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

Real-time Generation of Topic Maps from Speech Streams

Karsten Böhm, Lutz Maicher

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

Abstract. Topic Maps are means for representing sophisticated, conceptual indexes of any information collection for the purpose of semantic information integration. To properly fulfil this purpose, the generation of Topic Maps has to base on a solid theory. This paper proposes the Observation Principle as the theoretical fundament of a future scientific discipline Topic Maps Engineering. SemanticTalk generates sophisticated, conceptual indexes of speech streams in real-time. Reflecting the Observation Principle, this paper describes how these indexes are created, how they are represented as Topic Maps and how they can be used for semantic information integration purposes.

1. Introduction

The challenges addressed by this paper are described by the following example. Interviews are a very popular technique in qualitative research in a variety of fields, from social sciences to market research. For example, an automotive company observes the interests of its customers with the help of guided interviews. These interviews are continuously repeated with the same customers within intervals of one year. The task of the interviewer from the market research department is the observation of the rise and fall of trends within the time line. Additionally, he has to provide background knowledge about these trends for further investigations.

From our perspective we understand a Topic Map[1] *firstly* as a representation of a sophisticated index of any information collection. *Secondly* a Topic Map has to be generated in order to merge with others to allow semantic information integration.

This paper describes how sophisticated, conceptual indexes can be generated from speech streams automatically in real-time by a software-system called SemanticTalk. Topic Maps Technologies allow it to integrate the concepts observed in the speech-streams with known background knowledge to support the interviewer from the market research department. The paper is based on a solid theory how indexes of any in-

[1] To avoid ambiguities all terminology concerning Topic Maps Technologies is capitalized.

formation collection created by any means could be represented as Topic Maps to fulfil their purpose as integration tool.

Besides intellectual approaches to Topic Maps Engineering, the automatic generation of Topic Maps from textual sources is a relatively new area that has been covered by some research approaches already. The application of such techniques to speech streams in order to provide a real-time generation of Topic Maps from a spoken conversation is a new approach that is presented in this paper.

There are a number of interesting applications imaginable. The most evident is the real-time generation of Topic Maps for the comprehension of the core concepts of conversations in meeting situations. These Topic Maps should be used to link background knowledge to the key concepts of the current conversation by merging in Topic Maps Views (TMV, a set of Subject Proxies [2]) of other relevant information collections. By implementing this automatic process directly onto the speech stream it would be possible to build a Push-System that provides the user with relevant or related information without requiring an explicit search.

The generation of Topic Maps should base on a solid theory. To our best knowledge this theory is not available in the current literature. Therefore we introduce the Observation Principle. This principle is proposed to be the basis of a future scientific discipline Topic Maps Engineering, regardless whether automatic or manual methods are addressed. The Observation Principle bases on a deeper investigation of the Subject Equality Decision Chain. This chain discloses the nature of the semantic core of Topic Maps Technologies: the decision whether two Subject Proxies indicate identical Subjects. This paper discusses the generation of Topic Maps with SemanticTalk in the context of the Observation principle.

This paper makes the following contributions:

- Introduction of the Observation Principle based on a deep investigation of the Subject Equality Decision Chain in section 2,
- Introduction of SemanticTalk as means for the real-time generation of sophisticated indexes of speech-streams in section 3, and
- Reflecting the Observation Principle while representing these indexes as Topic Maps for semantic information integration in section 4.

2. The Observation Principle

Topic Maps are means for representing sophisticated indexes of any information collection for the purpose of semantic information integration. From this perspective, the generation of Topic Maps, whether it is done manually or automatically, needs a solid theory. Our proposal is called Observation Principle. It is introduced in brief by this section. Previously, the Subject Equality Decision Chain is introduced.

2.1 The Subject Equality Decision Chain

As discussed more detailed in [10] the *semanticness* of Topic Maps Technologies is determined through the following objective: Subject Proxies indicating identical Subjects have to be viewed as merged ones. The insight gained from the semanticness discussion in [10] is that the enforcement of this objective is the only semantic feature of Topic Maps. Consequently, Topic Map authors have to be strictly aware to *correctly* define the Subject of their proxies according to their exact intents.

