

# Subject Identification in Topic Maps in Theory and Practice

Lutz Maicher

Chair of NLP, Department of Information Science,  
University of Leipzig  
Augustusplatz 10-12, 04109 Leipzig  
maicher@informatik.uni-leipzig.de

**Abstract:** If Topic Maps should be exchanged in distributed environments a common semantic problem occurs: Do two Topics refer to the same Subject? If they describe the same Subject the given Topics have to be treated as one unit in the given context. Especially in integration scenarios this might be important. Within the Topic Map theory the merging paradigm and the description of Subjects of discourse is the central design criterion. These methods provided by the standard lead to sufficient results, but only if two distributed Topic Map authors share the same vocabulary for Subject description. But in our current research with automatic generated, hand-crafted refined Topic Maps we have to handle heterogeneous vocabularies. To solve the arising problems we introduced the Subject Identity Measure (SIM) which describes how closely related the Subjects of two distributed Topics are.

In the context of the doctoral consortium on the one hand we want to show how the omnipresent problem of Subject Identification is solved in the Topic Map theory, including the attendant shortcomings. On the other hand we want to discuss how our solution of automatic Subject comparison can be enhanced (with experiences from related technologies) and how our solution can be used to integrate the Topic Map world with other XML technologies based on the merging paradigm.

## 1 Problem

In our current research we are engaged in the automatic generation of Topic Maps<sup>1</sup> from unstructured text corpora with statistical means of NLP. Topic Maps are developed as powerful, exchangeable indexes for heterogeneous and distributed information resources.<sup>2</sup> This means, that we automatically create *indexes* for these collections of heterogeneous and distributed information resources.

---

<sup>1</sup> To avoid ambiguity all terminology concerning Topic Maps is capitalized.

<sup>2</sup> In contrast Topic Maps can be used in other application for machine based communication too. One example is the project Shark for location based Topic Maps in mobile environments [SG02].

For large corpora our approaches are nearly completely automatically [see HLQ01, BHQW02]. To get a similar precision and recall for small document collections too, we use a combination of these statistical methods, other methods derived from automatic term recognition and relevance feedback [see BMWC04]. For real application both contexts need means for further refinement on top because humans accept these automatic structures only if they have possibilities for correction [MHBG03]. In addition some foreseen ideas, like the cognitive semantic web [see Th02], bases on such handcrafted user refinements and remarks. The automatically extracted raw Topic Maps can be treated as a shared vocabulary but the user refinements lead to heterogeneous vocabularies. And they lead to new challenges.

Automatically generated, hand-crafted refined Topic Maps can be used for distributed knowledge management [see Cu03, Sc04]. With our application TOMATO [see for full detail BMWC04] distributed users can generate automatically Topic Maps from different text sources. In our application such a collection of texts, the resulting Topic Map and the included user refinements are called a Tomatlet. If the distributed users join a team (or a community of interests) they want to merge (unify) there Tomatlets.

Merging of Tomatlets leads to the issue of our contribution. Merging *is* the central paradigm of the Topic Map theory. If two distributed Topic Maps meet the process of merging must be applied. This means that if two Topics describe the equal Subject (this decision is supported by equality rules) within the same Topic Map they must be unified to one Topic (this process is defined by merging rules). In the final Topic Map only one binding point for all information concerning one Subject is allowed. Within in the Topic Map standards the fundamental design criterion “One Topic for One Subject” is well defined and applied. But problems occur if the Subjects of Topics aren’t described with a shared vocabulary. We are materially faced with this problem in TOMATO. For (content based) indexing purposes with user refinements the usage of a shared vocabulary is limited because of a low inter-indexer consistence [see FLGD87, Br90, SC04]. In this case all equality rules defined in the Topic Map have strong limitations or fail.

Summarizing we sketched the following problem to solve: Topic Maps are human centric indexes for heterogeneous and distributed information resources and are developed for interchange purposes. “One Topic for one Subject” leads to a consistent behaviour of Topic Maps in interchange scenarios, but only if Topic Map authors use a shared vocabulary for Subject Identification. But how this interchange can be applied if distributed Topic Map authors couldn’t use such a shared conceptualisation?

