

Subject Identification in Topic Maps -Theory and Practice-

XDWS 2004, Berlin
October, 11th 2004

Dipl.-Wirtsch.-Inf. Lutz Maicher

University of Leipzig, Germany

Contributions of Subject Identification in Topic Maps -Theory and Practice-

- Introduction of Topic Maps as a mean of knowledge representation.
- “One Topic for one Subject” - discussion of the central design criterion.
- “How to compare Subjects in heterogeneous and distributed environments?” - Discussion of Problems which occur in exchange and integration scenarios.
- Introduction, Assessment and Discussion of the SIM approach which addresses these problems.
- Challenges of further research.

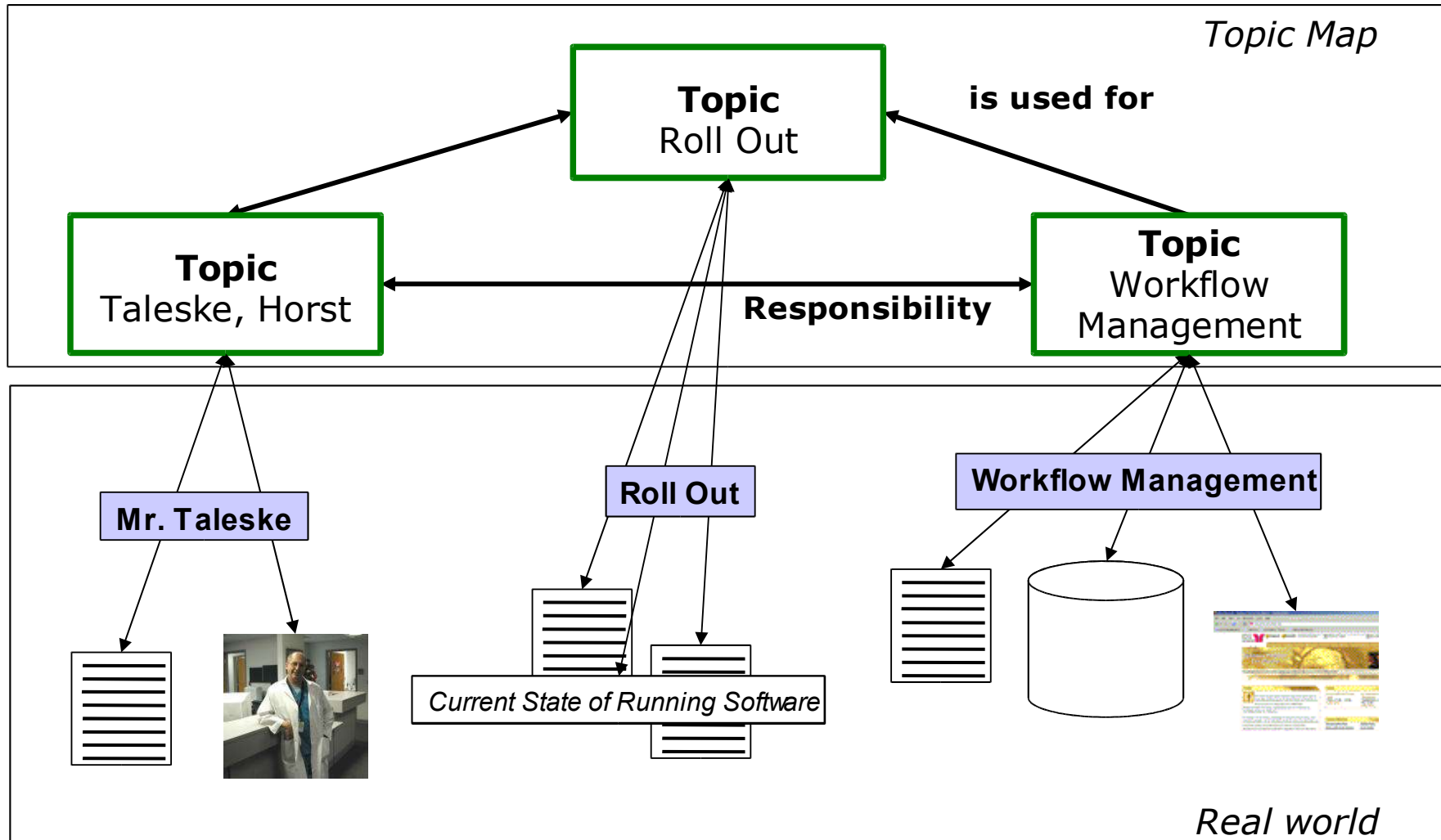
Agenda

- Contributions
- Introduction of Topic Maps
- Subjects and Topic Maps
- The Topic Map Standards
- One Topic for one Subject
- Subject Identity Measure (SIM) – Introduction, Calculation, Filtering
- Assessment of the SIM Approach – Testbed, Quality Measures, Results
- Further Research