But when do Subject Proxies *indicate identical* Subjects? This decision about Subject Equality is determined by a foregoing process of abstractions, simplifications and decisions. The following thought chain might help to understand the impact of this process on the semanticness of Topic Maps. This Subject Equality Decision Chain reveals the real nature of the decision whether two Subject Proxies indicate identical Subjects.

Basically, within the Subject Equality Decision Chain the separation of *Subject Identity* and *Subject Equality* is introduced. Subject Identity is based on the decision whether Subject Stages caught at different occasions belong to the same Subject. Subject Equality, however, is based on the decision whether Subject Proxies indicate identical Subjects. Simplified, Subject Identity targets to the real world whereby Subject Equality targets to the modelled world each Topic Map represent.

In the following the Subject Equality Decision Chain is described in detail. The subsequent summary defines the six steps of this thought chain:

1. Assumption of a world without any sensory systems,
2. Sensory systems come to stage, catching Subject Stages,
3. Decision about Subject Identity from a given perspective,
4. Decision about Subjectness from the given perspective,
5. Documenting the impressions from the given perspective,
6. Documenting the decision about Subject Identity, and
7. Decision about Subject Equality according to the governing Subject Map Disclosure (SMD, see [5], [10]) at consumption time.

The first step of the Subject Equality Decision Chain is a universe of Subjects. It is assumed, that within this universe all possible Subjects purely *exist*, but they are not recognised and deranged by any sensory system. This absence of any sensory system implies slightness to the very nature of the universe of Subjects. Any suppositions about these Subjects are speculations, because any empirical evidence is impossible to gain. Obviously, discussions about the very nature of Subjects should be left to the philosophy. Caused by the absence of any sensory system, within in the first step the question of Subject Identity or Subject Equality is irrelevant: Subjects solely exist, no one is (or can be) interested in their nature.

As a second step, sensory systems come on stage, trying to investigate the nature of Subjects (even if the "investigated" Subjects do not exist in the physical reality, like a unicorn or the abstract concept of liberty, etc.). Immediately, as a third step the problem of Subject Identity arises. In accordance to Quine's "momentary stages" [13] of things, we assume that the owner of sensory systems observes Subject Stages. *A*

Subject Stage is a momentary stage of a Subject. (Simplified, it is an observation of a Subject). For example, a ranger might observe moose at different occasions. He catches different momentary stages. The question is, whether both moose stages belong to the same moose. From this perspective *Subject Identity means that Subjects Stages caught at different occasions belong to the same Subject.* It is important to stress the fact that this definition disburdens from the constraint to make explicit what the Subject is (only its sameness is used as an identifying criteria). For the ranger, the decision about Subject Identity is a judgement whether the caught moose stages belong to the same moose.

Another example is an automatic traffic control indexing the traffic stream by taking pictures from each car passing different points. The question of Subject Identity is, whether two car stages caught at different occasions are the same car [14]. The question of Subject Identity is not compelling whether two car stages caught at different occasions belong to the car with a concrete registration number.

Obviously some philosophical questions remain open. For example, all Subjects alter in time, at least imperceptibly. In this case, is it possible to catch exactly the same Subject twice? The decision about Subject Identity is thus always a decision under uncertainty. As the first step of the thought chain revealed, only the observations gained from the sensory systems are available to judge about Subject Identity. It is compelling that this information is never sufficient to judge about Subject Identity with certainty. Uncertainty is an important aspect of the decision about Subject Identity.

Even in the case that the identity seems to be sure, this certainty is derived from the chosen perspective. Biezunski notes: each "interpretation of any subject is made within certain perspective." [2]. The chosen perspective defines the assertions which should be made later about the Subject. From a child's perspective all car stages caught by the traffic surveillance system belong to the same Subject ‘car’. From a toll system's perspective each car stage caught at different occasions belongs to the Subject of a specific toll debtor. The chosen perspective therefore heavily determines the decision about Subject Identity.

