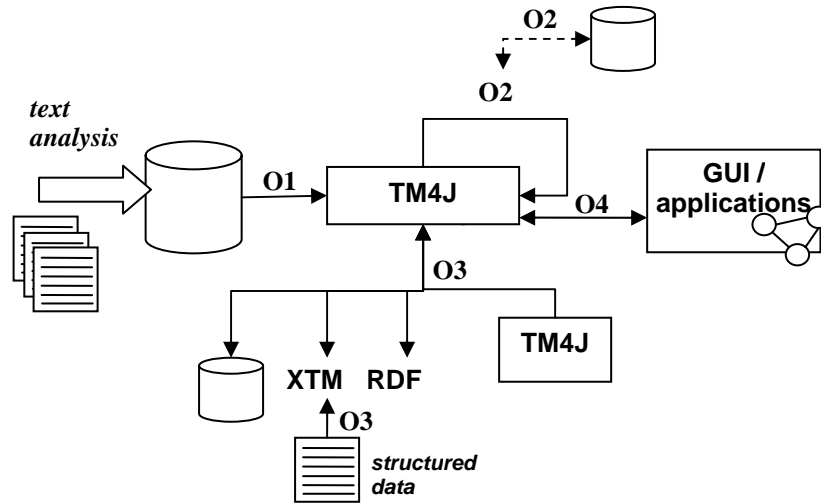


Figure 2 Operations for Topic Map Generation and Integration in GreenTOMATO

O2 Modifying Topic Maps

While users configure the Topic Maps they add some statements to it, such as new associations or new names for topics. For this use-case a basic set of atomic operations is needed, which covers these possibilities of manipulation. These operations are recorded and stored in a database. They are layered separately on top of the generated Topic Map and are only merged on demand. When the Topic Map is changed (because a document is added to the text collection), the operations done by the user can be reused, under the restriction of a test of validity. Because the application of operations is possible in a general way, external services (like automatic translation of the topic map) can be used simply.

GreenTOMATO is capable of modifying only the associations within a topic map. We decided to develop this characteristic first, since it is the most important kind of modification which can be applied to a topic map. The development endeavours were limited because we expect, with the evolution of TMQL [see <http://www.isotopicmaps.org/tmql/>] into a real retrieval *and* updating language, significant changes.

O3 Import, Export and Merging of the Topic Maps

Generated and modified Topic Maps can be exported into the standardized interchange syntax XTM (and via the Omnigator into other interchange syntaxes like RDF, LTM, HyTime, or CXTM).

In GreenTOMATO the import and merging of Topic Maps (as PSI directories or as “normal” Topic Map) will not be possible. However, the architecture of GreenTOMATO intends these features of advanced subject handling. In the future it

will also be possible to import other structured data (i.e. from databases) into TOMATO.

04 Integration in GUI and applications

In GreenTOMATO all Tomatlets can be administrated via a SWING GUI. The generated Topic Map can be viewed via a web-based interface. In TOMATO there will be well defined interfaces where applications and web-based GUIs can easily interact with the framework.

6 Conclusion and further research

We introduced our framework TOMATO for the integration of Topic Map generation in the users working environment. We illustrated the evolution of the method of automatic Topic Map generation into a more general approach, which led us to a reconstruction of the initial solution to overcome the initial outlined shortcomings. In [chapter 3] we introduced three tasks to solve if we want to integrate the underlying theory of Topic Maps in our approach. Our approach for terminology extraction is a solution for task 1 with adequate precision and recall. Except for possible enhancement of this approach, we have not solved task 2 and task 3. While task 2 might be solvable with little effort we foresee a lot of challenges for the generation of real interchangeable Topic Maps.

The use of Topic Maps is helpful for all possible information resources which can be digitalized, not only for texts. If we extend TOMATO we can open it to other data like pictures etc. The Topic Map based information access will alter the communication in organisations (see [MAIC⁺03A]), as will future technological enhancements. Discussing these influences will be very important for our aim to move Topic Maps to mainstream. On the one hand, the introduced framework will help to create real life applications where these influences can be evaluated, and on the other hand the further development of our framework will profit from these thoughts.

References

[BIEM⁺03] Biemann, Christian; Quasthoff, Uwe; Böhm, Karsten; Wolff, Christian: "Automatic Discovery and Aggregation of Compound Names for the Use in Knowledge Representations."; J.UCS - Journal of Universal Computer Science 9, 6 (2003), 530-541.

[BERN02] Berners-Lee, Tim: "What do HTTP URIs Identify?" Available at: <http://www.w3.org/DesignIssues/HTTP-URI.html>

[BLAN97] Blank, I.: „Computerlinguistische Analyse mehrsprachiger Fachtexte“; PhD Thesis, University of Munich (1999).

