IFToMM-TERMINOLOGY – BASE FOR AN EFFICIENT INFORMATION RETRIEVAL IN DIGITAL LIBRARIES FOR MECHANISM AND MACHINE SCIENCE

Torsten BRIX [(1)], **Burkhard CORVES** [(2)], **Ulf DÖRING** [(1)]
(1) Ilmenau University of Technology, Postfach 100565, D-98684 Ilmenau, Germany
(2) University RWTH Aachen, Germany
torsten.brix@tu-ilmenau.de, corves@igm.rwth-aachen.de, ulf.doering@tu-ilmenau.de

1 ABSTRACT

Over the years the IFToMM dictionaries have been grown up to a comprehensive digital resource which can be used as dictionary as well as glossary with four languages in parallel. Regarding today's needs of information retrieval in digital libraries it makes sense to see the IFToMM-terminology dataset as a semantic network which can be used to solve different retrieval problems in digital libraries providing content on mechanisms and machine science. The article presents the DMG-Lib as such a digital library and shows, how the content of the IFToMM dictionaries can be used to improve the retrieval process. Furthermore suggestions are made, how the IFToMM dictionaries could be extended to be more powerful and how the generation and maintenance process of the IFToMM dictionaries could be supported by online tools.

Keywords: Terminology, Digital Library in MMS, Information Retrieval, IFToMM dictionaries

2 INFORMATION RETRIEVAL IN DIGITAL LIBRARIES

In several recent projects digital libraries are built up. Some of them contain content about mechanisms and machine science (e.g. [1], [2], [3]) in different form, for instance textual sources, photographs and videos of physical models, abstract model descriptions, CAD models, animations, links to software etc.
The content is heterogeneous according to its form (and its type respectively) as well as the language in which texts and according textual information like metadata are given. Especially multilingualism of the textual content causes a large problem for information retrieval, because the retrieval is usually implemented by means of a string matching algorithm which mostly fails when the languages of the searched text and the stored text differ. But not only differences between languages must be handled adequately. Even within one language there are usually differences depending on the time period the texts were written, on the school the authors attended and on the community the authors belonged to. On the other hand each user of a digital library has his specific background (e.g. scientific skills) which results in terms the user could select for his search queries. Therefore an adequate information retrieval is only possible with knowledge about the relations between the different terms (used by authors and users).
The DMG-Lib project is a digital library and information system in the field of mechanism and machine science (MMS) which aims to solve problems as described above. At the moment textual documents, biographical information, interactive animations, mechanism descriptions, videos, images and software descriptions are shown on the project web pages (www.dmg-lib.org), see fig. 1. Most of the textual content and metadata is still in German (87%). However, texts in other languages can also be found, e.g. English (9%), French, Spanish and Serbo-Croatian. Nevertheless the number of languages a user can sufficiently understand is of course limited. For some people it might be questionable if it is worth to give users the access to textual documents in a foreign language. Here two facts have to be taken into account:

- For most of the academic users as well as engineers (but also laymen or pupils) it makes sense to get texts related to MMS in foreign languages, because images, formulas, links to other resources or key words in the surrounding text often contain usable/understandable information.
- The number of languages (content languages as well as user languages) supported in the DMG-Lib will increase in follow-up projects. In this way, multilingualism becomes important for much more than the languages currently supported in the IFToMM dictionaries.

Therefore, in our case it is worth that information retrieval is designed in a way that it can serve as an intermediate layer which matches the language of the user with the language of the content to generate suitable/helpful search results. One approach to implement a multilingual retrieval is based of a semantic network.

Fig. 1.: Different types of content in the DMG-Lib

3 SEMANTIC NETWORK AS A TOOL TO SOLVE RETRIEVAL PROBLEMS

What is a semantic network?

A semantic network consists of certain concepts/topics (named by means of according terms) and relations between those concepts. Such relations have certain semantics. The most general semantics is "*is related to*". More special semantics are "*is part of*", "*is kind of*", "*is a synonym for*", "*is subtopic of*" etc. Semantic networks can be seen as graphs where the concepts are the nodes and the relations are the edges. Beside general information about concepts and relations semantic networks usually include further information, e.g.:

- textual definitions of concepts (terms) and graphical examples which explain the concept and support the comprehensibility,
- information about contexts (scopes) in which names and relations are valid (e.g. to handle changes of terms in the course of time),
- roles that the concepts play in certain relations and
- occurrences of the concepts (e.g. a related page in the internet).

Summarizing the above, semantic networks are used for knowledge representation where the degree of detail can be adopted to the knowledge space and the problem which has to be solved.

How can it be represented?