Topic Maps – Topics and Associations (1/2)

- A **Topic Map** is a set of **topics** and **associations** (relationships between these topics)
- Topic Maps are means of (human centric) **knowledge representation**
- Topic Maps are **exchangeable**, powerful **indexes** for **heterogeneous, distributed** sets of information resources

Topic Maps – Topics and Associations (2/2)



Subjects and Topic Maps (1/2)

- **Topic Map** - a set of **topics** and **associations**.
- **Topic** - a symbol used within a **topic map** to represent some **subject**, about which the creator of the topic map wishes to make statements
- **Subject** - 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
- **Merging** - Every **topic** represents one, and only one, **subject**. Whenever two topics are known to represent the same **subject** they must be merged.

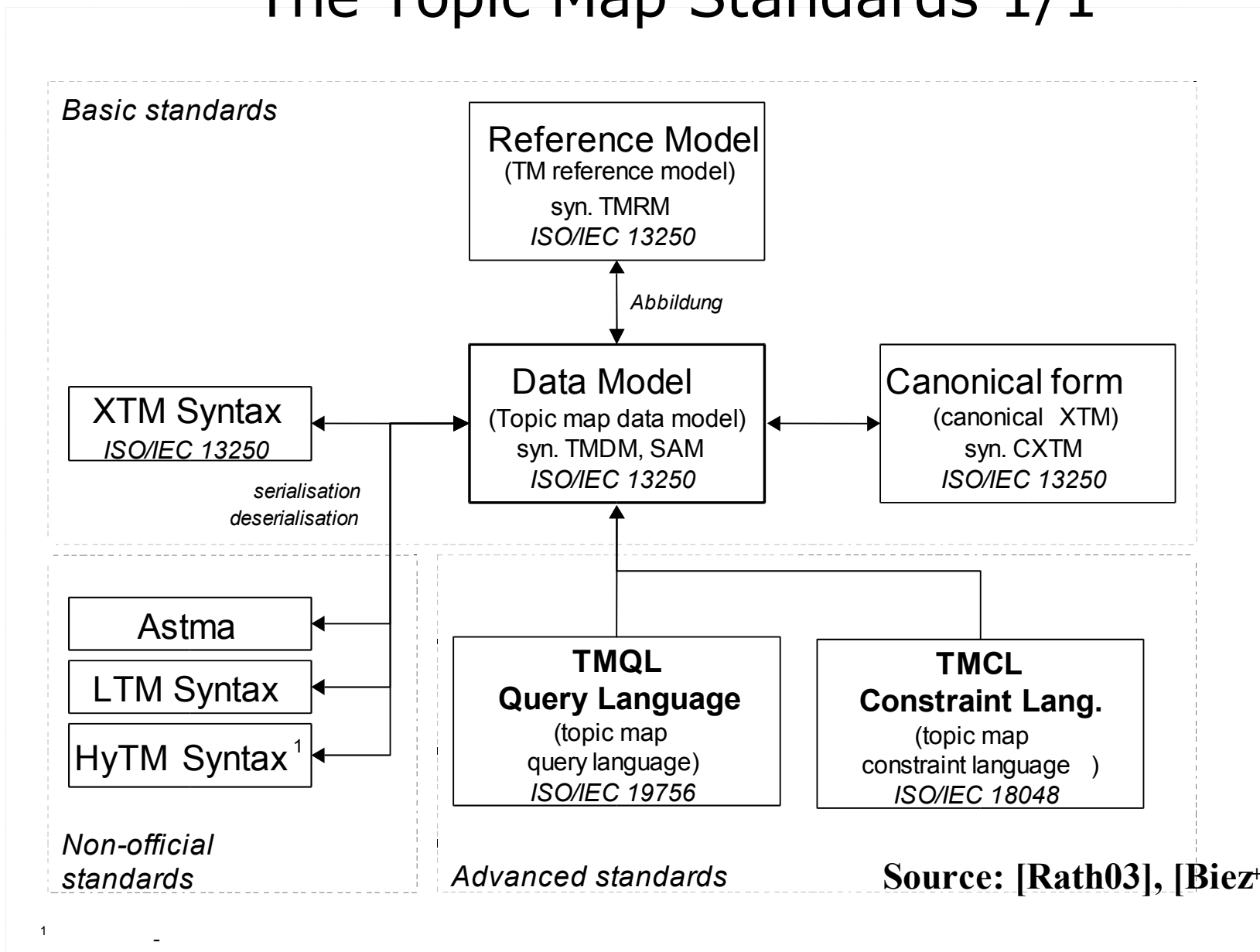
Source: ISO/IEC JTC 1/SC 34: "ISO/IEC 13250. Topic Maps – Part 2: Data Model."
Latest version available at: <http://www.isotopicmaps.org/sam/>