What does the fact that the child learns to differentiate car brands and models imply for Subject Identity? It reveals that the decision about Subject Identity is a process. Subject Stages are considered to belong to the same Subject as long as convincing information is received from the sensory system. Even if contradicting information is observed, Subject Identity might be considered as long as this information is not sufficient to disprove the present conviction. Looking at it as a process is an important characteristic of the decision about Subject Identity.

We conclude, that the decisions about Subject Identity is *a perspective dependent discovery processes under uncertainty*.

The next part of the Subject Equality Decision Chain is dedicated to the documentation of the impressions from the sensory systems: changing from the real world to the modelled world, changing from the question of Subject Identity to the question of Subject Equality. It is important to stress the fact, that modelling always implies a loss of information. The fourth step of the thought chain represents the first decision concerning this loss of information: impressions about which Subjects (Stages) should be documented? This is a decision about the *Subjectness* of a given Subject in the current perspective. The child looking at the street catches a lot of Subjects besides cars (SOS-telephones, traffic signs). By heavily repeating the word "car", the child shows that from its current perspective only the caught car stages are noteworthy.

After the decision about Subjectness the impressions should be documented. (This documentation is equal to modelling the caught impressions from the current perspective with the available vocabulary.) For each Subject Stage with ‘proved Subjectness’ in the current perspective a Subject Proxy has to be created. *Observation Subject Proxies* might be an appropriate name for those proxies. Afterwards, all impressions which should be documented about the recognised Subject Stages have to be documented using the available vocabulary: SMD ontology, Subject Map (SM) ontology and SM vocabulary (see [10] for a further discussion of these terms).

It might be evident to document the decision about Subject Identity simultaneously in the recently created Observation Subject Proxies. The decision about Subject Identity is a process under uncertainty. The conviction about Subject Identity of Subject Stages might alter in time. Therefore, the creation of additional *Integration Subject Proxies* is proposed. These Subject Proxies only assert that two Observation Subject Proxies indicate identical Subjects. Generally, the decision about Subject Identity has to be documented by using the Subject Indication approach of the governing SMD.

For the moose example this implies that for all times the ranger observes moose stages it has to be created an own Observation Subject Proxy to document the impressions at these occasions. Additionally, an Integration Subject Proxy has to be created to assert that these Proxies indicate identical Subjects from the current perspective. Integration Subject Proxies can be compared to Hubjects as introduced by Vatant [17].

The last step of the Subject Equality Decision Chain is the decision of a Topic Map Processing Application (TMPA) whether two Subject Proxies indicate identical Subjects. As discussed in [10] in more detail, this decision about Subject Equality depends on the governing SMD at consumption time. If the Topic Map Data Model (TMDM, [15]) is the governing SMD, these decisions are simple string comparisons.

2.2 The Observation Principle

The Observation Principle might be a theoretical fundament for a future scientific discipline Topic Maps Engineering. We assume that Topic Maps created according to the Observation Principle are well suited for semantic information integration, the core functionality of Topic Maps Technologies. The Observation Principle postulates that each observation (caught by any provided sensory system) of a Subject Stage *in interest from the chosen perspective* has to be documented by a proper Observation Subject Proxy. The documentation of the decision about the Subject Identity by means of Integration Subject Proxies allows defining the perspective in a proper way. All documentations are constrained by the Subject Indication Approach and the additional vocabulary provided by the governing Subject Map Disclosure (SMD).

1. Observe the information collections in interest (texts, video streams, etc.) by any means and detect Subject Stages.
2. Decide about the Subject Identity of the observed Subject Stages from the chosen perspective.
3. Decide about Subjectness from the chosen perspective.
4. Create Observation Subject Proxies for each Subject Stage in interest.
5. Document all observed information about the Subject Stage with the given vocabulary: SMD ontology, SM ontology, and SM vocabulary.
6. Document the decisions about Subject Identity by the means of Integration Subject Proxies with the provided Subject Indication Approach.

