To appear 2006 in:
Maicher, Lutz & Park, Jack (eds.): "Charting the Topic Map Research and Applications Landscape".
Proceedings of TMRA'05 – International Workshop on Topic Map Research and Applications: Leipzig, Germany, October 6-7, 2005; LNCS 3873, © Springer-Verlag, see http://www.springeronline.com/lncs

Topic Map Exchange in the Absence of Shared Vocabularies

Lutz Maicher

University of Leipzig, Department of Computer Science,
Augustusplatz 10-11, D-04109 Leipzig, Germany
maicher@informatik.uni-leipzig.de

Topic Maps are the international industry standard for semantic information integration. Appropriate means for Topic Map exchange are crucial for its success as integration technology. Topic Map exchange bases on the governing Subject Equality decision approach, the decision whether two Subject Proxies indicate identical Subjects. This paper discusses the 'absence of shared vocabularies' in the context of these decisions. Thereby, a differentiation between Referential and Structuralist Subject Equality decision approaches is introduced. All existing approaches to Topic Map exchange base on the TMDM. This implies a Referential Subject Equality decision approach and bound to a concrete Subject Map Disclosure (SMD) ontology and Subject Map (SM) vocabulary. This paper introduces a Structuralist Subject Equality decision approach which is called SIM. It allows the exchange of Topic Maps in the absence of a shared SM ontology and SM vocabulary.

1. The challenge in an example

Within a cooking peer-to-peer network remote peers exchange recipes documented as Topic Maps[1]. To collect information, peers send Topics which represent the Subjects of interest to remote peers. In the cooking network a Subject might be 'roasted lamb loin'. The remote peers check the availability of information about this Subject and respond with an according Topic Map Fragment. Afterwards, the requesting peer integrates all remote recipes about roasting lamb loins into its local recipe collection.

This works fine if all peers made agreements about how to describe lamb cuts correctly. What happens if a remote peer uses the term lamb saddle instead? Or roasted lamb leg chops? The resulting meals are identical, but the requesting peers will never receive their recipes from distance. This shows that two critical points arise, if semantic agreements are not made by all peers logging into the network: How to request knowledge from remote peers if shared vocabularies are not available? How to integrate (merge) the received information into the local Topic Map?

The solution proposed in this paper allows peers to interact in networks without having the overhead of centrally enforced vocabularies. Our solution detects

[1] To avoid ambiguities all terminology concerning Topic Map Technologies is capitalised.

similarity between Subjects through the similar usage of their proxies. Even if lamb lag chops and lamb loin are represented by different Subject Proxies, in recipe collections these proxies will be used similarly: with bean and rosemary proxies, etc. And the chef will cook roasted lamb loins according to this very good traditional French recipe even if the recipe's author roasted lamb leg chops.

2. Introduction

Peer-to-peer systems for Topic Map exchange envisaged in the introducing example already exist as well as approaches and protocols to Topic Map exchange. But all of them base on the agreement about shared vocabularies within the exchange network.

Our premise is that in practice the centralised enforcement of shared vocabularies has strong limitations. Only the semantic web search engine "swoogle" lists 763 different class definitions of 'person' found in divers ontologies[2]. Because all of the existing Topic Map exchange approaches completely fail if the peers use proprietary vocabularies, solutions for these environments have to be developed.

This paper makes the following contributions:

- Systematisation of the 'absence of shared vocabularies' in the context of Topic Maps Technologies in section 3.
- Description of existing approaches to Topic Map exchange and discussion of their limitations in the absence of shared vocabularies in section 4.
- Discussion of alternative Subject Equality decision approaches besides the Topic Maps Data Model (TMDM, [34]) in section 5.
- Introduction and Assessment of the SIM, a structuralist approach to Subject Equality decisions, which allows the exchange of Topic Maps in the (particular) absence of shared vocabularies in section 6.

3. The Absence of Shared Vocabularies

As sketched in the motivating example, Topic Map exchange is faced with the problem of the 'absence of shared vocabularies'. From a lazy point of view the 'absence of shared vocabularies' is the non-existence of mutual agreements about syntax and semantics of means for assertions about Subjects. This section systematises the notion 'absence of shared vocabularies'.

In section 3.1 the *semanticness* of Topic Maps Technologies is discussed. The semantic kernel of Topic Maps Technologies is examined in respect to the semantics of the vocabulary which will be shared. This supports the discussion about the nature of the necessary mutual semantic agreements. In section 3.2 the nature of Subject Equality decisions is further investigated. In section 3.3 the previous sections are

[2] http://swoogle.umbc.edu [requested: 15th April 2005]

summarised by systemizing the notion ‘absence of shared vocabularies’ in the context of Topic Map exchange.

3.1 Semantics in Topic Maps Technologies

Topic Maps are the international industry standard for *semantic* information integration. In a first step the *semanticness* of this technology will be depicted.

From an information science point of view *semantics* means that information systems are aware of the functionality which has to be applied to given data. There have to exist a well defined mapping from the syntax[3] to the semantic domain [18]. The difference between a *semantic technology* and a *non-semantic technology* is that in contrast to the latter one the semantic technology *reveals* the functionality which should be applied to data. In fact, the mapping from syntax to the semantic domain really exists to a sufficient extent. For example, an information system governed by a non-semantic technology applies to the string "<name>Leipzig</name>" an application specific functionality arbitrarily. A semantic technology, however, *reveals* the functionality which has to be applied to such a string.

