

Distributed Knowledge Management in the Absence of Shared Vocabularies

Lutz Maicher

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

Thomas Schwotzer

(University of Technology, Berlin, Germany
thsc@cs.tu-berlin.de)

Abstract: Distributed Knowledge Management Systems (DKMS) are often faced to heterogeneous environments associated with the absence of shared vocabularies. DKMSs realise Knowledge Flows between autonomous Knowledge Nodes as parts of social networks. Schwotzer models the KNs' individual policies for input relevance and output strategy as Knowledge Ports. Topic Map Technologies are well suited for the semantic integration of distributed, heterogeneous knowledge. But current implementations base on pure naming approaches to Subject Identity in connection with the use of shared vocabularies. Maicher's SIM Approach helps to use Topic Map Technologies for the semantic integration of distributed, heterogeneous knowledge in the absence of shared vocabularies. To detect Subject similarity it exploits the Topics' usage in the current context. Our contribution is the liaison of the Knowledge Port Approach and the SIM Approach. This leads to DKMSs which significantly better deal with the absence of shared vocabularies.

Keywords: Distributed Knowledge Management, Ontology, Topic Map Technologies, Semantic Integration, Knowledge Port Approach, Subject Identity Measure Approach (SIM)

Categories: C 2.4, H 1.2, H 4.2, H 4.3, I 2.4

1 The Challenge in an Example

A R&D employee examines a new technology like RFID. She collects research reports, scientific articles and patent specifications. After a while her company decides to join an accordant research project with partners from industry and universities. How to share the knowledge between all partners across organisational and spatial borders? The solution might be a Distributed Knowledge Management System¹ (DKMS). It is shaped according to the arising social networks within the research project and allows decentralised knowledge exchange between all peers. The R&D employee determines her input and output policies for knowledge exchange. The DKMS supports to exchange the right knowledge with the right people in the project network.

But two critical points arise: How to request knowledge from remote peers if shared vocabularies for knowledge description are not available? How to integrate

¹ To avoid ambiguities terminology anent the Knowledge Port Approach and Topic Maps is capitalised.

knowledge gained from remote peers into the personal knowledge base? This article proposes a solution based on Topic Map Technologies. This solution allows the R&D employee to exchange knowledge within the research project without the overhead of centrally enforced vocabularies.

2 Distributed Knowledge Management Systems, Topic Map Technologies and the Absence of Shared Vocabularies

Often the output of classical knowledge management projects is a centralised Knowledge Management System (KMS), which solely can be *accessed* by users in a distributed, decentralised way. These inherently centralised approaches tend to ignore that knowledge is usually distributed in and among complex knowledge-based organisations. Distributed KMSs try to overcome these limitations by splintering a centralised KMS into a network of cooperating Knowledge Nodes (KN). Introduced by Bonifacio, Bouquet and Cuel [BBC02, Cu03] a Knowledge Node is an abstraction of formal (e.g. division) or informal (e.g. community of practice) organisational units which are parts of social networks. Knowledge Flows (KF) take place between these nodes.

Schwotzer [Sc04] showed how a DKMS can be realised with Topic Map Technologies. The international industry standard "Topic Maps" is well suited for such knowledge exchange and integration problems. Schwotzer introduced Knowledge Ports (KP) which realise these Knowledge Flows between KNs. The behaviour of these ports is defined by the current context (interests, location etc.) and the KN's individual policy for input relevance and/or output strategy. As described in ch. 3 all these parameters are completely described with the help of Topic Maps. Besides the context and the strategy the knowledge of a peer is stored in a Topic Map, too. The therefore needed creation of Topic Map Views about the local information is out of this article's focus [see Ba04, BMWC04 for more information].

Schwotzer's Knowledge Port Approach bases on the exchange of Topic Maps in distributed, heterogeneous environments. For environments where all communication parameters of a KF (contexts and exchange policies) can be described by a shared vocabulary (ontology) Topic Map Technologies provide a well defined and implemented fundament for the semantic integration of distributed, heterogeneous knowledge. The underlying theory discussed in ch. 4 is called "One Topic for One Subject".