The modus operandi of observations can be a set of NLP methods (like those in SemanticTalk), a human hand-crafting a Topic Map, a mapping approach integrating data from a legacy system like a relational data base or any other function creating any kind of information which can be interpreted as "observations of Subject Stages". It is important to outline, that Topic Maps do only *represent* indexes created for any information collection by any means.

In the following it is discussed, how the SemanticTalk System can be used to apply the Observation Principle. As a first step SemanticTalk is introduced.

3. The SemanticTalk System

SemanticTalk (see figure 1) observes speech streams in real-time with a number of different Natural Language Processing (NLP) methods. It generates sophisticated conceptual indexes of speech streams in real-time. According to the Observation Principle, this bouquet of NLP methods defines a proper SemanticTalk perspective.

This section introduces the technical background of SemanticTalk. The following section discusses how SemanticTalk can be used to represent the created indexes as Topic Maps with the objective of semantic information integration.

SemanticTalk can be used on a single computer. It is implemented as platform-independent client-server architecture to support a multi-user scenario. Several GUIs as well as independent speaker input units, consisting of a standard headset and a client notebook, can be locally distributed and connected via standard networks, for example in a wireless local area network (WLAN). It thus enables the use in dynamic meetings scenarios. A more detailed description of the system can be found in [1].

SemanticTalk provides various interfaces to both data sources and sinks, the most important of which is the input module for free speech, a text mining component called Graph Distiller to automatically assemble static indexes, and the XML-based interface for the bi-directional exchange of structured data.

For speech-to-text conversion, the commercially available dictation system Voice-Pro from Linguatec is used. Although speech recognition is not error free the performance is noteworthy and for our requirements the generated word stream is more than sufficient. Term extraction is performed using a large-scale reference terminology corpus (about 10 million phrases mainly from newspaper papers, automatically generated) as background information (comparable to the general linguistic competence of native speakers). This database can be extended by domain specific terminology. This can be derived by processing relevant sets of documents using text mining algorithms. In the case highly specialised domains are investigated this extension is quite useful. While the term ‘cancer’ might by a key concept in a conversation about everyday life, it is not sufficient selective in a medical conversation about leukaemia. Adding domain specific terminology alters the perspective SemanticTalk takes.

Filtering relevant concepts is done by extracting nouns that exceed an adjustable relevance threshold for the domain of discourse. In the next step, relationships between terms selected from the text stream are determined based on term relations in the underlying database. The term is discarded, if no additional terms can be associated with this term within a defined time frame. Otherwise it will be added to the growing conceptual structure (white nodes in figure 1). This index will be continuously updated, as long as the speech stream is analysed. Furthermore, the network is expanded by additional terms that are not present in the voice stream. They are integrated if the strength of the association is above a threshold (grey nodes in figure 1).

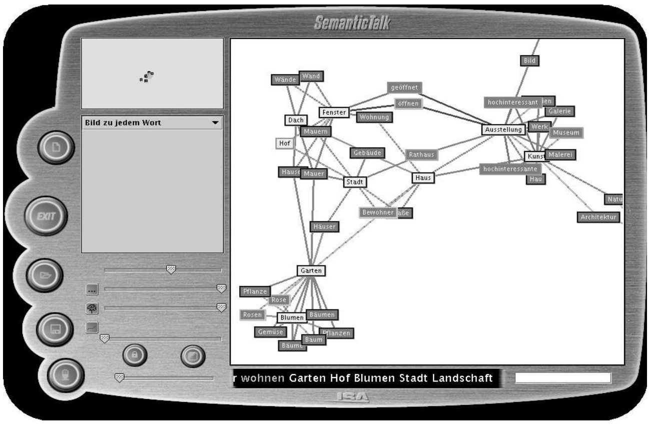

Fig. 1. The user interface of SemanticTalk

In order to generate a conceptual background structure, a module called Graph Distiller is used. It takes one or more text documents as input and generates a corresponding conceptual view on the data. The interconnection of concepts is based on the same principles as described above, but no additional concepts are associated. The output is only a network of domain specific concepts and their relationships. However, there is a significant difference to the process described above, as at any point during the processing only the current word and the history of words is available, no assumptions about the future of the ongoing discourse can be made. In contrast the Graph Distiller benefits from the availability of the complete text material to construct the graph and it is not required to work in real-time. Hence, a more sophisticated and thus more costly algorithm is applied to extract the relevant concepts. Obviously, using the Graph Distiller alters the SemanticTalk perspective slightly.