The *semanticness* of Topic Maps Technologies is defined by the Topic Map Reference Model (TMRM, [9]). Generally, a Subject Map Disclosure[4] discloses (the examples for the TMDM, the common SMD, are given in parenthesises):

1. *SMD Ontology* (defines that Topics have Base Names, Occurrences)
   - *Subject Indication Approach* (defines that Topics indicate the Subjects they represent by Subject Locators and Identifier)
2. *Subject Equality Decision Approach* (defines that Topics having identical Subject Locators or Identifiers indicate identical Subjects)
3. *Subject Viewing Approach* (defines, in example, that the set of Topic Names of a merged Topic is the union of the Topic Name sets of the original Topics).

The only generic semantic functionality of Topic Maps is the following objective: Subject Proxies indicating identical Subjects have to be viewed as merged ones. *Only* this functionality constitutes the *semanticness* of Topic Maps Technologies.

Additionally to this *generic functionality*, a Topic Maps Processing Application (TMPA) performs *application specific functionality*: for example showing a Base Name as a string in the left corner of the screen. The semantics of all those application specific functionality is not revealed by the SMD itself.

This implies that Topic Maps Technologies do not define the semantics of the represented facts (the assertions belonging to Subject Proxies).[5] The definition of

---

[3] In our cases a specific syntax implies a specific kind of instances of the data model. Therefore the existence of a mapping between these instances and the semantic domain is necessary.

[4] The latest proposal of the TMRM [9] replaces the term “Topic Maps Application”.

[5] One might argue, that the creator of a Subject Map Disclosure have to describe the semantics of the Property Classes of the Subject Proxies, i.e. the meaning of the concept ‘Occurrence’. But there is no structured way for this semantic modelling and its non-existence does not

these semantics is left to the ontology engineers, which are appropriate for that task. But the ontology engineers should heavily exploit the fact that in Topic Maps all relationships between proxies and their subjects have well defined semantics. That's the uniqueness of Topic Maps which makes them to a real semantic technology.

As depicted in the listing above the generic semantic functionality of Topic Maps is split into two parts: *Subject Equality Decision* (deciding that Subject Proxies indicate identical Subjects) and *Subject Proxy Viewing* (viewing Subject Proxies indicating identical Subjects as merged ones).

Why this has to be discussed in the context of Topic Map exchange? Section 4 shows that this exchange bases on the request of Subjects. A remote peer requests information by indicating the Subject of interest. The requested peer has to decide whether it can provide a Subject Proxy indicating the identical Subject. This request scenario is the context of this paper. Therefore the Subject Equality decisions will be discussed in further detail.

3.2 The Subject Equality Decision in the Absence of shared vocabularies

A Topic Maps Processing Application, an application which processes Subject Maps according to given disclosures, has to do the Subject Equality Decisions as follows[6]:

$$\textbf{Subject Equality Decision}_{SMD_i}\big(\text{Subject Indication}_{SMD1}(\text{Subject Identity}_{\text{Subject Stage }1}),\ \text{Subject Indication}_{SMD2}(\text{Subject Identity}_{\text{Subject Stage }2})\big) \Leftrightarrow$$
$$\textbf{Subject Identity}_{\text{integration perspective}}(\text{Subject Stage}_1, \text{Subject Stage}_2)$$

The formalisation asserts, that a TMPA should decide that two Subject Proxies indicate identical Subjects (Subject Equality holds) iff from the current integration perspective the Subject Stages represented by these Subject Proxies belong to the same Subject. Thereby, each Subject Proxy documents the decision about its own identity with the means of the governing Subject Indication approach at documentation time.

As discussed in more detail in [6] section 2.1, Subject Identity is not an absolute "quality" due to the vague nature of Subjects. Rather it is the result of a perspective dependent decision process under uncertainty whether Subject Stages caught at different occasions and from different perspectives [5] belong to the same Subject. (These thoughts are strongly affected by Quine [28], [29]).

The TMPA is governed by a SMD_i which defines the Subject Equality Decision Approach that as to be applied. (The index i does indicate the integration perspective.) This decision has two parameters: the documentation of the Subject Identity of the first Subject Stage ($\text{Subject Indication}_{SMD1}$) and the documentation of the Subject Identity of the second Subject Stage ($\text{Subject Indication}_{SMD2}$). It is important to

---

influence the independent behaviour of a TMPA. Obviously, the definition of the semantics of an Occurrence item (in TMDM) does not influence the behaviour of a TMPA.

[6] For simplification, in the following the Subject Equality Decision concerning only *two* Subject Proxies is discussed.

outline, that the used Subject Indication Approach for the documentation of the decisions about Subject Identity at documentation time can be governed by a different SMD than the Subject Equality decisions at consumption time. A SMD based on the SIM introduced by this paper might imply such a situation.

