

# mArachna – Ontology Engineering for Mathematical Natural Language Texts

Sabina Jeschke, Nicole Natho, Sebastian Rittau, Marc Wilke  
Technische Universität Berlin  
MuLF-Center Faculty of Mathematics & Natural Sciences  
10623 Berlin, Germany  
{sabina.jeschke, natho, rittau, marc}@math.tu-berlin.de

## Abstract

*The knowledge contained in the growing number of scientific digital publications, particularly over the internet creates new demands for intelligent retrieval mechanisms. One basic approach in support of such retrieval mechanisms is the generation of semantic annotation, based on ontologies describing both the field and the structure of the texts themselves. Many current approaches use statistical methods similar to the ones employed by Google to find correlations within the texts. This approach neglects the additional information provided in the upper ontology used by the author. mArachna, however, is based on natural language processing techniques, taking advantage of characteristic linguistic structures defined by the language used in mathematical texts. It stores the extracted knowledge in a knowledge base, creating a low-level ontology of mathematics and mapping this ontology onto the structure of the knowledge base. The following article gives an overview over the concepts and technical implementation of the mArachna prototype.*

## 1 Background

Information and knowledge are central concepts in today's society. Numerous publications, books and the World Wide Web create an "info glut" that is not easy to manage. Furthermore, manual information processing is a very time-consuming process. A reasonable approach to this problem is the development of knowledge bases integrating the knowledge from several different sources. However, this creates a new challenge: the stored knowledge has to be made accessible for users, requiring intelligent search and retrieval mechanisms. Many of the current ap-

proaches are based on the statistical analysis of the correlations between terms and knowledge elements. This approach, while very successful, neglects the advantages offered by the additional information of the ontology used by the authors and the ontology of the field of knowledge itself. Taking advantage of this additional information requires semantic annotation of the knowledge elements following the used ontologies. However, annotation of elements with metadata places an additional workload on the creator. Therefore, a mechanism is required to generate an ontology from text sources automatically. Going one step further, an extended ontology spanning several sources would support the user in gaining an overview over a wider field of knowledge.

## 2 Deployment Scenarios

The knowledge base and the corresponding ontology can benefit users on several different levels:

1. A user wants to find specific information on a mathematical concept such as a definition or an example of a given mathematical term. A query of the knowledge base will be able to return not only the desired information based on the upper ontologies of the sources, stored as annotations to knowledge granules, but also related terms and the context of the information based on both the upper and low-level ontologies of the database.
2. A user wants to gain an overview over a field of mathematics. The graphic representation of the knowledge base maps the structure of the knowledge base itself and therefore the ontology of the field presented.
3. A user required more detailed understanding of a particular mathematical concept new to

him. The retrieval mechanism will not only be able to provide him with information in direct reply to his initial query, but also with related knowledge helpful or even necessary to comprehending the topic.

### 3 Basic Concept

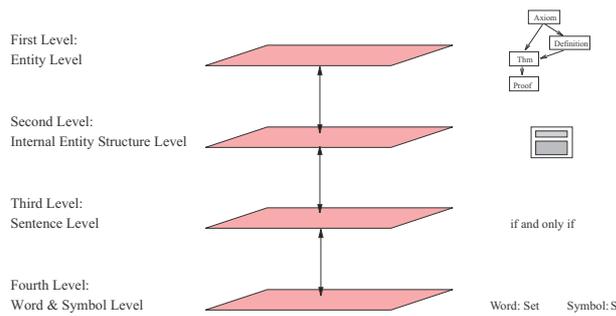
**mArachna** is a system for automatically extracting knowledge from mathematical natural language texts. **mArachna** creates a mathematical knowledge base representing elements of mathematical knowledge as well as the relations and interconnections between them. In the process, the system generates an ontology of the extracted mathematical knowledge, mapped onto the structure of the knowledge base. Mathematical texts show a distinctive structure, both on the linguistic level and in the presentation of knowledge chosen by the author. This structure is characterized by typical text elements, such as definitions, theorems and proofs. In the following, we will refer to these text elements as entities. Entities are commonly used to describe mathematical objects and concepts and represent the key-content of mathematical lectures. These entities form a complex network of relationships that defines an ontology. Since, following the ideas of Hilbert [?] and Bourbaki [?] mathematics as a whole can be derived from a small set of axioms using propositional logic<sup>1</sup> it can be said that mathematics possess an inherent structure, or, in other words, an inherent ontology. The network of mathematical terms and their relations as created by **mArachna** closely recreates that structure inherent to mathematics itself. As a result, **mArachna** is able to integrate mathematical entities from very different sources, such as different mathematical textbooks, articles or lectures, independent of the upper ontology preferred and used by the authors of those sources. This approach based on natural language processing offers several distinctive advantages over the more common approach based on purely statistical methods as used in e.g. Google. It utilizes both the ontology inherent in mathematics itself and the ontology defined by the structure of the text chosen by the author. **mArachna** provides mechanisms for the creation of retrieval networks from mathematical texts, such as textbooks, digital lectures or articles, as well as mechanisms for navigation on these networks. These networks reflect the contextual rela-

<sup>1</sup>This approach is valid within the context of **mArachna**, despite Gödel's Incompleteness Theorem [?], since we map existing and proven mathematical knowledge into a knowledge base and do not want to prove new theorems or check the consistency of the theorems we store.

tions between mathematical terms and concepts. They are further annotated regarding their role in the ontology of the original text, thus preserving the upper ontology used by the authors, and can map relationships in a very fine-grained way. In addition, these networks can be used to create an overview of mathematical content. Therefore, they offer different levels of information detail.

