

# Subject Identification in Topic Maps in Theory and Practice

Lutz Maicher

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 represent the same Subject? If they describe the same Subject the according Topics have to be merged. Within the Topic Map theory the merging paradigm and the description of Subjects is the main theoretical design criterion. Normally, these methods provided by the standard lead to sufficient results, but only if distributed Topic Map authors share a common vocabulary for Subject description. To solve the arising problems for distributed, autonomous environments we introduce the Subject Identity Measure. The SIM describes how closely related the Subjects of two distributed Topics are. The approach is independent from a shared vocabulary, from a specific natural languages and uses only data which is available inside the according Topic Maps.

## 1 Problem - Subject Identification in distributed Topic Maps

A Topic is a binding point for all information concerning *one* Subject within a Topic Map. A Subject is anything of the real world where an author wants to discourse about within his Topic Map. The main theoretical design criterion of Topic Maps is called “One Topic for one Subject”. This means, if two Topics describe the equal Subject (this decision is supported by equality rules) within the same Topic Map they must be merged to one Topic (this process is defined by merging rules). “One Topic for One Subject” is well defined and applied within the Topic Map standards.

Problems occur if the Subjects of Topics aren’t described with a shared vocabulary; especially in distributed environments. In these cases the defined equality rules have strong limitations or fail because Subjects are only defined as identical if they are described with identical strings. We don’t share the optimism, that centralised repositories for Subject description (so called PSI repositories) will be widely adopted. Rather we expect that the sameness of two Subjects can be inferred from the content of its Topics. Therefore we introduce the Subject Identity Measure which describes how closely related the Subjects of two Topics are. The level of the SIM supports humans or machines to decide whether these Topics should be merged or not.

In general, we foresee interesting applications for the usage of the SIM. Distributed knowledge management might be the main application [see Cu03, Sc04, and Si04]. Topic Maps are part of the Semantic Web efforts [see Th02] and are translatable into RDF or OWL [discussed in Pepp, Gars, and PS03]. This enables the reuse of the SIM approach in a variety of Semantic Web applications. At least, the SIM approach can be used for the integration of unstructured and structured information in business processes.

In this paper we are making the following contributions:

- We discuss the arising problems of Topic Maps' central theoretical criterion „One Topic for one Subject“ (see section 2).
- We describe the Subject Identity Measure approach to address these problems. Additionally, we assess its quality yielded for a testbed in brief (see section 3).
- We sketch the challenges of further research (see section 4).

## 2 Subject Identification in the Topic Map Theory

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” [TMDM]. Shortly, a Topic describes a Subject (which is anything on which the creator of a Topic Map chooses to discourse) from the perception of the current Topic Map. This implies, within each Topic the Subject must be declared.

While declaring Subjects, 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 pure naming approaches because they hardly handle indefiniteness, openness and ambiguity [see FLGD87].

But how a Topic can declare its Subject? Within the Topic Map Data Model (TMDM) two means are implemented which are more or less pure naming approaches:

- 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. The URI names the Subject.
- 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 (which names the Subject) of this information resource is called *Subject Identifier*.

To obtain “One Topic for one Subject”, two Topics having the same Subject Locator or a pair of identical Subject Identifiers have to be merged. These rules work well if all authors of Topic Maps have made agreements about a shared vocabulary for Subject naming. These agreements are called *Published Subject Indicators (PSI)* [Oasis]. 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 Topic Maps. Examples in the literature which discuss merging of distributed Topic Maps (or Topic Maps and RDF documents) exclusively use PSIs [see CPV03, Gr02, Sc04]. This is due to the absence of solutions for open vocabularies.

However, in distributed environments with a high autonomy, the mechanism of PSIs has its shortcomings. PSIs are only used if they are visible to the regarding Topic Map authors. Additionally, PSIs are faced with the philosophical problems of naming approaches discussed above. In contrast to the naming approach we follow up an description approach. We assume that a Subject is indirectly determined by the content of its Topic. We don't name or stringently delimit a Subject, we only decide whether two Subjects are quite identical. The level of this “identity” supports humans or machines to decide, whether these Topics should be merged. If they chose merging, these Topics get an identical Subject Identifier to apply the merging inside the Topic Map standards.

### 3 The Subject Identity Measure (SIM) Approach

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 propose a Subject Identity Measure.