To solve these problems we introduced the Subject Identity Measure (SIM) which describes how closely related the Subjects of two distributed Topics are. This measure is calculated automatically from the content of each Topic. In general we foresee for the usage of Topic Maps interesting application based on their matured Subject Identification skills. This might be distributed knowledge management systems [Sc04], a fruitful cooperation with Semantic Web technologies [Gars] and the integration of unstructured and structured information in business processes. We assume that the usage of the SIM enhance these possibilities additionally.

In the context of the doctoral consortium we want to show on the one hand how the problem of Subject Identification, which is omnipresent in all integration scenarios, is solved in the Topic Map theory, including the attendant shortcomings. On the other hand we want to start a vital discussion how our solution can be enhanced (with experiences from related research areas) and how our solution can be used to integrate the Topic Map world with other XML technologies.

## 2 Some thoughts about the Topic Map Theory

The best known representation of Topic Maps is their serialization in the XML based notation XTM [see XTM]. Besides the existing (but very rarely used) HyTime<sup>3</sup> notation XTM is the kernel of the current standard. But for the 2<sup>nd</sup> and substantially matured version of the Topic Map Standard Family (which is under ISO control) the Topic Map Data Model [TMDM], which uses the XML information set model [Infoset] as meta model, will be the central theoretical kernel.<sup>4</sup> Although XTM will be the solely exchange format in the following we concentrate on the TMDM.

The main theoretical design criterion of Topic Maps is called “One Topic for one Subject”. In order to understand this criterion, we need to explain the notions of Topic, Subject and their relationship. A Topic is “a symbol used within a topic map to represent some subject, about which the creator of the topic map wishes to make statements” [TMDM]. A Subject is “anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever. In particular, it is anything on which the creator of a topic map chooses to discourse.” [TMDM] Shortly, a Topic describes a Subject (which is any possible idea or artefact of discourse) from the perception of the current Topic Map. This implies that within each Topic its Subject must be declared. Using this terminology we can describe our automatic generation methods for Topic Maps within the Topic Map standard. We have to detect interesting Subjects in the given corpora, create their representations (Topics) in a Topic Map and name them in the context of the given Topic Map. In addition relationships between Subjects must be detected automatically, represented as Associations between the regarding Topics.

Before the Subject of a Topic can be declared, the Topic Map author must be sure of the according Subject. Important philosophical questions arise: What is identifiable? What constitutes the boundaries of a thing in respect to its identity? Can identity evolve in time? Is identity situational or relative? How must properties of a thing change to alter its identity? What about versions and copies? These questions (discussed in detail in [Ke78], [Ke03]) show the limits of purely computational approaches to merging because they hardly handle indefiniteness, openness and ambiguity.

---

<sup>3</sup> The differences between these two notations are discussed in [N277].

<sup>4</sup> The Topic Map Data Model is an Application (as defined in the [TMRM]) of the Topic Map Reference Model. Because we are only interested in these Applications of the Topic Map Reference Model we are not leave it aside at the moment.

“The process of merging ensures that whenever two topics are known to represent the same subject, they are merged.” [TMDM] But how a Topic can declare its Subject? Within the TMDM two (objectively analyzable) means are implemented:

- The *Subject Locator* is used whenever the Subject of the Topic *is* an addressable information resource. In this case, the URI of this resource is used as a Subject Locator.
- Because Subjects can be anything (not only addressable resources) a Topic can declare its Subject with the help of a *Subject Indicator*, too. A Subject Indicator is an information resource which *describes* the Subject. The URI of this information resource is called *Subject Identifier*.

To obtain “One Topic for one Subject”, two Topics which have the same Subject Locator or a pair of identical Subject Identifiers have to be merged. The distinction between Subject Locators and Subject Identifiers doesn’t exist in the RDF / OWL world, but an integration might be fruitful. See as a starting point for this discussion [PS03].