In general it is possible to use a proprietary representation within an own implementation of a semantic network software. Nevertheless it makes sense to store the network in a standardised format or provide at least a conversion tool. That supports the reuse of the network in other projects and ensures the possibility to use common semantic network tools to maintain or use the own network. An often used standardised format for the implementation of semantic networks is called Topic Maps [4].

Which problems can be solved with a semantic network?
There are a lot of possibilities to use the knowledge stored in a semantic network during information retrieval. Some examples of problems and according solutions are given in table 1.

Problem	Solution
A user is not satisfied with the received information. He wants other related information.	After the identification of the current topic (e.g. by looking for terms in the semantic network) the user can get an overview about related topics (e.g. as a textual list or in a graphical view).
The search for certain terms produced no/too view results.	Find related topics and use according terms to enlarge the search query.
There are too many search results listed.	Results should be grouped. During the grouping process the knowledge stored in the semantic network can be used, e.g. when the classification criteria are stored in it and the entries in the search result list match to values of the criteria.
A user wants to get information in another language.	As far as terms in the scope/context of this language are stored these terms can be used to generate the output.
The content includes technical terms which are unknown to the user or the text is written in a foreign language.	When terms are included in the semantic network then according synonyms and textual as well as graphical definitions could be displayed (e.g. by means of a "mouse over"-technique). If more than one piece of additional information is available then the one with the best fitting scope/language could be selected.

Table 1: Information retrieval problems and according solutions which are based on the use of a semantic network

4 THE IFToMM TERMINOLOGY AS A SEMANTIC NETWORK

Remarks to the current state of the IFToMM dictionaries
The current content of the IFToMM dictionaries includes information which can serve as a basis for a semantic network. In fact it is a semantic network when the topics (the technical terms which are identified by numbers) are seen as nodes which contain the definitions in a certain scope (one of four languages) and the html-links are seen as relations.
Analysing the dictionaries in more detail one can see:

- That different chapters of the IFToMM dictionaries (e.g. Chapter 1 for "*Structure of Machines and Mechanisms*" and Chapter 2 for "*kinematics*") define scopes. The term "*pressure*" is an example for a homonym which is defined in different scopes (see IFToMM Chapters 7 and 12).
- That IFToMM dictionaries combine the advantages of a glossary and a dictionary and are a very helpful instrument (Figure 2). At the moment it is especially useful for:
    - Manual translation of texts related to mechanisms and gears (can be used as a dictionary)
    - Understanding the meaning of certain terms (can be used as a glossary)
    - Promoting the standardisation process (e.g. translators will use the terms in the dictionary)
- That context information for certain terms is often given, e.g. by the placement in a certain section or by the sentence where the word occurs.
- That the quadrilinguality (English, French, German and Russian) of the IFToMM dictionary is partly incomplete. This problem will later occur more often, as soon as more than 4 languages are provided. Algorithms must consider this fact to be robust.

Additional features which could be useful for information retrieval
As described in Chapter 3 a semantic network is able to store information which can be utilised during the information retrieval process very successfully. To reach this it is necessary to include more information in the semantic network than the

information which can be currently extracted from the IFToMM dictionaries. Regarding the demands of information retrieval following would be desirable:

- Definition of classification criteria
  - Such a definition must include a property which serves as criteria and possible values.
  - „*Dimension of a mechanism*“ may serve as an example for a property and „*planar*“, „*spherical*“ and „*spatial*“ as possible values for this property.
  - That would be for instance usable for browsing. Here a subdivision of a search space and of search results according to certain points of views could be applied.
- Unique matching of important terms in the sentences which are written in the different languages.
- More information according to the grammar, e.g. singular and plural as well as articles, which may be important for translations.
- Images and occurrences which explain the definitions in more detail and show the context in which certain terms are used.
- Some hints, how errors can be avoided – e.g. hints discussing the correct use of synonyms or hints which show how to distinguish between homonyms.
- Less formal or obsolete terms should be linked as synonyms to the according terms in the dictionaries. It should be recognizable that the terms are less formal or obsolete. In this way both can be reached – the users can find good terms (the dictionary provides synonyms which should be used in publications) and the users can find matching definitions more often (the glossary can explain old terms too).