Alternative approaches for the exchange of Topic Maps are proposed (and implemented), too. These are Ontopia's Topic Map Remote Access Protocol [PG04], Ahmed's TMSHare [Ah03] and Barta's Federated Topic Map Approach [Ba04]. These technologies are suited for DKMS. But in the absence of shared vocabularies they suffer from the same limitations in semantic integration scenarios as discussed below.

Problems occur in environments where shared vocabularies are not available for any reasons. The flexibility and newness of knowledge based processes often cause such an absence of shared vocabularies, especially in the case of interaction of autonomous entities [FN⁺05]. Maicher introduced the Subject Identity Measure (SIM)

Approach to allow a Topic Map based semantic integration in the absence of shared vocabularies [MW04, Ma04]. Its further developments are introduced by this article. Though, the SIM Approach completely relies on the Topic Map Theory.

Our contribution is the liaison of the KP Approach and the SIM Approach. The SIM Approach allows the exchange and integration of distributed, heterogeneous Topic Map Fragments in the absence of shared vocabularies. Because of that KNs can request knowledge of interest from remote peers without having any information about the used vocabulary in distance. Proven by first empirical results, the yielded DKMSs would significantly better deal with the absence of shared vocabularies.

In detail, this article makes the following contributions:

- Description of the usage of Topic Map Technologies for DKMS,
- Introduction of the SIM Approach for the exchange of Topic Maps in the absence of shared vocabularies, and
- Liaison of the Knowledge Port and SIM Approach.

3 DKMS, Knowledge Flows and the Knowledge Port Approach

Bonifacio, Bouquet and Cuel introduced the idea of modelling DKMS with the help of Knowledge Nodes [BBC02]. Each KN represents the abstraction of a formal or an informal organisational unit. Individuals, groups (communities of interests) and (virtual) enterprises are represented by KNs which act as *autonomous* entities in dynamic, social networks. Between these KNs several communication channels realise Knowledge Flows. Schwotzer proposes the Knowledge Port Approach for the implementation of these KFs in DKMS [Sc04] (see Figure 1).

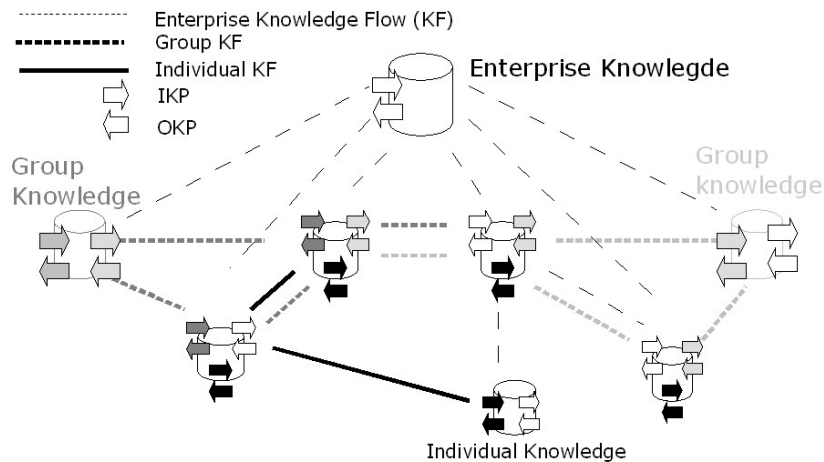


Figure 1 Knowledge Flows in DKMS

The approach bases on Knowledge Ports. Figure 2 sketches a Knowledge Port which is described in full detail in [Sc04]. A KP describes a Knowledge Node's policy of input relevance and/or output strategy for participating in Knowledge Flows.

It is governed by the insight that social systems reduce complexity of input streams from environment based on relevance filters and of output streams based on strategy policies [Wi02, SH03]. Both are described inside the KPs with the help of Topic Maps. A KF between two KPs will be realised if their knowledge exchange policies matches.