These rules work well if all authors of Topic Maps have made agreements about a shared or centralised conceptualisation of the represented knowledge. These agreements are called *Published Subject Indicators* (PSI) [Oasis]. These PSIs are published (but not necessarily public) descriptions of Subjects which should be reused by as much Topic Map authors as possible to obtain a broad interoperability of their Topic Maps. Examples in the literature which discuss the merging of distributed Topic Maps (or Topic Maps and RDF documents) exclusively use PSIs [see CPV03, Gr02, Sc04]) because of the absence of solutions for open vocabularies.

However, in distributed environments with a high autonomy of the clusters, the mechanism of PSIs has its shortcomings. A PSI will only be used if it is visible to a Topic Map author. In TOMATO our automatic Topic creation leads to similar PSIs for similar Subjects even if the Tomatlets are distributed. But the user refinements and the Topics inserted by hand annoy this systematic heavily and we have to deal with heterogeneous vocabularies.

### **3 SIM – The Subject Identity Measure**

But if no PSIs are used, merging of Topic Maps becomes impossible because there will probably be no common Subject Identifiers/Locators. And this might happen even if the Topic Map authors made assertions about the same Subjects in their private Topic Maps: If the distributed authors used different Subject Indicators to indicate the same Subject, the regarding Topics, which should theoretically be merged, rest apart.

But “Merging beyond the minimal rules [defined in the TMDM] is freely allowed. Most commonly, this will be done by inferring the subject of the topics from their characteristics.” [TMDM]. Therefore, we introduced a Subject Identity Measure (SIM). The SIM describes how closely related the Subjects of two distributed Topics are, even if the authors didn’t use a common vocabulary. This measure is calculated by comparing the content of two different Topics with statistical NLP methods. See for full detail and the lasting shortcomings [MW04]. If this measure is 1 the regarding Topics definitely represent the same Subject (according to the rules defined in the [TMDM]). If the measure is 0, the regarding Topics definitely represent different Subjects. All values between 0 and 1 support a human being to decide whether two Topics represent the same Subject. All pairs of Topics that have a SIM which is higher than a certain threshold will automatically be proposed for merging.

Recapitulating, the SIM helps to propose merging of distributed Topics which represent the same Subject but where the centralised Subject Identification fails.

#### **4 Conclusion –Discussions we want to stimulate at XDWS 2004**

We described the merging paradigm of Topic Maps in detail. In addition we introduced the SIM which helps to abate shortcomings of this paradigm in distributed environments where the regarding Topic Map authors don’t use a shared vocabulary.

We foresee the merging paradigm as a central binding point between Topic Maps and other XML technologies, especially semantic technologies. We want to discuss how the idea of Subject Identification and Uniqueness is solved in other technologies and how these solutions can be used to create content based connectors between Topic Map based and other applications. Because Topic Maps are especially suitable in human centric applications and other Semantic Web technologies are especially suitable in machine interaction applications we envisage great benefits in the connection of these two worlds.

Besides we want to discuss how the SIM can be enhanced. On the one hand we are interested in ideas which especially address the structure of Topic Maps. On the other hand we are interested how the problem of automatic Subject detection and comparison is solved in other XML technologies.

#### **References**

- [BHQW02] Böhm, K.; Heyer, G.; Quasthoff, U.; Wolff, C.: Topic Map Generation Using Text Mining.”; J.UCS - Journal of Universal Computer Science 8, 6 (2002), 623-633.
- [BMWC04] Böhm, K.; Maicher, L.; Witschel, H. F.; Carradori, A.: Moving Topic Maps to Mainstream - Integration of Topic Map Generation in the User's Working Environment. J.UCS - Journal of Universal Computer Science (Springer), Volume 10, Special Issue I-Know 2004, pp. 241-251.