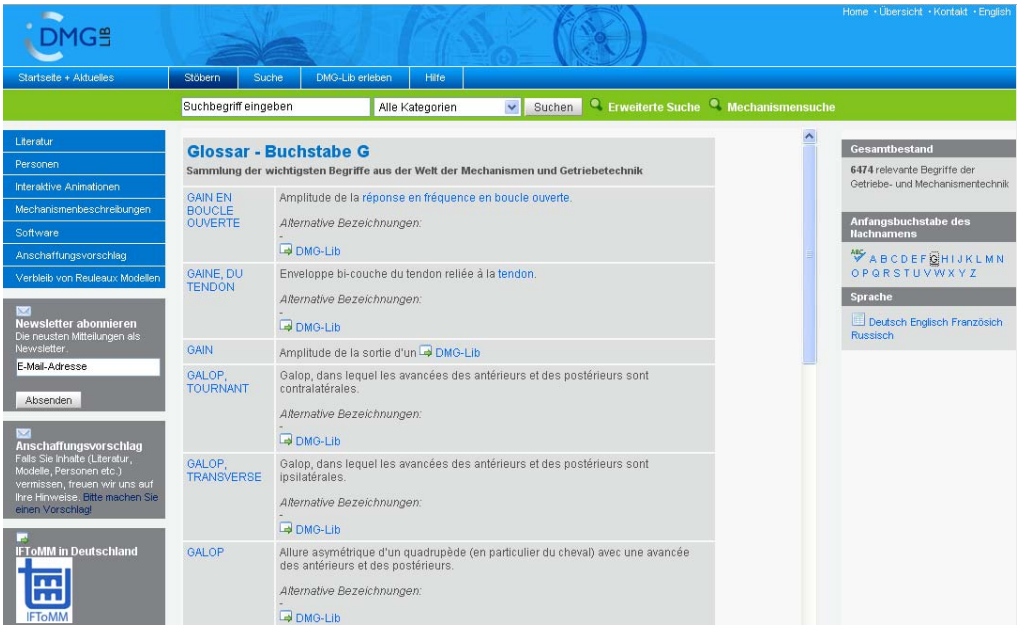

Fig. 2.: The IFToMM glossary in the DMG-Lib

5 POSSIBILITIES OF A SEMANTIC BASED INFORMATION RETRIEVAL

5.1 Support of guided information retrieval

The semantic network can be used as intermediate layer between users on one side and information sources on the other side. In this case it allows an efficient information retrieval in different ways. The following list is not complete, but it shows some possibilities how users can be supported.

- **Fast input of search terms**
  Some entered characters (e.g. “*Mob*”) are used for a comparison with technical terms stored in the semantic net. This may result in an automatic expansion of the search term or in a list of possible terms (e.g. “*mobility*”, “*mobile*”, “*mobilité*”, “*Mobilität*”). The scope and the order of the items in such a list should be controllable. Possible criteria are preferred language, preferred domain etc.

Comparisons between entered characters and terms in the semantic network can be executed with different algorithms, e.g. fuzzy similarity, equality of first characters etc.

- **Detection of appropriate search terms**
  Usually the users are forced to know or guess a fitting search term to find specific information in a knowledge base. This can lead to time consuming trial and error iterations. However, a semantic network can help users by proposing other (better) terms. It is sufficient when a user knows a term near by the search area (e.g. "*mechanism*"). With the aid of the semantic network all terms in relation to the entered search term (meronyms, hypernyms etc.) can be displayed. That allows the user to select a better search term for the search query or to display all related terms of the currently selected better term (e.g. "*dwell mechanism*", "*step mechanism*").
- **Controlled search**
  Please imagine a map with its areas, cities, villages and streets between them. The controlled search in a semantic network shall be explained now using such a map as metaphor. The cities and villages in the map represent important as well as not so important nodes in the semantic network and the different types of streets represent the different significance of relations between the nodes. If a user must find a way from a city A to a city B, he must decide what kind of way he wants to go, e.g. the shortest way, the time optimal way, a way with some intermediate stops and so on. Similarly the semantic network helps to navigate from one search term (e.g. "*kinematic chain*") to another search term (e.g. "*gear train*"). During the trip the user learns about further appropriate search terms (e.g. "*planetary gear*", "*geared linkage*"). The map metaphor is the base for graphical navigation tools.
- **Automatic expansion of database queries**
  If a search produces no results (e.g. search for "*Koppelgetriebe*") or users are not satisfied with the results then new search terms could be added automatically. Such terms may be synonyms in the language of the search term (e.g. "*Gliedergetriebe*", "*Kurbelgetriebe*") or synonyms in other languages (e.g. "*linkage*"). Of course the automatic use of related terms should be adjustable. In this way automatic search query expansion allows a successively wider search and (if wanted) the retrieval of foreign-language information sources (full text as well as metadata).
- **Logging of search queries and search paths**
  The (anonymised) recording of search queries (terms and according concepts) and search paths (used relations between the concepts) during the navigation/search process can be used as a basis for empirical studies. This helps to find out user intentions and can be used to derive new search heuristics and to support a controlled search.

