Fit for the future – a semiautomatic growing ontology to enhance university needs

R. Böving[1], U. Bleimann[1], C. Wentzel[1], P. Walsh[2]

[1] University of Applied Sciences Darmstadt, Darmstadt, Germany
[2] Cork Instiute of Technoloy, Cork, Ireland
roland@boeving.net

Abstract: Semantic technologies are an important tool for developing structured knowledge, for linking all types of content sources and making them available for evaluation. Knowledge is thereby developed through a network of topical terms, which, in themselves, are interrelated. The academic environment poses a special challenge for semantic knowledge development as its topic environments are often heterogenic and expand rapidly. The essential question is:
Which topics are the focus of students and university staff and how can access to the required data be provided in a structured manner?

By merging and mapping existing ontologies, the University of Applied Sciences Darmstadt has access to topic network comprising more than 200,000 terms that thematically link lectures, final papers, professional articles and internal publications, allowing these to be evaluated.

1 Introduction

By merging and mapping existing ontologies, the University of Applied Sciences Darmstadt has access to topic network comprising more than 200,000 terms that thematically link lectures, final papers, professional articles and internal publications, allowing these to be evaluated [Boe10].

In the course of merging the ontologies to form one ontology which is the basis for a semantically evaluable database, the references to the linked content of the original ontologies should not be lost. My approach creates a new combined topic landscape which nevertheless remains linked to the original ontologies. This topic landscape will continue to grow, firstly as a result of updates to the original ontologies, and also as a result of new suggestions of topics for lectures, specialist articles and theses.

In addition, the entire university structure is stored in semantic objects. Employees and students are linked to their faculties and publications. The evaluation of semantic

relationships between topics, publications and persons will provide the university with the opportunity to build an expert database in the future.

Precise terminological designations are essential for ontologies used by universities, where the specialist vocabulary must be of sufficiently high quality to ensure that topics are linked with professional articles, for example. Specialist terms are also continually created in this case, which only specialists are able to understand and place in their correct context. Preservation of quality and internal consistency, as well as providing a sufficient number of relationships to meet user expectations, pose the greatest challenges in the maintenance of ontologies used for scientific purposes.

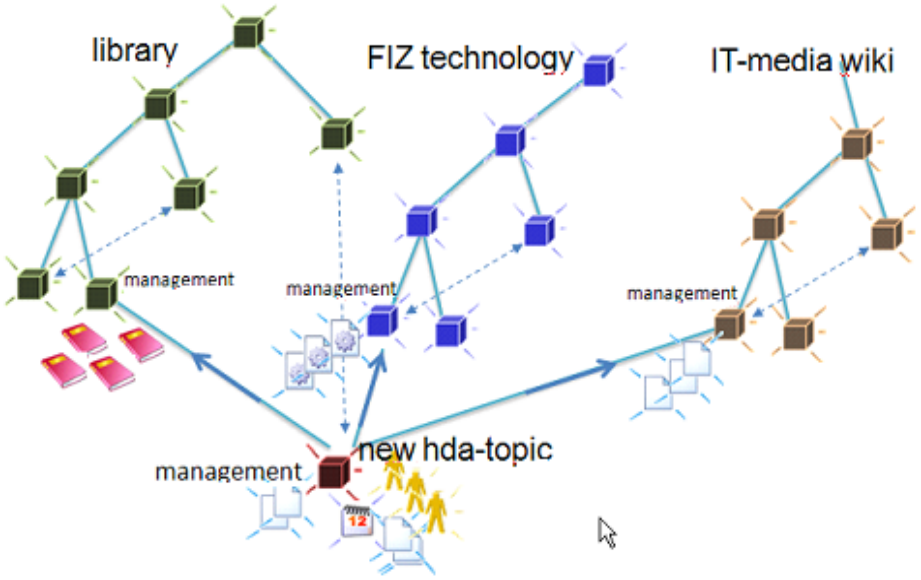

Figure 1: The solution – besides the independently imported ontologies, the ontologies already existing will generate a new h_da ontology, which will contain each knowledge object once. All additional content sources can be accessed via this h_da topic map, without impairing their structures.