[BÖHM⁺02] Böhm, Karsten; Heyer, Gerhard; Quasthoff, Uwe; Wolff, Christian: "Topic Map Generation Using Text Mining."; J.UCS - Journal of Universal Computer Science 8, 6 (2002), 623-633.

[CUNN⁺03] Cunningham H., et al.: Developing Language Processing Components with GATE (a User Guide), For GATE version 2.1, <http://gate.ac.uk/sale/tao/tao.pdf>, 2003

- [FREE02] Freese, Eric: "So why aren't Topic Maps ruling the world?"; Proc. of Extreme Markup Languages; Montreal, (2002).
- [GRØN02] Grønmo, Geir Ove: "Creating topic maps from existing data sources." Available at: <http://www.ontopia.net/topicmaps/materials/automagic.html>
- [HEYE⁺01] Heyer, Gerhard; Läuter, Martin; Quasthoff, Uwe; Wittig, Thomas; Wolff, Christian: "Learning Relations using Collocations"; Proc. of the IJCAI Workshop on Ontology Learning (2001), S. 19-24.
- [HEYE⁺03] Heyer, Gerhard; Maicher, Lutz: "Persönliche und gemeinschaftliche Wissensräume. Erfüllen Topic-Maps die technologischen Anforderungen?"; Proc. LIT⁺03, Leipzig (2003), 43-54.
- [HOLS⁺02] Holsapple, Clyde W.; Joshi, K. D.: "A Collaborative Approach to Ontology Design."; Communications of the ACM 45, 2 (2002), 42-47.
- [KENT03] KENT, William: "The unsolvable identity problem."; Proc. of Extreme Markup Languages, Montreal (2003).
- [KERK03] Kerk, R., Groschupf S.: "How to Create Topic Maps"; Working report; Available at: <http://www.media-style.com/gfx/assets/HowtoCreateTopicMaps.pdf>, Halle (2003)
- [LEGR⁺02] Le Grand, Bénédicte; Soto, Michel: "Visualisation of the Semantic Web: 'Topic Map Visualisation'"; Available at: <http://csdl.computer.org/comp/proceedings/iv/2002/1656/00/16560344abs.htm>
- [MAIC⁺03A] Maicher, Lutz; Heyer, Gerhard: „Erstellung und Nutzung von Inhalten für das Semantische Web. Entwicklung eines Ordnungsschemas.“; Proc. LIT⁺03, Leipzig (2003), 112-122.
- [MAIC⁺03B] Maicher, Lutz; Heyer, Gerhard; Böhm, Karsten; Grahn, Olaf: "Automatische Erstellung individualisierter, domänenspezifischer Topic-Maps zur nachhaltigen Nutzung von Projektdokumentationen.“; Proc. KnowTech 2003, München (2003).
- [MAYN99] Maynard, D.: "Term Recognition Using Combined Knowledge Sources"; Phd. Thesis, Manchester Metropolitan University (1999).
- [OASIS03] OASIS: "Published Subjects: Introduction and Basic Requirements."; Available at: <http://www.oasis-open.org/committees/download.php/3050/pubsub-pt1-1.02-cs.pdf>
- [PEPP00] Pepper, Steve: "The TAO of Topic Maps. Finding the way in the age of infoglut"; Available at: <http://www.ontopia.net/topicmaps/materials/tao.html>
- [PEPP⁺03] Pepper, Steve; Schwab, Sylvia: "Curing the Web's Identity Crisis. Subject Indicators for RDF"; Available at: <http://www.ontopia.net/topicmaps/materials/identitycrisis.html>
- [RATH03] Rath, Holger: "Topic Maps Standard Update."; 2. Deutscher Kongress für XML Topic-Maps in der Praxis, Darmstadt (2003).
- [THOM02] Thompson, Bryan: "The Cognitive Web. Presentation to the Semantic Web Interest Group."; Available at: <http://www.cognitiveweb.org/publications/CognitiveWeb-SWIG-NASA-1.nov.2002.pdf>
- [TMDM] ISO/IEC JTC 1/SC 34:"ISO/IEC 13250. Topic Maps – Part 2: Data Model." Draft available at: <http://www.isotopicmaps.org/sam/sam-model/>
- [WITS04] Witschel, H. F.: "Text, Wörter, Morpheme — Möglichkeiten einer automatischen Terminologie-Extraktion"; diploma thesis, Leipzig (2004).
- [WÜST91] Wüster, E.: „Einführung in die allgemeine Terminologielehre und terminologische Lexikographie“; Romanistischer Verlag, Bonn (1991).