5.2 Increasing the readability and comprehensibility of texts

The provision of definitions and descriptions of terms (in textual as well as in graphical form) results in a considerable better readability and comprehensibility of texts. For instance it is possible to mark all terms in a text which occur in the semantic network. For unknown terms the user can force the presentation of definitions and descriptions, e.g. by means of tool tips or a JavaScript window. Such techniques enable easier understanding of the content in known languages (when unknown terms are used) as well as in unknown languages, because the language of the displayed definitions and descriptions is adaptable to the needs of a certain user and to the scope of the technical term in a certain text. For the selection of the correct scope metadata like author or publishing date can be used.

5.3 Controlled Semantic Tagging

For several user groups (engineering designers, laymen), for several domains (medical technology, sports equipment engineering), etc. often well-established terms exist which do not occur in the IFToMM dictionaries. This hinders the information retrieval in particular for non-experts in MMS. In such cases it becomes custom in the internet that the users are allowed to attach their new terms to the according information objects (documents, pages, videos etc.). This process is called tagging and the tags are a kind of new metadata. The following three phases describe the process of semantic tagging and especially the advantages of a controlled semantic tagging.

Phase I:
A user finds an interesting information object (e.g. a photograph, a book, a paper) and attaches a tag (e.g. the term "fidget equipment" to a certain mechanism on a photograph). In general the user can select the tags from his own context/background (terms from other scientific domains, colloquial language, slang) regardless of the set of preferable technical terms. Because tags are searchable they may directly support the information retrieval for non-experts.

Phase II:
A group of experts checks the user-defined tags (terms) to detect meta information which is suitable to expand the semantic network. The tags (terms) selected by the experts become part of the semantic network. A scope shows the context in which the new terms are valid (e.g. certain scientific domains, colloquial language, slang). Now a controlled tagging was carried out, because experts have evaluated the tags and made a choice. This results in a domain overlapping dictionary and glossary with connection to colloquial language.

Phase III:
Users from different user groups can benefit from the tags on different ways. For instance they can search on the set of controlled tags (stored in the semantic network) as well as search on the set of controlled tags (not part of the semantic network). Searching on the set of controlled tags additional information (e.g. scopes) may improve the search results. Because controlled tags became an integral part of the semantic network, they are available for all the improvements described above (see section 5.1). This allows an easier access to domain specific sources of mechanism and machine science for all user groups.

6 Collaborative development of the semantic network

A more general dictionary (or a semantic network with information as described in the previous chapters) could be build up by the integration of terms and relations from different sources. Because of the outstanding quantity and quality of the current IFToMM dictionary it would play a key role in the selection of the initial content. Possible other sources are:

- Originally printed dictionaries, e.g. [5]
- International and national standards, e.g. [6], [7] and [8]
- Online dictionaries like [9] and [10]
- Online word lists, e.g. [12]
- Databases like internal database from www.dmg-lib.org

A certified subset of terms and definitions could be selected as the content of the new release of the IFToMM dictionaries. For the semantic network expansion a collaborative development is useful. For this, several tools and features are needed. The following subjects below give an overview about necessary tools and features:

Expansion of the semantic network

- Use of a web-based editor tool helps to a collaborative work in the semantic network
- Consideration of user rights management (e.g. which experts have the right to change the terms in a certain language)
- Generation of word tables with language specific columns (e.g. editable for maintenance or non-editable for dictionaries and glossaries)
- Visualisation tool for the semantic network

Communication between maintainers

- Establish different internet-based communication links (e.g. fast generation of an e-mail, internal/public discussion forum)
- Possibility to attach notes to certain entries in the semantic network
- Base for discussions in the IFToMM permanent commission “Terminology” (e.g. IFToMM commission decides about new releases of the semantic network)

Documentation of the work process

- Recording of changes during the maintenance process of the semantic network (the protocol is part of the database),
- Generation of reports and statistics

An editor as well as a tool, which visualises the semantic network, is currently developed (Figure 3). These open source tools could serve as a basis for collaborative work.