Furthermore it is important to outline, that the perspective of the decisions about Subject Identity $_{\text{Subject Stage 1}}$ (at the time of creating the Subject Proxy belonging to Subject Stage 1), Subject Identity $_{\text{Subject Stage 2}}$ (at the time of creating the Subject Proxy belonging to Subject Stage 2) and Subject Identity $_{\text{integration}}$ (at the time of the decision about Subject Equality) might differ fundamentally. In [6] section 4, the evolution from a more technical perspective at documentation time to a special integration perspective at consumption time is discussed in detail.

The applied approach to Subject Equality decisions defines the semantics of the vocabulary (used to create the Subject Proxies) in respect to the only generic *semantic* functionality of Topic Maps Technologies: viewing Subject Proxies indicating identical Subjects as merged ones.

To understand the semantic implied by the approaches to Subject Equality decisions a side glance to linguistics is useful. Linguists distinguish between the referential and the structuralist paradigm. (Their differences are roughly reflected by the shifting from Wittgenstein's early thoughts to its late ones.) In referential semantics the meaning of a word (as a symbol) is defined by a referent (mostly outside the language) it refers to. According to the structuralist paradigm the meaning of words is only defined by their usage within the language.

Adopting this spadework we will differ between *Referential Subject Equality Decisions* and *Structuralist Subject Equality Decisions.*

The TMDM is a popular SMD adopting an approach to referential Subject Equality decision. If Subject Proxies' sets of Subject Identifiers/Locators comprise identical URLs, they have to be viewed as merged ones. Referring to a discrete 'thing' is the only mean for indicating the intended Subject. This approach enforces a Proxy to make explicit the Subject it intends to represent.

The premise of structuralist Subject Equality decision approaches is that the Subject depends on other Subject Proxies in the Subject Map. For example, the SIM introduced by this paper assumes, that whenever two Subject Proxies are used similarly, the probability that both indicate identical Subjects increases. The Subject is non tangible by any means, because it is emergently defined by relationships between Subject Proxies.

Summarised, the Subject Equality decision has the following structure:

Subject Equality Decision $_{\text{SMDi}}$ **(**
Subject Indication $_{\text{SMD1}}$, Subject Indication $_{\text{SMD2}}$,
Subject Map $_{\text{Subject Proxy1}}$, Subject Map $_{\text{Subject Proxy2}}$**)** ➔ **true | false**

The differences between the formalism introduced above have the following rationale. At the point of time the decision about Subject Equality is made, none information about Subject Identity is available. Only the documentation of the result of these decisions can be used. Additionally, the Subject Maps which are the origin of the according Subject Proxies are introduced as parameters. The rational is that at least structuralist Subject Equality approaches might rely on all Subject Proxies from these Subject Maps. At the moment, the decision about Subject Equality is a binary

one, whether equality holds or not. In future probabilistic or fuzzy approaches should be investigated.

3.3 Topic Map Exchange and the absence of shared vocabularies

In the following the previous insights are summarised to sketch the possibilities of an absence of shared vocabularies in the context of Topic Map exchange.

As shown in Figure 1, the chosen Subject Equality decision approach defines at consumption time the *semantics* of the vocabulary used by the Subject Proxies.

The competition of SMDs between the time of the Subject Equality decision (SMD_i) and the time the according Subject Proxies were created ($SMD_{1,2}$) implies different SMD ontologies which have to be handled. The 'absence of shared vocabularies' can be interpreted as the absence of a shared SMD ontology.

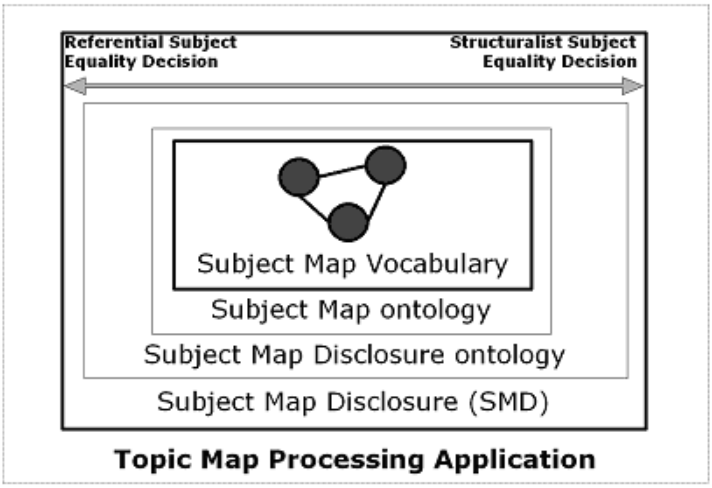

Fig. 1 Vocabularies and the Subject Equality Decision

Furthermore each Subject Map (governed by a SMD and its ontology) is restricted by an application specific ontology. For example, the type 'person' can be defined including further constrains for its instances (i.e. by a schema language). This specific ontology is called the SM ontology. The 'absence of shared vocabulary' might include the absence of a shared SM ontology, too.

Finally, inside a Subject Map the vocabulary at the instance level can be constrained, too. The concept of PSIs (Published Subject Identifiers, [33]) enforces, that if two Topic Map authors intend to refer to exactly the same Subject (i.e. a specific book is referred by using the according ISBN), they have to share these published vocabularies. The 'absence of shared vocabulary' might even include the absence of a shared SM vocabulary. The absence of a shared SM vocabulary might be

more important in the case a Referential Subject Equality Decision Approach is applied.