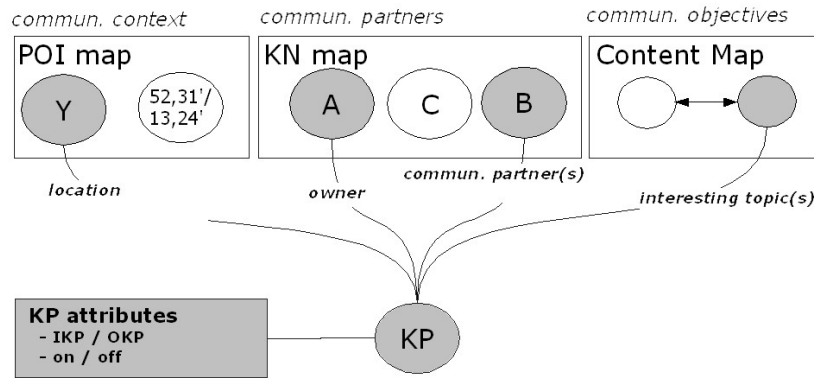


Figure 2 Parameters of a Knowledge Port

So far all communication parameters (context, partners, objectives and attributes) are defined by a shared vocabulary (PSIs) with the help of a Knowledge Exchange Protocol (KEP) [SH03]. While this might be inevitable for the description of a location or a defined person, for the description of the communication objective (the knowledge of interest) a PSI based approach often leads to avoidable empty matches.

It is required a solution which allows a KN the submission of the current interest by sending the according Topic from its Content Map. The remote peers extract similar Topics from their Content Map as the result of the request which have to be integrated by the requesting KN in its Content Map. The SIM Approach fits these requirements in the absence of shared vocabularies.

4 The Topic Map Theory and Topic Map Exchange

The KP Approach proposed in [Sc04] does exploit the main Topic Map Theory which is called “One Topic for one Subject”. 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 in the context of the current Topic Map. This implies that each Topic has to declare its Subject. The Subject is the identity of a Topic on which semantic integration takes place.

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 do properties of a thing have to 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.

According to the Topic Map Data Model [TMDM] Subjects are explicitly declared. We call this a pure naming approach to Subject Identity:

- 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 of this information resource is called *Subject Identifier*.

To obtain “One Topic for one Subject”, two Topics having the same Subjects (a pair of identical Subject Identifiers or Subject Locators) have to be merged. These rules work well if all authors of *all* 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. However, in heterogeneous, 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.

Examples in the literature which discuss the merging of distributed Topic Maps (or Topic Maps and RDF documents) exclusively use PSIs [see CPV03, Sc04, Ba04]. This is due to the absence of solutions for not-shared vocabularies.

All of these approaches work similarly. A requesting peer sends a Topic to the remote peers. This Topic names the Subject of interest. The remote peers check whether a Topic with an identically named Subject is available. In this case, a Topic Map Fragment around the according Topic is send to the requesting peer.

5 The SIM Approach and the Integration of Distributed, Heterogeneous Knowledge

In contrast to the pure naming approach to Subject Identity discussed above Maicher proposed the SIM Approach [MW04, Ma04] which implements a descriptive approach to Subject Identity. Within this paper a further developed SIM is introduced. The SIM bases on the assumption that a Subject is indirectly determined by its Topic's usage in the current context: the content of this Topic and the surrounding of that Topic. In contrast to naming approaches the SIM doesn't name or stringently delimit a Subject. Solely it decides whether two Subjects might be treated as identical, deduced from their Topics' similar usage in the current context.

According to the SIM Approach a requesting peer sends a Topic Map Fragment around the Topic of interest to the remote peers. These peers check the availability of Topics which are similarly used. In this case, it is assumed that both Topics represent sufficiently similar Subjects in the current context.

We have to underline that using the Topic Map Reference Model [TMRM] the implemented descriptive approach to Subject Identity does completely rest inside the Topic Map Theory, but coevally extends the current approach applied in the [TMDM].

5.1 The SIM Approach – An Overview

For brevity, in the following the SIM Approach is introduced in limited detail. The SIM Approach bases on the assumption that if both Topics from a requesting and a requested Topic Map similarly interact with similar Topics, the probability of the similarity of these Topics in the current context increases, too. Roughly, this assumption is derived from Melnik's et al. insights from schema matching [MGR02].

The requesting Topic will be called T. F is the Fragment of the requesting Content Map around T. Shortly, this fragment consists of all Topics and Associations which are influenced by T. After the receiving of F the remote peer compares each Topic from F with each Topic of its own Content Map in two levels. The first level does not comprise the similarity of Topics in the environment, whereby in the second level the similarity calculated in the first level is exploited. After the second level, T's most similar Topic from the requested Topic Map is outputted.