- [Br90] Bruza, P. D.: Hyperindices: a novel aid for searching in hypermedia. In: Proceedings of the 1990 ACM Hypertext, pp. 109-122 (1990).
- [CPV03] Ciancarini, P.; Pirruccio, M.; Vitali, F. et al.: Metadata on the Web. On the integration of RDF and Topic Maps. In: Proceedings of "Extreme Markup Languages 2003", Montreal, 2003.
- [Cu03] Cuel, R.: A New Methodology for Distributed Knowledge Management Analysis. Proceedings of I-KNOW '03, Graz, (2003), 531-537.
- [FLGD87] Furnas, G. F., Landauer T. K., Gomez L. M., Dumais S. T.: The Vocabulary Problem in Human-System Communication. Communications of the ACM (CACM), 30, pp. 964-971, (1987).
- [Gars] Garshol, L. M.: Living with topic maps and RDF. Topic maps, RDF, DAML OIL, OWL, TMLC. Available at: [www.ontopia.net/topicmaps/materials/tmrdf.html](http://www.ontopia.net/topicmaps/materials/tmrdf.html)
- [Gr02] Grønmo, G. O.: Automagic Topic Maps. Available at: <http://www.ontopia.net/topicmaps/materials/automagic.html>
- [HLQ01] Heyer, G.; Läuter, M.; Quasthoff, U.; et al.: Learning Relations using Collocations; Proc. of the Workshop on Ontology Learning (2001), pp. 19-24.
- [Infoset] Cowan, J.; Tobin, R.: XML Information Set (Second Edition). W3C Recommendation Verfügbar unter: <http://www.w3.org/TR/xml-infoset/>.
- [Ke78] Kent, W.: Data and reality. Basic Assumptions in Data Processing Reconsidered. North-Holland Publishing, Amsterdam, New York, Oxford (1978).
- [Ke03] Kent, W.: The unsolvable identity problem." In: Proceedings of "Extreme Markup Languages 2003", Montreal, (2003).
- [MHBG03] Maicher, L.; Heyer, G.; Böhm, K.; Grahn, O.: Automatische Erstellung individualisierter, domänenspezifischer Topic-Maps zur nachhaltigen Nutzung von Projektdokumentationen. Proceedings of KnowTech 2003.
- [MW04] Maicher, L.; Witschel, H. F.: Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM). To appear in: Proceedings of Leipziger Informatiktage (LIT) 2004.
- [N277] ISO/IEC JTC 1/SC34: Differences between XTM 1.0 and the HyTime-based meta-dtd. Available at: <http://www.y12.doe.gov/sgml/sc34/document/0277.htm>.
- [Oasis] OASIS: "Published Subjects: Introduction and Basic Requirements." Available at: <http://www.oasis-open.org/committees/download.php/3050/>
- [PS03] Pepper, S.; Schwab, S.: Curing the Web's Identity Crisis. Subject Indicators for RDF. Available at: <http://www.ontopia.net/topicmaps/materials/identitycrisis.html>
- [SG02] Schwotzer, T.; Geihs, K.: Shark – a System for Management, Synchronization and Exchange of Knowledge in Mobile User Groups. Proceedings of I-KNOW '02, Graz, (2002).
- [Sc04] Schwotzer, T.: Modelling Distributed Knowledge Management Systems with Topic Maps. J.UCS - Journal of Universal Computer Science (Springer), Volume 10, Special Issue I-Know 2004, pp. 53-60.
- [SC04] Stumpf, S.; Zini, C.: An Investigation into Sharing Metadata: "I'm not thinking What you are thinking." J.UCS - Journal of Universal Computer Science (Springer), Volume 10, Special Issue I-Know 2004, pp. 252-260.
- [Th02] Thompson, B.: The Cognitive Web. Presentation to the Semantic Web Interest Group, 07.04.2003. 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." Latest version available at: <http://www.isotopicmaps.org/sam/>
- [TMRM] ISO/IEC JTC 1/SC34: Topic Maps – Reference Model. Editor's Draft, Revision 3.1. 01.12..2003. Available at: <http://www.isotopicmaps.org/TMRM/TMRM-latest-clean.html>
- [XTM] TopicMaps.Org Authoring Group: XML Topic Maps (XTM) 1.0. Available at <http://www.topicmaps.org/xtm/1.0/>.