The nested relationships between all different kinds of vocabularies imply that the semantic (in the context of Topic Maps Technologies) of a specific vocabulary depends always on all higher layers.

4. Topic Map Exchange – The state of the art

Topic Map exchange is governed by a one-to-many-to-one problem (1:N:1) [19]. One master requests from N remote peers information about a Subject in interest (1:N). These remote peers extract their answer set, usually a Subject Map Fragment, from their local Subject Map. After receiving, the master has to integrate these different results into its local repository (N:1). *Request* and *Integration* are the tasks of Topic Map exchange to be solved in the 'absence of shared vocabularies'. Requesting is a retrieval task: retrieve the most appropriate Subject Proxy from a repository.

Request means, that the remote peers might receive Subject Map Fragments with unfamiliar SMD ontology, SM ontology or SM vocabulary. Under this uncertainty they have to decide about Subject Equality. The second part of the request is the specification of the Subject Map Fragment which has to send to the requesting peer.

Integration means, that the master has to decide about Subject Equality in respect to the received Subject Proxies in uncertainty about the used SMD ontology, SM ontology and SM vocabulary. This paper does only focus on the Subject Equality decisions. It leaves out the functionality of Subject Viewing.

In the following existing approaches to Topic Map exchange are introduced, whereby the arising problems in the absence of shared vocabularies are emphasised.

4.1 The Topic Map Remote Access Protocol (TMRAP)

The Topic Map Remote Access Protocol (TMRAP) [10], [14], [25], [27] is proposed by Ontopia[7]. It addresses requirements from distributed Topic Maps Portal integration. If a Topic Map Portal knows other TMRAP supporting Topic Map Portals it is enabled to request all information concerning a given Subject from these applications. The TMRAP bases on the TMDM (as common SMD and SMD ontology). In [16] TMViews as "mechanism for describing what to include when extracting a fragment from a topic map" is introduced. Besides being bounded on the TMDM, TMViews bases on the knowledge about the used SM ontology.

How does TMRAP address the Subject in interest? TMRAP enforces the usage of a shared SM vocabulary. If a Topic Map Portal requests information about a given Subject, it has to declare it by a (set of) Subject Indicators or one Subject Locator.

[7] http://www.ontopia.net

Furthermore, it has the opportunity to request information from a Topic with a specific Source Locator.[8]

Problems arising in the absence of shared vocabularies. That implies that all communicating Portals have to share a SM vocabulary.

4.2 TMShare

TMShare [1] is a P2P information sharing application based on Topic Maps Technology using the JXTA framework[9]. The aim of TMShare is to allow the exchange of Topic Map Fragments in a group of interacting peers. Each peer hosts a set of 'private' Topic Maps in designated back ends. Additionally, it hosts cached Topic Maps which were received from remote peers. TMShare bases on the TMDM.

How to address the Subject of interest? From our perspective, requesting a remote peer is quite similar to the TMRAP. Furthermore it may request for all Topics which satisfy a tolog query [15]. (The latest version of the TMRAP [14] does define this opportunity, too.)

Problems arising in the absence of shared vocabularies. As already discussed concerning the TMRAP all peers have to share the SM vocabulary. Using tolog queries is useful for customising requests. But at least in all cases where dynamic predicates or class definitions are used, the usage of tolog implies that the requesting peer is familiar with the remote peers' SM ontologies. A peer is only able to request the statement

```
performed-by($A : performer, aisha : song)
```

if it is familiar with the Association Type 'performed-by' and the Role Types 'performer' and 'song'.

4.3 The Knowledge Port Approach

Inspired by Bonifacio et al. [7], [8] Schwotzer proposed the Knowledge Port Approach (described in more detail in [31], [21]). Through the Knowledge Port Approach the Topic Map exchange is contextualized. Simplified, Knowledge Ports (KP) are end points of Topic Map exchange channels with the function of input/output filters. The peers store all information as Topic Maps.

How to address the Subject of interest? A peer stores three kinds of Topic Maps. The first reifies the known network structure. The second, called content map, is a Topic Map View about all local information. Additionally, information is useful in dedicated contexts, especially spatial coordinates. Therefore a Point of Interest (POI) map is introduced. Generally, each context should be modelled like the POI map.

---

[8] Requesting a Topic by its Source Locator is used to in the case the local ID is already known, e.g. from previous requests. For *semantic* integration the request of distributed Topic Fragments by their local IDs is out of interest.

[9] http://www.jxta.org

The Topic Map exchange takes place between the peer's Knowledge Ports. A requesting peer describes its demand with Topics from its local Topic Maps: its Subject in interest, its current POI, the allowed communication partners within the network. The Knowledge Ports of the requested peers match these demands with their offer. If all communication parameters fit, Topic Map exchange takes place. The Knowledge Port Approach bases on the TMDM.

Problems arising in the absence of shared vocabularies. All communication parameters (context, partners, Subjects in interest) are defined by PSIs within these ports. This is a shared SM vocabulary. Whereby for some parameters PSIs are inevitable (i.e. within the POI map), the definition of the Subject in interest with the help of PSIs delimit the power of the approach. Therefore, in [21] its liaison with the SIM approach is proposed.