Two Topics are compared as follows. Each Topic has a state of interaction with its environment which we will call simDNAType. For example, the simDNAType "x13tn" characterises a typed Topic having a Base Name, a Source Locator and a Subject Identifier which is used for typing purposes in one other Topic of the given Topic Map Fragment. A Topic's simDNAType is valid according the following regular expression:

`/x*y*z*w*s*1*2*3*t*n*(\o)*(\a\)*`

x,y,z,w – the Topic is typing a Topic (x), an Association (y), a Topic Characteristic (z), or an Association Role (w)

s – the Topic is scoping a Topic Characteristic

1,2,3 – the Topic has a Source Locator (1), a Subject Locator(2), or a Subject Identifier (3)

t – the Topic is typed

n – the Topic has a TopicName

o => /(v|l)t?s*/ – the Topic has an Occurrence (with OccDNAType)

a => /a(tp)*/ – the Topic takes part in an Association (with AssDNAType)

The similarity of a pair of Topics called simDNA is calculated for each digit of the simDNAType. The simDNAType of the *requesting* Topic constrains the simDNA of this pair. For example, in the first level a digit of type "t" can have the values "X" and "1". "X" specifies that the requested Topic is not typed, "1" specifies that the requested Topic is typed, too. In the second level the value "3" is attainable and

specifies that the typing Topic of the requested Topic and the typing Topic of the requesting Topic gained sufficient similarity in level 1.

For each digit of the simDNAtype similar rules are defined. The higher the sum of digits of a simDNA, the higher is the similarity of two Topics. For each requesting Topic, the requested Topic with the highest sum of digits should be the response of a request.

5.2 Evaluation of the SIM Approach

For brevity, only some insights from the evaluation are given. Imagine a Topic Map which is requested by its own Topics. This test we call self assessment. For each requesting Topic the SIM Approach has to response with its "twin" in the requested Topic Map. If for all Topics the twins are returned the recall is 1. The question is the behaviour of the SIM Approach if the requesting Topic and its submitted surrounding are pruned randomly? What happens if only 40 percent of all Names and 60 percent of the Associations are left in the submitted fragments? What happens if all Names and all Associations are pruned in the submitted fragments?

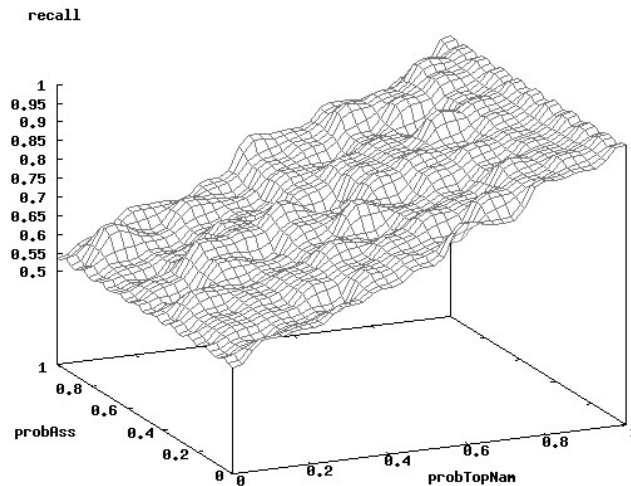


Figure 3 Iteration: probTopName [0,1] , probAss [0,1]

Figure 3 shows the result of an experiment with a small Topic Map of 20 Topics. The probability of non-pruning Topic Names (probTopNam) and Associations (probAss) is iterated in the interval [0,1]. To yield statistically firmed results the calculated recall is the mean of 10 self assessments.

As already predicted, if probTopNam and probAss are 1 the recall is 1, too. But, if both probabilities are 0, the recall is still 0.53. This implies, even if *all* Topic Names and *all* Associations are pruned, the typing information (typing of Topics, typing of Occurrences etc.) and the information inside the Occurrences are sufficient to get the half of all Topics correctly. One has to bear in mind that the algorithm

neither has knowledge about the used vocabulary (including any used Subject Locators and Subject Identifiers) nor about the human languages used in the Occurrences and Topic Names. In addition, the results already drastically improve if only some typing information bases on shared vocabulary.

This experiment sketches the abilities of the SIM Approach for DKMS. In the absence of shared vocabularies Subjects can be requested by submitting Topic Map Fragments which represent the Subject in interest.