Subjects and Topic Maps (2/2)

- **Association** - a representation of a relationship between one or more **subjects**.
- **Topic Name** - a name for a **topic**, consisting of the base form, known as the base name, and variants of that base form, known as variant names
- **Occurrence** - a representation of relationship between a **subject** and an information resource
- **Topic Characteristic** - a **topic name**, **occurrence**, or **association** role belonging to some topic
- **Scope** - the context (a set of **topics**) within which a **topic characteristic** assignment is valid

Source: ISO/IEC JTC 1/SC 34: "ISO/IEC 13250. Topic Maps – Part 2: Data Model."
Latest version available at: <http://www.isotopicmaps.org/sam/>

The Topic Map Standards 1/1



„One Topic for one Subject“ 1/2

- A Topic declares its Subject via an URI.
- But Problems arise:
 - Subjects can't be discriminated sharply! What is a Subject?
 - Moreover: Does the URI „http://www.KMmTM.com/eMeyer“ declare the Subject „Ernst Meyer“ or the Subject „Homepage of Ernst Meyer“?
- **Solution:** Subject Locators vs. Subject Indicator
 - **Subject Locator:** A locator (URI) that refers to the information resource that *is* the subject of a topic.
 - **Subject Indicator:** An information resource that is referred to from a topic map in an attempt to unambiguously identify the subject of a topic to a **human being**.
- **Merging:** if two Topics have identical Subject Locators or a pair of identical URIs of Subject Indicators.

„One Topic for one Subject“ 2/2

- **Goal:** Exchange of Topic-Maps in distributed environments.
- **Goal:** Integration of heterogeneous data (texts, relational data, processes, system states) via distributed Topic Map views.
- How can two distributed Topic Map authors assure that Topics which represent the same Subject really merge?
 - **Published Subject Indicator (PSI):** A **subject indicator** that is published and maintained at an advertised location for the purposes of supporting topic map interchange and mergeability.
Example: <http://www.topicmaps.org/xtm/1.0/core.xtm#sort>
 - **Subject Identity Measure (SIM):** A measure (calculated by language independent algorithms based on NLP) of the similarity of the Subject of two distributed Topics. (*our work*)

The Subject Identity Measure – Introduction 1/1

- *The Subject Identity Measure (SIM) describes how closely related the Subjects of two Topics are.*
- **Goals of (simple) SIM**
 - rather Subject similarity (equality test) than Subject identity
 - recognition of Subject similarity with high precision
 - to be simple and efficient
 - to be (natural) language independent
 - baseline for (more expensive) approaches
- SIM makes us of
 - Topic Names (only strings; without types)
 - Occurrences (only Strings; without types)
 - only Topic Maps which bases on TMDM

The Subject Identity Measure - Calculation 1/3

- Starting with Calculation of the $SIM(T1,T2)$
- For two Strings $S1, S2$ a similarity measure $s(S1,S2)$ is calculated as follows:
 - Delete all special characters and numbers from $S1, S2$
 - Delete all words with less than k letters (i.e. $k=4$)
 - Let $|S1|=m, |S2|=n$ and $m < n$:
 - $c = 0$
 - For each t in $S1$
if t in $S2$ then $c = c + 1$
 - $s(S1,S2) = c/m$