4.4 From Federated Topic Maps to TMIP

Barta introduces an approach to federate distributed materialised and non-materialised Topic Maps [3]. This approach was further developed to TMIP, a RESTful Topic Maps Interaction Protocol [4]. TMIP bases on the TMDM.

How to address the Subject in Interest? While introducing Map Spheres TMIP always addresses the Subject in interest by using path expressions of the (future) Topic Maps Query Language (TMQL).

Problems arising in the absence of shared vocabularies. Similar to the tolog requests in TMShare and TMRAP, the path expressions of TMIP are bound to an overall knowledge of the SM ontologies and SM vocabularies of the requested peers.

5. Subject Equality decision approaches besides the TMDM

As shown in Figure 2 different approaches to Subject Equality decisions are imaginable. One has to outline, that each Subject Equality decision approach besides the TMDM implies a proper SMD.

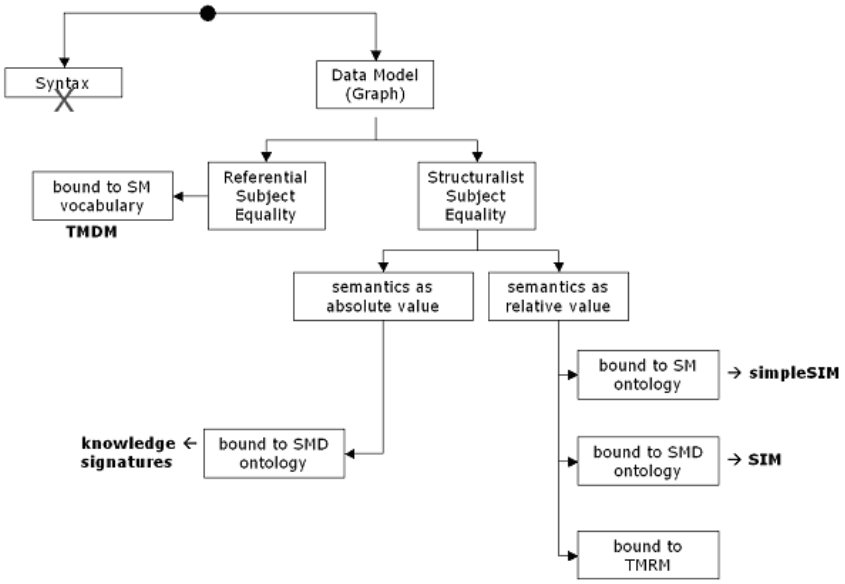

Fig. 2 Approaches to Subject Equality decisions besides the TMDM

Naturally, all approaches should operate on the data model level instead of the syntax level.

The first important decision is the differentiation between structuralist Subject Equality decision approaches and referential Subject Equality approaches. The latter is materialised by the TMDM. As discussed above, the TMDM enforces that all communication partners have to share the SM vocabulary.

In general, two kinds of structuralist approaches are imaginable. The first interprets a Proxy's Subject as a relative value. The SIM introduced by this paper materialises this approach. Being a relative value means that the Subject Equality between two Subject Proxies depends on the Subject Equality of all other Subject Proxies (which in turn depends on the Subject Equality which has to be decided) in the Subject Maps. Those algorithms do hardly scale.

For effective retrieval of conceptual graphs Sowa and Majumdar proposed the calculation of feature vectors representing the Subject of a conceptual graph as a concrete value [32]. These vectors are called knowledge signatures of a conceptual graph. The spatial distance between two knowledge signatures define the semantic closeness of the according Subjects. Retrieving of Subjects becomes very efficient. Those knowledge signatures interpret a Proxy’s Subject as absolute value.

6. The SIM Approach

We have shown that all existing approaches to Topic Map exchange are bound to the TMDM. To gain more flexibility, we propose the SIM Approach. This is a

structuralist Subject Equality decision approach ($SMD_i=SMD_{SIM}$). The SIM approach is independent of a shared SM ontology and SM vocabulary. But all Subject Proxies which are the input of the SIM have to be governed by the TMDM ($SMD_1=SMD_2=TMDM$).

Subject similarity is a weak kind of Subject Equality. The SIM Approach bases on the assumption that if two Subject Proxies interact with similar Subject Proxies in similar ways, the probability of their Subject similarity in the current context increases, too. And if the Subject similarity exceeds a specific threshold, Subject Equality holds.

The SIM Approach has strong relationships to Gentner's Structure-mapping theory. "This structural view of analogy is based on the intuition that analogies are about relation, rather than simple features. No matter what kind of knowledge (causal models, plans, stories, ect.), it is the structural properties (i.e., the interrelationships between the facts) that determine the content of an analogy" [10]

Furthermore, the SIM approach uses insights from schema matching gained by Melnik's et al. [23].

For brevity, the SIM Approach will be introduced in limited detail. The requesting Topic will be called *T*. The fragment of the requesting Topic Map around *T* will be called *F*. The fragment *F* consists of all Topics and Associations which are influenced by *T*. Our premise is that the fragment *F* indicates the Subject which is represented by *T*. (In future, TMView introduced by [16] might be used to define the fragment properly.)