*The SIM describes how closely related the Subjects of two Topics are.* If the SIM is 1 the regarding Topics definitely represent the same Subject (according to the rules defined in the TMDM). If 0, the regarding Topics definitely represent different Subjects. All values between 0 and 1 support the decision whether two Topics represent the same Subject.

Whenever two Topic Maps meet, the SIM Approach performs the following steps:

1. *Calculation.* The SIMs for TopicNames, Occurrences, and Subject Indicators for each pair of Topics must be calculated.
2. *Filtering.* According to different thresholds and coefficients, the overall SIM will be calculated. For each Topic, a suitable counterpart in the other Topic Map will be chosen (but only if there is a SIM greater than 0).

**For calculation** only data inside each Topic is used, whereby structural (types, associations etc.) or external (content of information resources referenced from the Topic Map) information is left aside. We need similarity measures for URIs (for Subject Indicators, VariantNames and OccurrenceLocators) and strings (all TopicNames and OccurrenceData). For the calculation of the similarity of a pair of two strings ( $S1, S2$ ) we used a language and context independent measure  $c(S1, S2) \rightarrow [0, 1]$ . For more detail and the discussion of the calculation of the URI similarity we refer to [MW04].

In general, the approach inspects each possible pair of Topics ( $T1, T2$ ) where  $T1$  and  $T2$  belongs to two different Topic Maps. For each pair ( $T1, T2$ ) the measures  $SIM.Names$  and  $SIM.Occurrences$  are calculated as follows:

1. “Fillet” each Topic for Names. Take all property values from the property “value” of all Topic Name Items and Variant Items of  $T1$  and store them in a set  $Nam1$ . To get  $Nam2$  do the same for  $T2$ .
2. “Fillet” each Topic for Occurrences. Take all property values from the property “value” of all Occurrence Items of  $T1$  and store them in a set  $Occ1$ . Do the same for  $T2$ .
3. Calculate  $SIM.Names$ . If  $|Nam1| < |Nam2|$ , and  $|Nam1|=m$  then:
$$SIM.Names = \frac{1}{|m|} \sum_{n1 \in Nam1} \max_{n2 \in Nam2} s(n1, n2)$$
4. Calculate  $SIM.Occurrences$ . Do the same for  $Occ1$  and  $Occ2$ .

**For Filtering**, the  $SIM(T1, T2)$  of each pair of Topics ( $T1, T2$ ) is calculated as follows (whereby  $\lambda$  indicates whether TopicNames or Occurrences are more important):

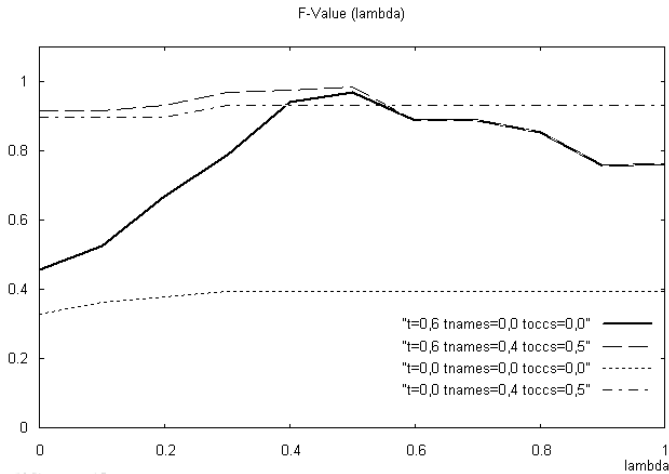
$$SIM = \lambda SIM.Names + (1 - \lambda) SIM.Occurrences$$

But the  $SIM$  is set to 0 if one of three thresholds isn't exceeded.  $t_{Name}$  has to be exceeded by  $SIM.Names$  to avoid overrating of  $SIM.Occurrences$ .  $t_{Occ}$  has to be exceeded by  $SIM.Occurrences$  to avoid overrating of  $SIM.Names$ . And finally,  $t$  has to be exceeded by  $SIM$  to avoid a great number of false positives.

In our case, we use the application-specific constraint that inside a Topic Map each Subject is represented by only one Topic. We iterate over the smaller Topic Map. For each Topic  $T1$ , we choose a merging candidate  $T2$  such that  $T2$  is the maximum over all  $SIM(T1, Ti)$ . If all  $SIM(T1, Ti)$  are 0, the  $T1$  remains without a merging candidate.