The Subject Identity Measure - Calculation 2/3

- For each Pair of Topics (T1,T2) calculate:
 - **SIM.Names(T1,T2):**
 - Extract the sets of all Names $N1$ and $N2$ from T1 and T2
 - Let $m = |N1| < |N2|$. Calculate SIM.Names:

$$SIM.Names = \frac{1}{m} \sum_{n_1 \in N_1} \max_{n_2 \in N_2} s(n_1, n_2)$$

- **Sim.Occurrences(T1,T2):** same algorithm, but for sets of Occurrences $O1$ and $O2$

The Subject Identity Measure - Calculation 3/3

- Choose $0 \leq \lambda \leq 1$ and calculate $SIM(T1, T2)$:

$$SIM(T1, T2) = \begin{cases} 0 & \text{falls } (SIM.Names < t_{Name}) \vee \\ & (SIM.Occurences < t_{Occ}) \vee (SIM < t) \\ \lambda SIM.Names + (1 - \lambda) SIM.Occurences & \text{sonst} \end{cases}$$

- $SIM.Names$, $SIM.Occurences$ and SIM have to exceed specific threshold!

The Subject Identity Measure - Filtering 1/1

- ... and now find the merging pairs!
- If two Topic Map Fragments TM1 and TM2 join, merging pairs have to be found:

(conceptual constraint:

a Topic has only one or zero merging partner)

- Let TM1 be the smaller Topic Map (less Topics)
- For each $T1 \in TM1$ find

$$T2 = \arg \max_{T_i \in TM_2} SIM(T1, T_i)$$

- If $SIM(T1, T2) > 0$, $(T1, T2)$ is proposed as a merging pair; else T1 has no „merging partner“.

Assessment of SIM - Testbed 1/2

- **Testbed:** „artificial“ Topic Maps generated from online library catalogs (different descriptions of one Subject)
 - „Gemeinsamer Bibliotheksverbund“ (GBV)
 - „Deutschen Bibliothek“ (DDB)
- **Topics are:**
 - all publications of Springer published in 1997
 - Topic Names = title, subtitle
 - Occurrences = keywords, subject headings, publisher etc.
- no Types, Associations etc.
- **objective criterion for equality test: ISBN**

Assessment of SIM – Testbed 2/2

- choose randomly 300 Topics (books) from GBV and DDB
- obtain 90.000 possible pairs of Topics (T1,T2)
 - **Complexity**: 90.000 times $SIM(T1,T2)$!
- in our example: 25 pairs (out of 300 in maximum) are valid matches (have the identical ISBN)

Assessment of SIM – Quality Measures 1/1

- Let G the set of all pairs obtained by the SIM and
- Let I the set of (valid) identical pairs (i. e. $|I| = 25$):