The extraction algorithm is based on *Difference Analysis* [6]. It applies pure language statistics: the distribution of terms in the given text documents is compared to the distribution of the same terms in the general use of a specified language, for example German. If a term occurs significantly more often in the examined input text documents than in the general language use, it is called a 'key concept of the domain'. The conceptual graph is constructed of only these key concepts. As based on pure statistics, Difference Analysis is language independent. However, it requires a large, representative collection of digital documents. Beyond Difference Analysis, the Graph Distiller optionally provides algorithms as described in [18]. For morphologically rich languages like German, complex noun decomposition and lemmatization particularly improve the output of the Graph Distiller.

SemanticTalk provides a bi-directional interface (import/export) for the Resource Description Framework (RDF), a language representing statements about resources and their relationships. Hence, virtually all kinds of graphs generated in different applications can be used as background information for SemanticTalk. The results of a SemanticTalk sessions can be exported to RDF and as shown in section 4 be further processed to refine and augment the generated structures.

4. SemanticTalk and the Observation Principle

SemanticTalk generates sophisticated conceptual indexes of speech streams. This section provides an illustrative example from the automotive industry. After a brief description of the example setting, section 4.1 depicts how SemanticTalk generates indexes by taking its own perspective. Thereafter, the representation of these indexes as Topic Maps for the purpose of information integration is described in 4.2 and 4.3.

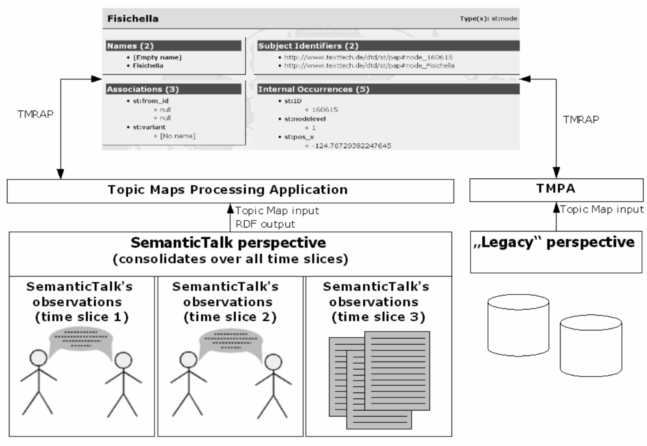

Fig. 2. SemanticTalk and the Observation Principle

The example setting described in figure 2 was used to index ten conversations about the automotive domain at different time epochs (time slices). The goal of the experiment was to show the changes of subjects during time although the domain remains fixed. Additionally to speech streams, other available documents (such as a year's publication of an automotive magazine) were used to extract similar indexes.

4.1 Getting the SemanticTalk Perspective

What does SemanticTalk actually observe? Does it, as the figure 2 might indicate, recognise *Fisichella* as a key concept of the conversation?

To answer to this question accurately, the application of the Observation Principle in SemanticTalk should be investigated in more detail: SemanticTalk observes speech streams and recognises words (or phrases). From the SemanticTalk perspective each word (or phrase) can be interpreted as a Subject Stage. With the help of some linguistic methods (stemming, etc.) a decision is made about the identity (simplified its base form) of each Subject Stage (the observed inflected word form). During the observation of the speech stream the frequency and the use within the context of the utterance is observed. Based on these observations, noticeable usage of a word implies its Subjectness from the current SemanticTalk perspective. As already described in section 3, this perspective depends on the methods and corpora which are applied in detail. If Subjectness is decided, a Subject Proxy is created for each Subject (a node) and all observed impressions, which are in interest from the current perspective, are documented. (In contrast to the proposal in section 2.2, our implementation does not differentiate between Observation and Integration Subject Proxies.)