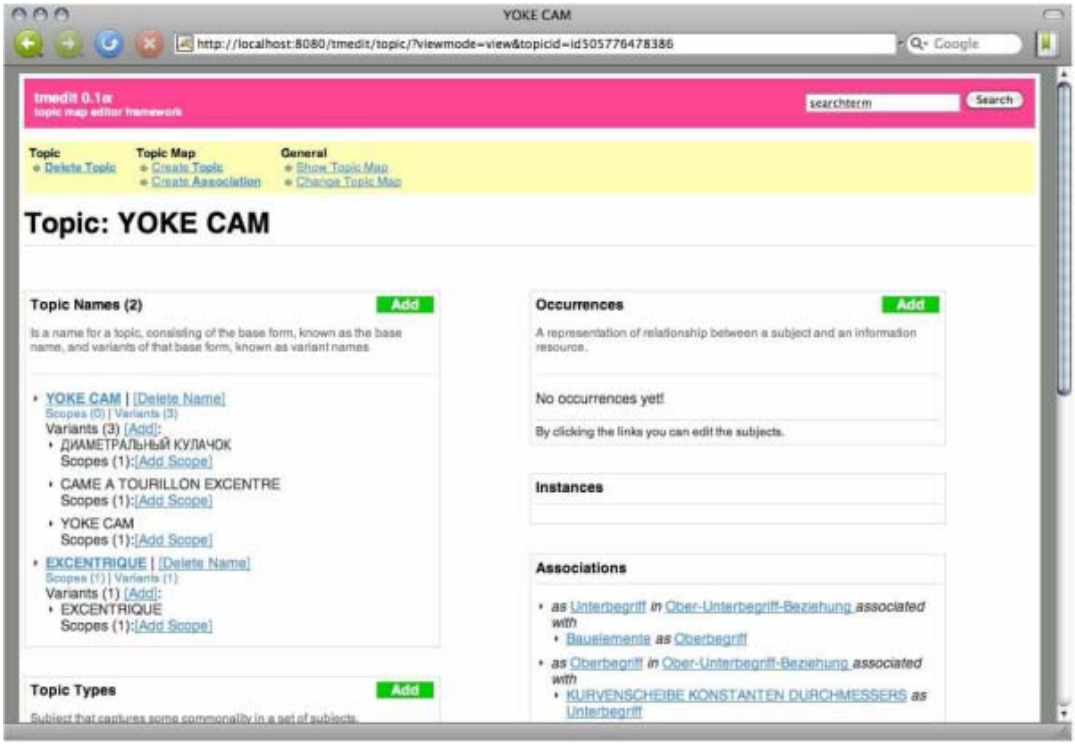

Fig. 3.: Prototypical implementation of the semantic network editor

7 CONCLUSIONS

The paper discusses the use of the IFToMM dictionaries as basis for a semantic network which can be utilised to improve the information retrieval process in digital libraries and information systems (e.g. DMG-Lib). The paper starts with an overview of known retrieval problems in this context. It is shown how a semantic network can be used to solve such problems. Furthermore a semantic network which is based on the IFToMM dictionaries as well as some use cases are presented.
The presentation of ideas for a collaborative development of an IFToMM dictionary based semantic network closes the paper.

8 ACKNOWLEDGEMENT

The authors would like to thank Hendrik Thomas and Tobias Redmann for their work on semantic networks in the DMG-Lib project. Furthermore, thank goes to Doreen Altinay for her support during the preparation of this paper. The work was partially financed by the DFG (Deutsche Forschungsgemeinschaft).

9 REFERENCES

[1] DMG-Lib - Digital Mechanism and Gear Library, http://www.dmg-lib.org
[2] KMODDL - Kinematic Models for Design Digital Library, http://kmoddl.library.cornell.edu
[3] Moscow State Technical University n.a. N.E. Bauman (MSTU), description of the local mechanism collection, http://tmm-umk.bmstu.ru/index_3.htm
[4] Topic Map standard ISO/IEC 13250:2003.
[5] MEYER zur CAPELLEN, W. (Ed.): Kinematik und Dynamik der Getriebe; terminologisches Wörterbuch; Deutsch, Englisch, Französisch, Russisch, Bulgarisch, in: Industrie-Anzeiger, 1969. http://www.dmg-lib.org/dmglib/handler?docum=2935009
[6] ISO Standard 701:1998, International gear notation - Symbols for geometrical data.
[7] ISO Standard 1122-1:1998, Vocabulary of gear terms -- Part 1: Definitions related to geometry.
[8] German standard DIN 3998, Denominations on Gears and Gear Pairs.
[9] Italian standard UNI 4760, Gearing - Glossary and geometrical definitions.
[10] IATE ("Inter-Active Terminology for Europe"), EU inter-institutional terminology database with over 20 supported languages. http://iate.europa.eu/iatediff/SearchByQueryEdit.do

[11] LEO, Online dictionary in different languages. http://dict.leo.org
[12] Static Word list of Schray GmbH concerning gear technology.
http://www.schray-antriebstechnik.de/deutsch-englisch.htm