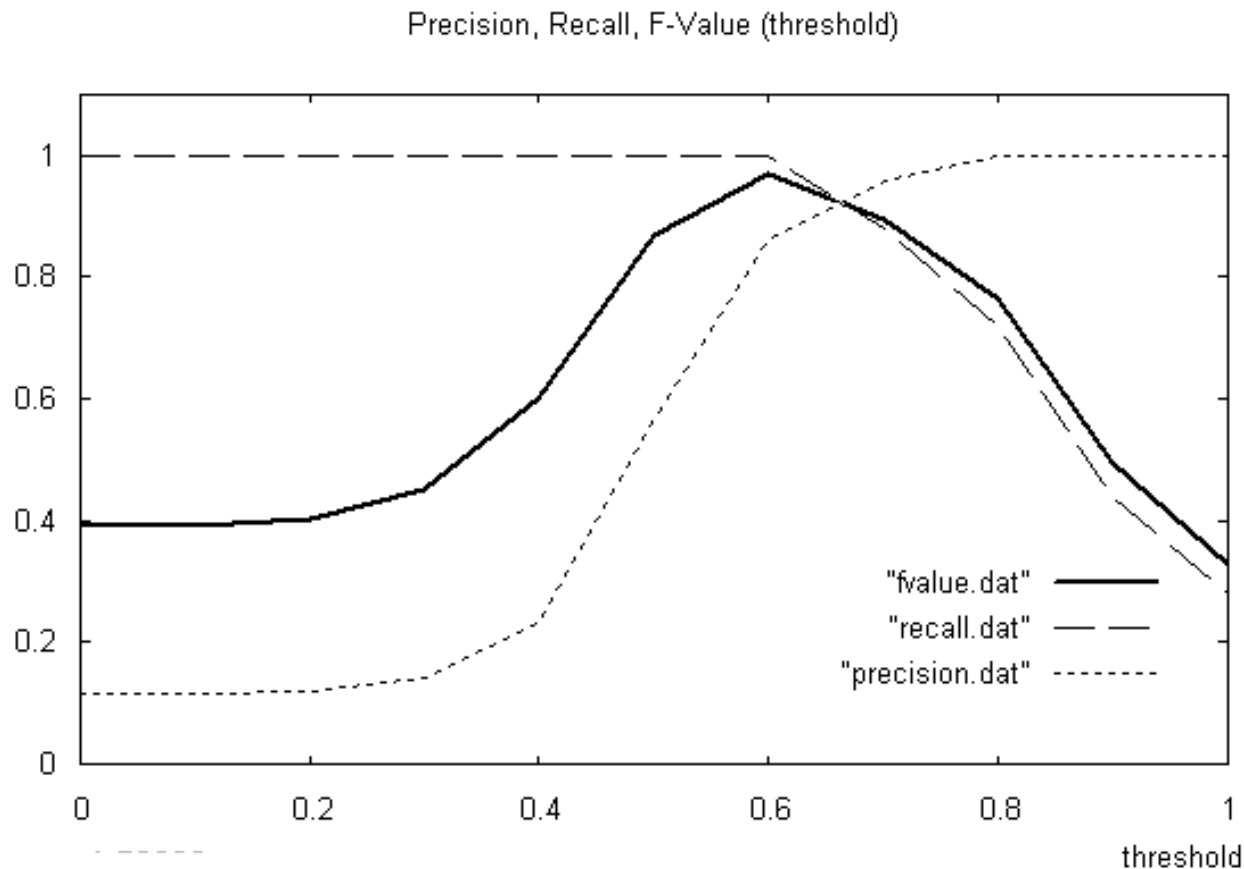
– Precision: $P = \frac{|G \cap I|}{|G|}$

– Recall: $R = \frac{|G \cap I|}{|I|}$

– F-Wert: $F(\beta) = \frac{(1 + \beta^2)PR}{\beta^2 P + R}$ (we use $F(2)$)

Assessment of SIM – Results 1/2

- Results obtained with $\lambda = 0.5$, $t_{Occ} = t_{Name} = 0$, t varies:



Assessment of SIM – Results 2/2

- (simple) SIM yields very good results (for our testbed)
- **Maximum:** $F(2) = 97\%$ ($R = 100\%$, $P = 86.2\%$)
with $t = 0.6$
- Further tests with a lot of different, randomly created subsets from the testbed showed:
 - always the maximum of $F(2)$ is similar and is reached with similar thresholds
 - usage of t_{Occ} und t_{Name} has a positive influence to quality (and t can be replaced)
 - the parameter λ can be canceled ($\lambda = 0.5$ is optimum)

SIM – Further research 1/1

- **SIM can be improved** (advanced SIM):
 - Types: Types of Topics and Topic Characteristics might be a good indicator of similarity
 - Associations: if two Topics have similar neighbourhoods they might be similar (attention: recursive)
 - References: URIs and the content of referenced information resources are out of focus
- **and other challenges**
 - Create Topic Maps (as testbeds) from texts, processes, relational data and system states (integration scenarios!)
 - Adoption of other promising approaches (i. e. from schema matching) to Topic Maps
 - Topic Map metrics for calculation of correlations between matching techniques and matching quality
 - ... and establish Subject-centric thinking

Contributions of Subject Identification in Topic Maps -Theory and Practice-

- Introduction of Topic Maps as mean of knowledge representation.
- “One Topic for one Subject” - discussion of the central design criterion.
- “How to compare Subjects in heterogeneous and distributed environments?” - Discussion of Problems which occur in exchange and integration scenarios.
- Introduction, Assessment and Discussion of the SIM approach which addresses these problems.
- Challenges of further research.

Discussion

Lutz Maicher

(maicher@informatik.uni-leipzig.de)

University of Leipzig