Coming back to the question about what SemanticTalk really observes? It does not recognise the concept ‘Fisichella’ as a key concept of the conversation. This might be an interpretation from another perspective which is in interest for integration purposes as discussed in 4.3. SemanticTalk does only observe a noticeable usage of the term ‘Fisichella’ in the speech stream and this 'noticeable usage of the term Fisichella’ is the Subject of an according Subject Proxy.

The same procedure holds for the relationships between the terms, because relationships are represented as Subject Proxy, too ([5]). The Subject of those proxies, the Associations, is the observed noticeable relationship between two terms.

4.2 Creating the SemanticTalk Topic Maps View

As mentioned above, SemanticTalk generates RDF output for the further usage of the created index. As described in figure 2, the next step is the generation of a Topic Map from this output:[2] the SemanticTalk Topic Maps View (SemanticTalk TMV) of

[2] At the moment, SemanticTalk does only create the indexes in real-time. The generation of Topic Maps based on these indexes is done with the result of conversation observations afterwards. In future, it is desirable that a TMPA becomes the kernel of SemanticTalk.

the observed conversation. This TMV must be usable for further semantic integration tasks, like integrating background knowledge from marketing, customer and employee databases. The following example assumes the TMDM as the governing SMD.

As discussed in literature, the mapping of RDF and Topic Maps governed by the TMDM should be a semantic mapping [12]. Today, such a semantic mapping is provided by Ontopia's Omnigator[3]. In future a Q engine, as proposed by Garshol [7], might be able to process RDF and Topic Maps simultaneously.

SemanticTalk's RDF output consists of a set of nodes (Subject Proxies for the observations of noticeable usage of terms) and a set of edges (Subject Proxies for the observations of noticeable simultaneously usage of noticeable terms) connecting these nodes. The following listing shows (a part of) the description of a node in RDF:

```
<st:node rdf:ID="node_Fisichella">
  <st:ID>160615</st:ID>
  <st:label>Fisichella</st:label>
  <st:nodelevel>1</st:nodelevel>
  <st:ref_wort_nr rdf:resource="#node_160615"/>
  <st:variant st:index="3" st:type="4" st:weight="0.3176"/>
</st:node>
```

The node with the internal ID ‘160615’ has the label ‘Fisichella’. The `<st:variant>` tag asserts, that the term ‘Fisichella’ was observed with sufficient relevance only in time slice number 3. This node is semantically mapped into LTM[4] (Linear Topic Maps Notation) as follows. The `<st:label>` tag is mapped to a Basename and the `<st:ref_word_nr>` and `<rdf:id>` tags are mapped to Subject Identifiers. This Subject Indication determines that if a noticeable usage of the same term is detected by the same technique twice, Subject Equality holds. All information about a time slice represented by each `<st:variant>` tag is represented by a single Topic which is connected to the Topic ‘Fisichella’ with a typed Association.

```
[id7406 : id7276 = "Fisichella"
    @"http://www.texttech.de/dtd/st/pap#node_160615"
    @"http://www.texttech.de/dtd/st/pap#node_Fisichella"]
   {id7406, id7650, [[160615]]}
id7549( id7406 : id463, id464 : id2195 )

[id464]
   {id464, id1636, [[0.31766722453166335]]}
   {id464, id4378, [[3]]}
   {id464, id787,  [[4]]}
```

Similar semantic mappings are done for the edges. In the example, the only edge of the node ‘Fisichella’ is a relationship to the node ‘Schumacher’. Due to the success of Michael Schumacher in several competitions it is not surprising, that in contrast to ‘Fisichella’ the term ‘Schumacher’ was observed in all time slices.