2 Approach

Ontologies dedicated to technology and natural sciences are already in use in many large companies today. The benefits of providing and exchanging information, in a web-based format independent of location, by means of ontologies are indisputable. In contrast to the distinct commercial focus and product range of companies, the range of topics dealt with in universities at the highest level is very broad. The difficulties of building and maintaining an ontology are disproportionately greater. Different meanings and interpretations of terms and abbreviations will make automatic matching more difficult. In the social sciences, too, new words will certainly evolve over time. The precise language of engineers is confronted by a world of relatively loose terms. There are as

yet no ontologies in the university sector in Germany which would be sufficiently comprehensive to link all sources of knowledge in an interdisciplinary manner, in order to provide a real solution for everyday operation. This work will partly be devoted to this issue, and it leads to a further complex problem that has not yet been tackled, namely: How can a complex multilingual ontology, for use in a university context, be maintained with respect to its growth, quality and consistency so that it will meet future needs?

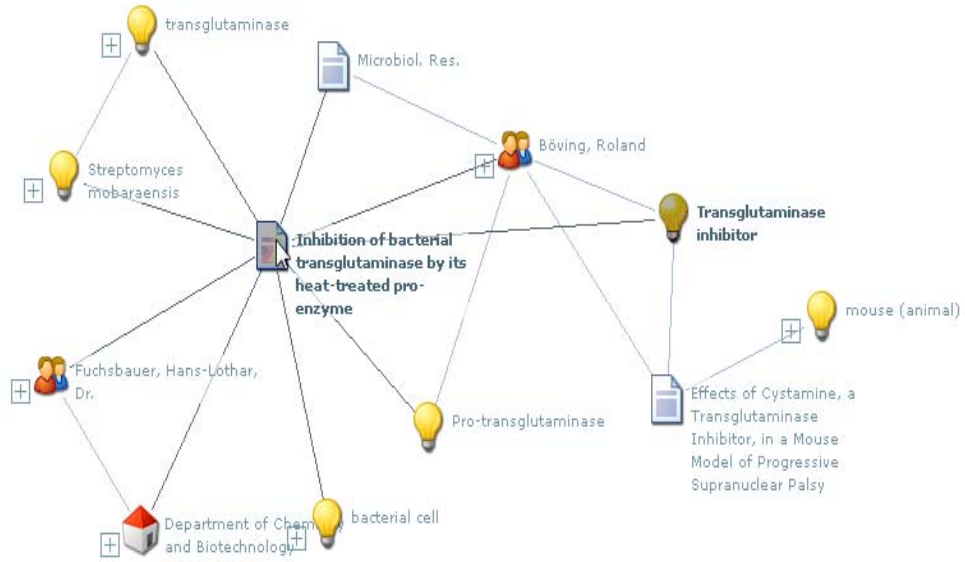

Figure 2: h_da-semantic web - showing relations between topics persons and content.

3. What future tasks does this reveal to be necessary?

The maintenance and the expansion of the fundamental ontology "h_da topics", which can be seen as the heart of the system, should be given particular attention in the future. If searched topics for tagging an article are not present, or if topics are insufficiently linked with one another, then this will have an immediate effect on the acceptance of the entire system. Maintenance of the topic area is essential for the system to remain attractive for an academic context in the future, too. In order to achieve this, it is necessary to design as many automatic or semi-automatic processes as possible for maintaining and expanding the system. This includes regular topic updates of the original ontologies FIZ-Technik and h_da library. The result of this is that h_da topics also performs a merging and mapping update and is thus kept up to date.

In spite of the updates, to begin with many newly proposed topics for tagging articles and theses will not appear in h_da topics, and will therefore exist initially without any

further linking. The aim must therefore be to create a significant, logical semantic link between all newly proposed topics and the existing terms in h_da topics. This will likely lead to better semantic search results, not only for new topics but overall.

4. Comparison of newly proposed topics with external sources

In order to achieve this, the aim is to perform a direct comparison with an external, free and fast-growing ontology, in which it can be expected that new topics will quickly be proposed and will subsequently be discussed by an international community and internally linked. One international reference ontology which lends itself to this is Wikipedia, as it has been shown that not only general topics but also specialist scientific terms are represented here in sufficient depth and with detailed descriptions in the content. There will likely be good matches for technical and scientific terms here. However, it must be expected that other fields such as social sciences or economics will increasingly see newly coined words which are not connected to a nomenclature. Many of these 'soft' new topic proposals can, with the necessary background knowledge, very probably be assigned to an existing topic as a new synonym. Some of the newly coined words will in the future actually be considered to be distinct concepts. In addition to newly coined words, articles will also be tagged with abbreviations. In order to solve this problem in part, a comparison of newly proposed topics with the ontology of the DUDEN publishing company could be conducted. The Duden publishing company is the largest centre of competence in the German-speaking world and has an elaborately updated ontology of all German terms and synonyms. The database contains more than 500,000 topics and is designed to filter out linguistic relationships. A cooperation has already been offered by the Duden publishing company, which likewise maintains its database using K-Infinity and generates 260 new book releases from it annually.