6 Summary and Outlook

We showed the usage of Topic Map Technologies for DKMS. Additionally, we discussed its limitations in the absence of shared vocabularies.

Derived from these limitations we introduced the SIM Approach, which abandons as a descriptive approach to Subject Identity from pure naming approaches to Subject Identity. In liaison with the Knowledge Port Approach, the SIM helps to implement DKMS for distributed, heterogeneous environments in absence of shared vocabularies. Our empirical results show the positive effect of that liaison.

The SIM Approach is similarly usable for other techniques of Topic Map Fragment Exchange [PG04, Ba04]. Caused by considerable intersections to challenges in the field of DKMS, the gained insights about the exchange and integration of distributed knowledge in the absence of shared vocabularies should be exploited in Enterprise Information Integration (EII) scenarios, too [Ba04], [PG04].

While the first empirical results are encouraging, the proposed liaison of both approaches has to be tested in real life applications. We foresee that the SIM Approach has to be customised for the varying contexts. It is crucial, that it *always* rest on the Topic Map Theory which is the fundament of the semantic integration.

As a long-term vision, the resulting DKMS may help to realise the vision of a cognitive web as the human centric layer of the Semantic Web [Th02].

Within the research project of the cover example, the R&D employee will get a DKMS which allows the exchange of knowledge without the overhead of defining shared vocabularies. She gains productivity and flexibility: switching between projects will be drastically simplified.

References

- [Ah03] Ahmed, K.: TMSHare – Topic Map Fragment Exchange in a Peer-to-Peer-Application. In: *Proceedings of XML Europe 2003*, London, (2003).
- [Ba04] Barta, Robert: Virtual and Federated Topic Maps. In: *Proceedings of XML Europe 2004*, Amsterdam (2004).
- [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. In: *Proceedings of I-KNOW '04*, Graz, (2004), pp. 241-251.

- [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).
- [BBC02] Bonifacio, M.; Bouquet, P.; Cuel, R.: Knowledge-Nodes: the Building Blocks of a Distributed Approach to Knowledge Management. In: *Proceedings of I-KNOW '02*, Graz, (2002), pp. 191-200.
- [Cu03] Cuel, R.: A New Methodology for Distributed Knowledge Management Analysis. In: *Proceedings of I-KNOW '03*, Graz, (2003), pp. 531-537.
- [FN⁺05] Fröhner, T.; Nickles, M.; Weiß, G.; Brauer, W.; Franken, R.: Integration of Ontologies and Knowledge from Distributed Autonomous Sources. In: *Künstliche Intelligenz*, No. 1/2005, pp. 18-23, (2005).
- [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).
- [Ma04a] Maicher, L.: Subject Identification in Topic Maps in Theory and Practice. In: Tolksdorf, Eckstein (eds.): *Berliner XML-Tage 2004*, Berlin (2004).
- [MGR02] Melnik, S.; Garcia-Molina, H.; Rahm, E.: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: *Proceedings of the 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). In: *Proceedings of Leipziger Informatiktage*, Leipzig (2004).
- [Oasis] OASIS: Published Subjects. Introduction and Basic Requirements. Available at: <http://www.oasis-open.org/committees/download.php/3050/>
- [PG04] Pepper, S.; Garshol, L. M.: Seamless Knowledge – Spontaneous Knowledge Federation using TMRAP. Presentation at: *XML Europe 2004*, Amsterdam (2004). Available at: http://www.ontopia.net/topicmaps/learn_more.html
- [SH03] Schulz, S.; Herrmann, K.; Kalcklösch, R.; Schwotzer, T.: Towards Trust-based Knowledge Management in Mobile Communities. In: *Proceedings of AAAI Spring Symposium on Agent-Mediated Knowledge Management*, Stanford, (2003).
- [Sc04] Schwotzer, T.: Modelling Distributed Knowledge Management Systems with Topic Maps. In: *Proceedings of I-Know '04*, Graz (2004), pp. 53-60.
- [Th02] Thompson, B.: The Cognitive Web. Presentation to the Semantic Web Interest Group. Available at: <http://www.cognitiveweb.org/publications/>
- [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>
- [Wi02] Willke, H.: Systemtheorie I: Grundlagen. 6th revised edition, Lucius & Lucius, Stuttgart (2000).