After the reception of *F*, the remote peer compares each Topic from *F* with each Topic in the requested Topic Map and calculates a similarity measure (simDNA') for each pair of Topics.

The calculation is done in two iteration steps. In the first step only the similarity of the topology is exploited. In the second step additionally the similarity of Topics calculated in the first step is used. After the second step, Subject Equality holds for *T*'s most similar Topic from the requested Topic Map, if simDNA' exceeds a specific threshold.

The similarity of two Topics is calculated 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. The 'x' in the simDNAtype indicates that this Topic 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*. It 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 iteration 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 iteration step 1.

For each digit of the simDNAtype similar rules are defined. The complexity of these rules would go beyond the scope of this paper. The simDNA' is the sum of the digits of the simDNA. Basically, the higher the simDNA', the higher is the similarity of two Topics. Subject Equality holds for a pair of requesting and requested Topics if they gain the highest simDNA’ and this simDNA’ exceeds a specific threshold.

6.1 Assessment 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 environment are pruned randomly. What happens if randomly 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? The higher the recall, the better the SIM Approach allows to retrieve Topics in environments with unfamiliar vocabularies.

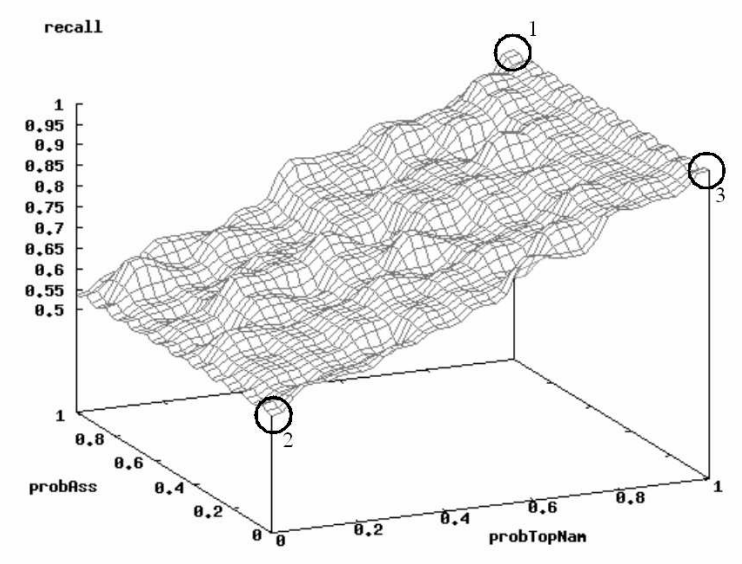

Fig. 3 Iteration: probTopName [0,1], probAss [0,1]

Fig. 3 shows the result of an experiment with a small Topic Map of 20 Topics. The probability of non-pruning Topic Names (probTopNam) and non-pruning 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 (see circle number 1). But, if both probabilities are 0, the recall is still 0.53 (see circle number 2). This implies, even in the case of a massive loss of information, when *all* Topic Names and *all* Associations are pruned, the typing information (typing of Topics, typing of Occurrences etc, whereby the typing Topics are pruned, too) and the information inside the Occurrences is sufficient to get the half of all Topics correctly.

Furthermore, the naming information has more influence on a high recall than the Topics' participation in Associations (see circle 3).

One has to bear in mind that the algorithm neither has knowledge about the used SM ontology and SM vocabulary 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 a shared SM ontology. This result implies that a combination of TMDM and SIM might be useful. In a first step, the Subject Equality decisions according the TMDM will be applied. In a second step, this information will be used, to decide for all Subject Proxies were the first decisions failed, whether Subject Equality governed by the SIM might hold.

This experiment sketches the abilities of the SIM Approach for Topic Map exchange in the absence of a shared SM ontology and SM vocabulary.

Problems of the SIM Approach. The SIM Approach has a number of limitations which should be introduced in short detail.

- If *F* and the requested Topic Map grow, complexity increases significantly.
- Applying the SIM Approach to non-materialised Topic Maps [3] is not possible.
- The SIM Approach does only yield good results, if the assertions of requested and requesting Topics are similar (i.e. the requested Topic Map provides small new information). If the requested Topic Map provides only new information, the SIM Approach fails.
- In cases a requested Topic Map can objectively not provide a Topic similar to a requesting Topic, the SIM Approach tends to post a false Topic. While the recall tends to be high, the precision tends to be low.
- In contrast to the TMDM, the decisions about Subject Equality are not deterministic. The result always depends on the whole requested Topic Map which might change randomly.

7. Related Work

"For computing the similarities, we rely on the intuition that elements of two distinct models are similar when their adjacent elements are similar." [23] Melnik et al. introduce and apply a graph matching approach for schema matching based on flooding of similarity through the graph [23]. Further they give a broad overview about existing matching techniques (which we do not want to rehash), mainly restricted to schema matching [30]. In contrast to the most efforts in information integration working at the schema level, solutions for Topic Maps Technologies should explicitly target to the integration at the instance level.