---

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

[4] http://www.ontopia.net/download/ltm.html

4.3 Integration of background knowledge with the SemanticTalk TMV

As shown in figure 2, the last step is the integration of background knowledge (represented as TMVs) with the SemanticTalk TMV. To remind again, the perspective of the generated SemanticTalk TMV is defined as follows: a Subject Proxy documents a noticeable usage of a term detected by SemanticTalk. Subject Identity is given, if this observation of noticeable usage of the same term is done twice by the same algorithm at different occasion.

For integration purposes, the perspective of the integration has to be defined. From the integration perspective the noticeable usage of a term observed by SemanticTalk should be interpreted as the observation of a key concept of the conversation.

For example, the SemanticTalk TMV should be merged with a hand-crafted Topic Map about motor sports. This edited Topic Map might have a Topic which documents observations about Giancarlo Fisichella, the formula one driver. This Topic uses the Published Subject Identifier (PSI) "http://www.formula1-fansite.org/Fisichella" for Subject Indication. In the example the integration perspective is defined as follows: whenever information about the person Giancarlo Fisichella is observed the Subject 'Fisichella' is caught.

From this perspective, Subject Identity holds between the observations documented in the 'Fisichella' Topics of both Topic Maps. This decision should be documented in an Integration Topic Map. The creation of this Integration Topic Map should be governed by the Observation Principle, too.

According to the Observation Principle, an Integration Subject Proxy has to document the decided Subject Identity with the means of Subject Indication provided by the governing SMD. From the current integration perspective, the same Subject is caught, if SemanticTalk observes a noticeable usage of the term 'Fisichella', and if the author of the hand-crafted Topic Map indicates a Subject by the PSI for Giancarlo Fisichella. An appropriate Integration Topic Maps looks as follows:

```
[id @"http://www.formula1-fansite.org/Fisichella"
    @"http://www.texttech.de/dtd/st/pap#node_Fisichella"]
```

The merging of all three Topic Maps (the SemanticTalk TMV, the hand-crafted Topic Map and the Integration Topic Map) governed by the TMDM integrates all relevant information automatically. The advantage of the Integration Topic Map is obvious: in the case the integration perspective changes, the Integration Topic Map can be switched off or changed accordingly to the alterations in the integration perspective. It is apparent that in this connection the concept of Observation and Integration Subject Proxies introduced in 2.2 is applied on a higher semantic level.

For the integration of distributed sources the usage of the Topic Maps Remote Access Protocol (TMRAP) [8] or similar techniques discussed in [10] is recommended.

In most cases, the automatic generation of Integration Topic Maps is not straightforward. In these cases the application of heuristic methods for the detection of Subject Identity respectively Subject Equality [11], [10] is recommended.

5. Related Research

There exists a substantial body of work on text classification and topic or metadata extraction from multimedia data (information extraction). All of these approaches define their own perspective in respect of the Observation Principle.

Research on sophisticated real-time indexing of speech streams in group meetings as one possible application has so far been very limited. Jebara et al. [9] describe a system that classifies the current focus of the conversation according to a limited number (12) of pre-defined topics. The classifier is trained by initially providing training sets of documents associated with each topic. Real-time analysis of spoken conversation is also reported in DiMicco & Bender [4]. Their focus is on facilitating equal participation in group discussions by visualizing the contributions of each participant. Terms and sentence fragments are associated with a fixed number of categories by a classification component also based on machine learning. None of the above approaches, however, builds up a concept structure by relating the terms spoken. Also, no extraction of semantically related terms is reported.

To our best knowledge there exists no approach that focuses on Topic Maps, fostering the integration of background knowledge to the current conversation. To our best knowledge, there exists no deeper theory related to the creation of Topic Maps, similar to the Observation Principle proposed by this paper. The Observation Principle exclusively focuses on the semantic characteristic of Topic Maps Technologies: viewing Subject Proxies indicating identical Subjects as merged ones.

The work presented in this paper is the advancement of our previous research described in [3].

6. Conclusion and Outlook

We discussed how Topic Maps can be generated from speech streams in real-time by using SemanticTalk. Furthermore, we introduced the Observation Principle based on a deeper investigation of the Subject Equality Decision Chain.