The comparison of newly proposed h_da topics with the corresponding topics in Wikipedia leads in the first instance to the following realisation: is the topic already a real term which is used, or perhaps a newly coined word? If the compared topic already exists, then the obvious thing is to examine the content recorded for it on Wikipedia for categories and linked topics and to use these for independently networking the new h_da topics. Moreover, the majority of Wikipedia topics are stored in multiple languages. The language versions here are not translations of a main language, however, but rather are independent of one another. It may therefore be the case that a description in German contains more links than an English one and vice versa.

A multilingual approach has the advantages of achieving a greater depth of linking and also of extracting linguistic synonyms in order subsequently to complement the translations of the h_da ontology. This would provide the universities with an ontology which exists initially in bilingual form in all scientific and technical fields and could make an important contribution to cooperation between European universities.

5. Solution and requirements:

- creation of a hybrid system for the maintenance and further development of h_da knowledgeworld
- open for new applications/document types
- open for new topic areas
- open for new user groups
- the topic network 'h_da topic' should meet the requirements of the departments
- consistent, controlled quality
- internally consistent data scheme
- multilinugally conceived topic network: initially German/English
- regular topic updates from checked sources
- comparison of newly proposed topics with external sources and automatic networking with recognised topics already present
- semi-automatic proposal scheme for classifying language versions and abbreviations as synonyms

6. Validation of newly proposed topics by comparison with external resources

Uploading theses and internationally published articles produces the following initial situation. Not all of the necessary keywords for tagging the documents also have equivalent topics available in h_da topics. These keywords are set up as 'new topics' in the ontology 'h_da knowledgeworld'. As a result, all documents can be fully tagged. Over time, a large number of newly proposed topics accumulate under the topic node 'new topics', which are not validated and are not connected to any other topics in the topic network. Each of these 'new topics' has one or more relations with available documents. Moreover, the proposed topics are already available in two different languages and in many cases also as abbreviations. Fig. && shows a schematic drawing of a recent extract of proposed topics. It is obvious that some topics can be classified as synonyms of topics which already exist. It is also noticeable that very new topics from the English-speaking world, such as 'smartphone', have been incorporated into the German language. In principle, only those topics are represented in h_da topics which also have a real relationship to a book or to another publication. If there has until now been no article on Facebook, then there is also no corresponding 'facebook topic' in the network.

Figure 3: Typical example of newly proposed topics, of which some English topics obviously already have their equivalents in the German topic network and must therefore be entered as linguistic synonyms. The majority, however, do not exist in a foreign-language version in h_da topics.

A comparison with corresponding search terms in Wikipedia must first evaluate the categories and links of the first section of the description and second perform this process automatically for the second language variety to be validated. Searching through further sections of the Wikipedia descriptions is less worthwhile, because too many references are generated, which cannot be put into any meaningful context without the corresponding test. The comparison of the Wikipedia equivalents not only generates important relationships with relevant terms in German and English but also provides all of the synonyms of the core term. If identical terms are found in Wikipedia, as in Figure &&, the 'new topics' which initially were only proposed can then be transferred with their synonyms and relations to 'h_da topics'. Furthermore, existing topics could be expanded with the linguistic synonyms additionally found. Topics found which are not unambiguous – such as 'apple' or 'android', for example, which could also be referring to a company or to a product respectively – must be excluded for the time being. The fact is that not all relations represented in Wikipedia can be used. Overall, linking the topics from the first section of the description in Wikipedia will lead to a significantly better definition of the h_da data.