For the assessment of the matching quality we needed Topic Maps which describe fully qualified Subjects with uncontrolled vocabularies. We decided to use the online catalogues of two libraries, the German Library (“DDB”, <http://www.ddb.de>) and the compound catalogue of a network of German libraries (“GBV”, <http://gso.gbv.de/>). From both we extracted all entries which represent publications of Springer in 1997 (DDB 1.800, GBV 2.700). We automatically created Topic Maps from these datasets where each is represented by a Topic. Properties like “title” are transformed to Basenames, other properties (e.g. “keywords” or “editor”) are treated as typed Occurrences. We obtained a Topic Map without Associations. For each Topic, we put aside the ISBN or ISSN as an objective criterion helping to decide whether two publications in the two catalogues were indeed identical.

For assessment purposes we used three quality metrics. *Precision* tells how many of the merging candidates proposed by the SIM approach are really identical. *Recall* tells how many of the existing identical pairs were found by SIM. *F( $\beta$ )-Value* is a combination of precision and recall that yields high values only if both are high. We decided to set  $\beta=2$  to make Recall twice important as Precision.



**Figure 1** F-Value ( $\lambda$ ) for  $(t=0.0; t_{Name}=0.0; t_{Occ}=0.0)$ ,  $(t=0.0; t_{Name}=0.4; t_{Occ}=0.5)$ ,  $(t=0.6; t_{Name}=0.0; t_{Occ}=0.0)$ ,  $(t=0.6; t_{Name}=0.4; t_{Occ}=0.5)$

Figure 1 shows some results from the assessment process. It shows that  $\lambda=0,5$  yields the best results. This implies that  $\lambda$  can be eliminated. For the full discussion we refer to [MW04]. Summarising, we sketch the following results [see MW04 for full detail]:

- The SIM approach yields good results for both, recall and precision.
- We propose the usage of  $t$ ,  $t_{Names}$ , and  $t_{Occ}$ .
- If  $t_{Names}$  and  $t_{Occ}$  are chosen carefully, the usage of  $t$  isn't necessary.
- $\lambda$  doesn't augment the quality of the SIM approach and can be eliminated.

## 4 Conclusion and Further Research

We introduced and discussed the problem of Subject Identification in distributed environments. As corollary, we proposed the SIM approach to solve these problems. We state, that for the testbed our approach yields very good results. We want to emphasize, that our approach is independent from a shared vocabulary, from a specific natural language, and uses only data which is available inside the given Topic Maps.

In general, we consider the SIM approach as a lightweighted "sparring partner" for more advanced matching techniques. Therefore our further research comprises:

- *Topic Map Metrics.* We need objective criteria of Topic Maps (closeness, etc.) to calculate correlations between matching quality and the used testbeds.
- *Adoption of promising approaches to Topic Maps.* We are going to adopt matching techniques proposed in other contexts to Topic Maps. We envisage the Similarity Flooding Approach proposed by Melnik et al.[MGR02] and a matching technique proposed by Castano et al. [CFM04].
- *Advanced testbeds.* For quality assessment purposes we need additional testbeds.
- *Quality measures.* We have to discuss whether precision, recall and F-Value are the most suited quality measures in our problem scenario.
- *Improvement of the SIM approach.* Based on the yielded results we have to improve the proposed SIM approach, especially by using structural information.

## Acknowledgements

Thanks to Hans Friedrich Witschel for his substantial contribution to this work.

## References

- [CFM04] Castano, S.; Ferrara, A.; Montanelli, S.; Racca, G.: Matching techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions. In: Proceedings of “International Conference on Coding and Computing (ITCC04)”, IEEE Computer Society, Las Vegas, (2004).
- [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>
- [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).
- [MGR02] Melnik, S.; Garcia-Molina, H.; Rahm, E.: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: Proceedings of “18th International Conference on Data Engineering (ICDE'02)”, San Jose, California, (2002).
- [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”, Leipzig, (2004).

- [Oasis] OASIS: Published Subjects: Introduction and Basic Requirements. Available at: <http://www.oasis-open.org/committees/download.php/3050/>
- [Pepp] Pepper, S.: Ten Theses on Topic Maps and RDF. Available at: <http://www.ontopia.net/topicmaps/materials/rdf.html>
- [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>
- [Si04] Sigel, A.: *kPeer* as a Context-Aware Topic Map P2P Application for the Distributed Integration of Knowledge. Submitted to MRC 2004. Available at: <http://kpeer.wim.uni-koeln.de/~sigel/>
- [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, (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, (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>