Our experiments with Melnik's et al. similarity flooding approach in conjunction with Topic Maps revealed substantial problems to provide a structuralist Subject Equality approach which is independent of SMD ontology, SM ontology and SM vocabulary. If Topic Maps are translated into the required directed labelled graphs (i.e. using Garshol's Foundational Model for Topic Maps [26] which is today superseded by the Q model [13]) the number of nodes increases enormously, conjunct with complexity problems. Additionally, nodes which represent the TMDM ontology (i.e. "SOURCE_LOCATOR") exhaust the similarity from the nodes which represent the SM ontology and SM vocabulary. These results showed that Melnik's et al. approach might be very interesting for SMD ontology and SM ontology matching. For the more general case of providing a common structuralist Subject Equality decision method, we had to decide to modify the approach significantly and to bind the SIM to the TMDM ontology.

Falkenhainer et al. report the implementation of Gentner's Structure-mapping theory through the so called Structure-Mapping Engine. The implemented algorithm has the poor complexity of $O(N^2)$, too.

Newcomb introduces the Versavant Project[10] (in early versions at the moment of writing) which provides a Topic Maps Application bus acting as "Subject addressing engine". This bus allows aligning between different SMD ontologies (and probably

[10] http://www.versavant.org

shifting between referential and structuralist Subject Indication paradigms). Versavant is further described in [24].

Additionally, Vatant introduces the concept of 'Hubjects' [35]. A Hubject is "a hub connecting different representations of a subject inside the same or across different contexts. [..] Hubjects provide neither semantic interpretation of the representations they connect nor absolute indication of the subject." [35]. As far as the sparse literature about Hubjects allows a Subject Equality decision method can be interpreted as a Hubject *class*.

Guo and Yu proposes the idea "that schema mapping and data mapping might be carried out simultaneously in a mutually way." [17]. Encouraged by the positive assessment, the SIM does mutually enhance the matching quality of schema entities and their data instances, too.

Basically, the issues discussed in this paper are strong related to the idea of emergent semantics [1].

In [20] and [22] we introduced a more lightweight version of the SIM Approach (see simpleSIM in Figure 2). This version yielded very good results, but was bounded to a common SM ontology. The new version of SIM is more generic.

8. Conclusions and Further Research

We outlined, that Topic Map exchange heavily depends on the Subject Equality decisions. We discussed this decision in detail, differentiating between a referential and structuralist approaches to Subject Equality decisions. We depicted, that the 'absence of shared vocabularies' might include the absence of shared SMD ontology, SM ontology and SM vocabulary. We introduced the SIM as a structuralist Subject Equality decision approach which is only bound to a shared SMD ontology (the TMDM). In future, the SIM should be disclosed as a SMD_{SIM} on top of the TMDM.

The main challenge of the current SIM Approach is the unbounded complexity. Today, the SIM Approach resembles a broadcast search within the requested Topic Map. The requesting Topic Map Fragments will be compared with each Topic from the requested Topic Map. Inspired from [36], interpreting the request of an appropriate Topic as a retrieval task, it is imaginable that each Topic knows *k* 'similar' neighbours inside its Topic Map. A requesting Topic will be forwarded through this network until it reaches its merging partner. We assume that only a few hops are sufficient to find this Topic (in contrast to the broadcasting approach today).

Additionally, the idea of Knowledge Signatures introduced by Sowa might be interesting to reduce the complexity.

Furthermore, the usage of the SIM approach might be appropriate to evolve a future TMQL towards a probabilistic query language, like probabilistic Datalog [11]. Such a probabilistic query language might allow requesting remote Topic Maps like:

> $topic_{TM1}(\$TYPE, 0.5\ person_{TM2})?$ – bind all Topics in Topic Map 1 (TM1) to the variable $TYPE which are with a probability of at least 50% similar to the Topic with id 'person' in Topic Map 2 (TM2).

instance-of$_{TM1}$($TYPE, $TOPIC), topic$_{TM1}$($TYPE, 0.5 person$_{TM2}$)? – bind all Topics in TM1 to the variable $TOPIC which are of the types specified by $TYPE.

Besides these ideas of future research, the SIM applied in the introducing example's cooking network would bring the chef interested in roasting lamb loins to the traditional French recipe for roasted lamb leg chops. Both handles with rosemary, green beans, lamb …

References

[1] Aberer, K. et al.: Emergent Semantics Systems. In: Proceedings of ICSNW 2004; LNCS 3226, Springer, (2004).

[2] Ahmed, K.: TMShare – Topic Map Fragment Exchange in a Peer-to-Peer-Application. In: Proceedings of XML Europe 2003, London; (2003).

[3] Barta, R.: Virtual and Federated Topic Maps. In: Proceedings of XML Europe 2004, Amsterdam; (2004).

[4] Barta, R.: TMIP, A RESTful Topic Maps Interaction Protocol. In: Proceedings of Extreme Markup Languages 2005, Montreal; (2005).

[5] Biezunski, M.: A Matter of Perspectives: Talking About Talking About Topic Maps. In: Proceedings of Extreme Markup Languages 2005, Montreal; (2005).

[6] Böhm, K.; Maicher, L.: Real-time Generation of Topic Maps from Speech Streams. In: Proceedings of TMRA'05, Leipzig; LNCS 3873, Springer, (2006).

[7] 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; pp. 191-200, (2002).