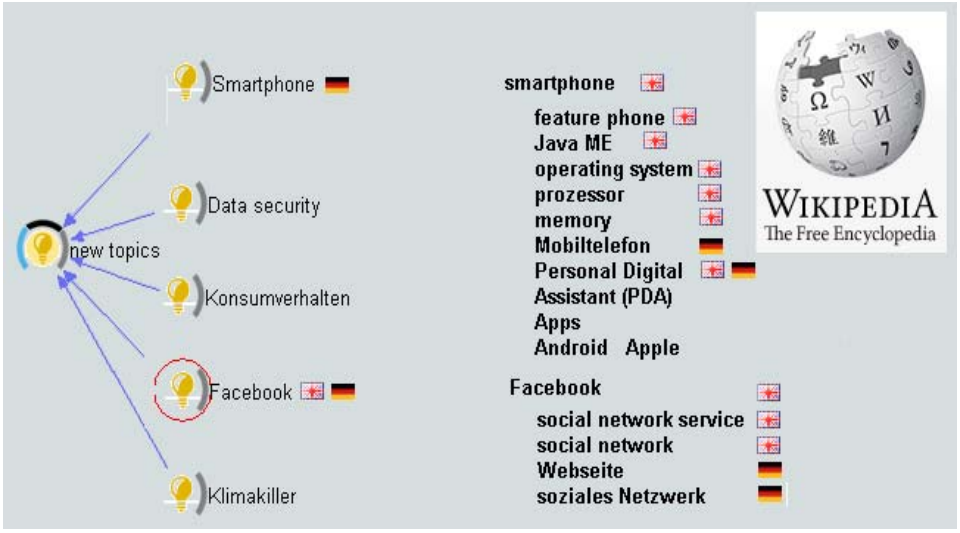

Figure 4: Results of a comparison of the newly proposed topics 'smartphone' and 'Facebook' with the first section of the description in Wikipedia, and extraction of the linked topics from the English and German versions of the description.

7. Research Plan

1. Implementation of a hybrid system which compares searched topics during the search process with equivalent topics from external international sources and automatically or semi-automatically transfers these topics to the university's ontology. In the process, as many relations as possible should also be created with other topics in order to expand the h_da topic network with important semantic relationships.
2. Validation of the comparison with different language versions
    - Quality audit relating to corresponding extracted topics from external sources
    - Quality audit relating to extracted master categories from external sources
    - Quality audit relating to extracted links from the first section of the description
3. Validation of the comparison with relationships between words
4. Validation of newly created semantic relations in the topic network of h_da
5. Validation of the semantic search results following integration of a hybrid comparison and correlation with the original condition.
6. Validation of a university-wide integration in a European context
7. Discussion of the current problems and limits involved in maintaining a fast-growing ontology in a university context and the resulting prospects, opportunities for development and challenges.

References:

[BL01] Berners-Lee, T.; Hendler, J.; Lassila, O.: “The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities”. Scientific American, May, 2001.

[Boe10] Böving, Roland; Bleimann, Udo; Wentzel, Christoph; Walsh, Paul: Semantic search scenarios to enhance University needs - based on a multilingual Next Generation Topic Map (NGTM), In: Conference Proceedings of the 5th Plymouth e-learning conference 2010 - Learning without limits: Facing the Challenges”, ed. Steve Wheeler, Plymouth 2010, 14

[Boe10] Böving, Roland; Bleimann, Udo; Wentzel, Christoph; Walsh, Paul: High Level Semantic Networking - Using K-infinity to Build a Multi-Ontological Learning Environment In: Proceedings of the 8th International Network Conference (INC 2010), ed. Paul Dowland und Steven Furnell, Plymouth 2010,

[Eh04] Ehric, M.; de Bruijn, J.; Manov, D.; Martin-Recuerda, F.: State-of-the-Art survey on Ontology Merging and Aligning. V1 SEKT Deliverable 4.2.1, DERI Innsbruck, 2004.

[Euz07] Euzenat, Jerome; Shvaiko, Pavel (2007): Ontology Matching; Springer

[Fa00] Falge, Clarissa; Cobos, Ruth; Groh, Georg (): Classification and Ontology Maintenance in Agent-based Knowledge Management Frameworks: A Prototypical Approach

[Fri00] Fridman Noy, Natalya; Musen, Mark (2000): PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment

[Hi06] Hitzler, Pascal; Krötzsch, Markus; Ehrig, Marc; Sure, York (2006): What Is Ontology Merging? – A Category-Theoretical Perspective Using Pushouts

[Ka09] Kayalı, Can (2009) Hybride, partielle OWL-Ontologieabgleich für Dienstselektierung im semantischen Web

[Pi09] Pinkwart, Niels; Dicheva, Darina; Mizoguchi, Riichiro (2009): Ontologies and Social Semantic Web for Intelligent Educational Systems. In conjunction with the 14th International Conference on Artificial Intelligence in Education Thistle Hotel, Brighton, July 6th-10th, 2009.

[Rei10] Reichenberger, Klaus: Kompendium semantische Netze: *Konzepte, Technologie, Modellierung*, Springer, Heidelberg 2010, ISBN 3-642-04314-3

[Fo08] El Jerroudi, Zoulfa; Ziegler, Jürgen (2008): Interactive Ontology-Mapping and Merging