This investigation showed that each automatic generation of Topic Maps is governed by a specific perspective for the observation of digital information collections. The resulting Topic Maps are only the documentations of the observations from these perspectives. It is important to outline, that the interpretation of these observations has to be done in the time of integration with other Topic Maps.

In future, large repositories of Observation Subject Proxies should be investigated by statistical means to detect emergent relationships and concepts inside.

We assume, that it is important to disclose for each Topic Map how the Observation Principle was exactly applied during its generation. This knowledge allows more accurate interpretation in the time the generated Topic Map is used for integration purposes. We foresee theses disclosures as an important feature for the further success of Topic Maps in information integration scenarios. The disclosure of the Topic Map Engineering Process (TMEP) as sketched in [16] is the consequential development in respect to these insights. We assume that further corresponding research will be the advent of the emergence of the scientific discipline *Topic Maps Engineering*.

References

[1] Biemann, C.; Böhm, K.; Heyer, G.; Melz, R.: Automatically Building Concept Structures and Displaying Concept Trails for the Use in Brainstorming Sessions and Content Management Systems, In: Proceedings of I2CS’04, Guadalajara, Mexico; Springer LNCS, (2004).

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

[3] 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; pp. 241-251, (2004).

[4] DiMicco, J. M.; Bender, W.: Second Messenger: In-creasing the Visibility of Minority Viewpoints with a Face-to-face Collaboration Tool. In: Procceedings of Conference on Intelligent User Interfaces, Funchal, Portugal; ACM Press, (2004).

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

[6] Faulstich, L.; Quasthoff, U.; Schmidt, F.; Wolff, C.: Concept Extractor – Ein flexibler und domänen-spezifischer Web Service zur Beschlagwortung von Texten. In: Hammwöhner, R; Wolff, C.; Womser-Hacker, C. (eds.): Proceedings of ISI 2002. Schriften zur Informationswissenschaft 40; Hochschulverband für Informationswissenschaft 2002, Regensburg, (2002).

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

[8] Garshol, L. M.: TMRAP – Topic Maps Remote Access Protocol. In: Proceedings of First International Workshop on Topic Maps Research and Applications (TMRA'05) Leipzig; Springer LNCS, (2006).

[9] Jebara, T.; Ivanov, Y; Rahimi, A.; Pentland, A.: Tracking conversational context for machine mediation of human discourse. In: AAAI Fall 2000 Symposium - Socially Intelligent Agents, Massachusetts; AAAI Press, (2004).

[10] Maicher, L.: Topic Maps and the Absence of Shared Vocabularies. In: Proceedings of First International Workshop on Topic Maps Research and Applications (TMRA'05) Leipzig; Springer LNCS, (2006).

[11] Maicher, L.; Witschel, H. F.: Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM). In: Proceedings of LIT'04; pp. 229-238, (2004).

[12] Pepper, S.; Vitali, F.; Garshol, L. M.; Gessa, N.; Presutti, V.: A Survey of RDF/Topic Maps Interoperability Proposal. W3C Consortium Working Draft. Available at: http://www.w3.org/TR/2005/WD-rdftm-survey-20050329/

[13] Quine, W. v. O.: Identity, Ostension, and Hypostasis. In: Journal of Philosophy, 47(22), pp.621-633, (1950).

[14] Russell, S.: Identity Uncertainty. In: Proceedings of IFSA-01, Vancouver; (2001).
[15] ISO/IEC: Topic Maps – Part 2: Data Model. Latest version available at: http://www.isotopicmaps.org/sam
[16] Sigel, A.: Report on the open space sessions. In: Proceedings of First International Workshop on Topic Maps Research and Applications (TMRA'05) Leipzig; Springer LNCS, (2006).
[17] Vatant, B.: Tools for semantic interoperability: hubjects. Working report available at: http://www.mondeca.com/lab/bernard/hubjects.pdf
[18] Witschel, H. F.: Terminologie-Extraktion: Möglichkeiten der Kombination statistischer und musterbasierter Verfahren. Ergon, Würzburg, (2004).