[8] Cuel, R.: A New Methodology for Distributed Knowledge Management Analysis. In: Proceedings of I-KNOW '03, Graz; pp. 531-537, (2003).

[9] Durusau, P.; Newcomb, S. R.: Topic Maps - Reference Model, 13250-5 version 6.0. Available at: http://www.isotopicmaps.org/tmrm/TMRM_6.0.pdf

[10] Falkenhainer, B.; Forbus, K. D.; Gentner, D.: The Structure-Mapping Engine: Algorithm and Examples. In: Artificial Intelligence, 41, pp. 1-63, (1989).

[11] Fuhr, N.: Probabilistic Datalog - a Logic for Powerful Retrieval Methods. In: Proceedings of SIGIR-95; (1995).

[12] Garshol, L. M.: XTM Fragment Interchange 0.1; Ontopia Technical Report 2002-09-23. Available at: http://www.ontopia.net/topicmaps/materials/xtm-fragments.html

[13] Garshol, L. M.: Q: A model for topic maps: Unifying RDF and topic maps. In: Proceedings of Extreme Markup Languages 2005, Montreal; (2005).

[14] Garshol, L. M.: TMRAP – Topic Maps Remote Access Protocol. In: Proceedings of TMRA'05, Leipzig; Springer LNCS 3873, (2006).

[15] Garshol, L. M.: tolog - a topic maps query language. In: Proceedings of TMRA'05, Leipzig; Springer LNCS 3873, (2006).

[16] Garshol, L. M.; Bogachev, D.: TM/XML - Topic Maps fragments in XML. In: Proceedings of TMRA'05, Leipzig; Springer LNCS 3873, (2006).

[17] Guo, M.; Yu, Y.: Mutual Enhancement of Schema Mapping and Data Mapping. In: Proceedings of the 10th ACM Knowledge Discovery and Data Mining, Seattle; (2004).

[18] Harel, D.; Rumpe, B.: Meaningful Modeling: What's the Semantics of "Semantics"? In: Computer, 37 (10) 64–72, IEEE (2004).

[19] Korthaus, A.; Hildenbrand, T.: Creating a Java- and CORBA-Based Enterprise Knowledge Grid Using Topic Maps. In: Proceedings of the Workshop on Knowledge Grid and Grid Intelligence, Halifax; pp. 207-218, (2003).
[20] Maicher, L.: Subject Identification in Topic Maps in Theory and Practice. In: Proceedings of Berliner XML-Tage 2004, Berlin; (2004).
[21] Maicher, L.; Schwotzer, T.: Distributed Knowledge Management in the Absence of Shared Vocabularies. In: J.UCS - Journal of Universal Computer Science, Volume 11, Special Issue I-Know 2005, Springer, (2005).
[22] Maicher, L.; Witschel, H. F.: Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM). In: Proceedings of LIT'04, Leipzig; pp. 229-238, (2004).
[23] 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).
[24] Newcomb, S. R.; Durusau, P.: Multiple Subject Map Patterns for Relationships and TMDM Information Items. In: Proceedings of Extreme Markup Languages 2005, Montreal; (2005).
[25] ISO/IEC JTC 1/SC34: Topic Maps Remote Access Protocol 0.2. Available at: http://www.jtc1sc34.org/repository/0507.htm
[26] ISO/IEC JTC 1/SC34: A Proposed Foundational Model for Topic Maps. Available at: http://www.jtc1sc34.org/repository/0529.htm
[27] Pepper, S.; Garshol, L. M.: Seamless Knowledge – Spontaneous Knowledge Federation using TMRAP. Presentation at: Extreme Markup Languages 2004, Montreal; (2004).
[28] Quine, W. v. O.: Identity, Ostension, and Hypostasis. In: Journal of Philosophy, 47(22), pp.621-633, (1950).
[29] Quine, W. v. O.: Word and Object. MIT Press (1960).
[30] Rahm, E.; Bernstein, P. A.: On Matching Schemas Automatically. Microsoft Technical Report MSR-TR-2001-17. Available at: http://www.research.microsoft.com/pubs/
[31] Schwotzer, T.: Modelling Distributed Knowledge Management Systems with Topic Maps. In: J.UCS - Journal of Universal Computer Science, Volume 10, Special Issue I-Know 2004, pp. 53-60, Springer, (2004).
[32] Sowa, J. F.; Majumdar, A. K.: Analogical Reasoning. In: Aldo, A.; Lex. W.; Ganter, B. et al.: Conceptual Structures for Knowledge Creation and Communication, LNAI 2746, Sprinter, pp. 16-36, (2003).
[33] Pepper, S.; Schwab, S.: Curing the Web's Identity Crisis. Subject Indicators for RDF. Available at: http://www.ontopia.net/topicmaps/materials/identitycrisis.html
[34] ISO/IEC JTC 1/SC34: Information Technology – Topic Maps – Data Model. Final Committee Draft. Available at: http://www.jtc1sc34.org/repository/0588.htm
[35] Vatant, B.: Tools for semantic interoperability: hubjects. Available at: http://www.mondeca.com/lab/bernard/hubjects.pdf, (2005).
[36] Witschel, F.: Content-oriented Topology Restructuring for Search in P2P Networks. Technical report, University of Leipzig, (2005).