### 4 Related Work

**mArachna** consists of several separate subprojects. While we are unaware of any current research mirroring **mArachna**'s global concept, there are a number of projects with similar aims (if different approaches) to the subprojects within **mArachna**. Fundamental research into ontologies for mathematics was performed by T. Gruber and G. Olsen [?]. Baur analyzed the language used in mathematical texts in English [?]. The linguistic analysis of mathematical language is based on TRALE [?]. MBASE is an example of a mathematical ontology created by humans. The structuring of mathematical language and the modeling of the resulting ontology is similar to the work performed by humans in the creation of several mathematical encyclopedia like MathWorld [?]. The ontologies represented in these projects can be used as an interesting comparison to the ontologies and knowledge bases automatically generated by **mArachna**. WordNet/GermaNet [?] [?] [?] are intelligent thesauri providing sets of suitable synonyms and definitions to a given queried expression. WordNet uses a similar approach to the automated semantic analysis of natural language compared to **mArachna**. However, WordNet does not provide detailed correlation between dissimilar or not-directly related terms. Helbig [?] has designed a concept for automated semantic analysis of German natural language texts. The Mizar [?] [?] [?] system uses a both human and machine-readable formal language for the description of mathematical content as input for automated reasoning systems. An example for a different approach towards information extraction from mathematical natural language texts, based on automated reasoning, would be the work of the DIALOG [?] project. DIALOG is aimed at processing natural language user input for mathematical validation in an automated reasoning system. Both Mizar and DIALOG, in difference to **mArachna**, require complex reasoning systems for their extraction of information. The HELM [?] project defined metadata for semantic annotation of mathematical texts for use in digital libraries. However, this metadata has to be provided by the origi-

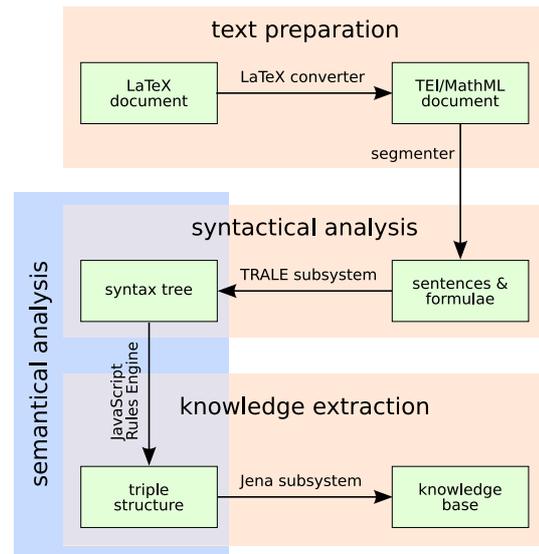


**Figure 1. Linguistic Classification Scheme**

nal authors. MoWGLi [?] extracts the metadata provided in HELM and provides a retrieval interface based on pattern-matching within mathematical formulae and logic expressions. The PIA [?] [?] [?] project extracts information from natural language texts (not limited to mathematics) using machine learning techniques. PIA requires an initial set of human expert annotated texts as a basis for the information extraction performed by intelligent agents.

## 5 Linguistic Approach

Entities are the principal carriers of information in mathematical texts. They are analyzed using natural language processing techniques, based on a linguistic classification scheme [?] [?]. This scheme defines four levels (see Fig. 1): relations between different types of entities are described on the entity level. On the internal entity structure level specifies the internal structure of an entity (i.e. the assumptions and proposition of a theorem). Characteristic sentence structures, which are commonly found in mathematical texts, are described on the sentence level. On the word and symbol level at the bottom single symbols and words and their relations between each other are schematized [?]. Mathematical information is extracted from a text using the structures and linguistic relations as defined by this classification scheme. The information is integrated into a knowledge base. This knowledge base consists of one or more directed graphs representing terms, concepts and their relations between each other. It is based on an ontology of the language of mathematics encoded with the web ontology language OWL [?]. The linguistic analysis of entities yields triples consisting of two nodes and one relation. Nodes represent mathematical terms and propositions, with the relation describing how they are connected to each other. It should be noted that triples themselves can be



**Figure 2. Processing of Natural Language Texts**

used as nodes in other triples, allowing the representation of more complex interrelations. In this context, different types of relations describe different types of linguistic phrases or key words in mathematical texts (e.g. two nodes corresponding to two propositions A and B, connected by the relation “is equivalent to”). These triples are then integrated into the knowledge base. This process closely maps the actual language structure, resulting in a very fine-grained knowledge base.

## 6 Technical Aspects of mArachna

mArachna consist of several separate modules (see Fig. 2: a preliminary analysis of the input, the syntactic analysis, the semantic analysis and the integration of the results into a knowledge base.

### 6.1 Design Decisions

The basic design of mArachna adheres to the following general guidelines:

- **Modularization:** Each component (input segmentation, syntactic analysis, semantic analysis, knowledge base, retrieval) have been designed as independent modules that can be replaced easily.
- **L<sup>A</sup>T<sub>E</sub>X as input format:** L<sup>A</sup>T<sub>E</sub>X was chosen since it is the standard format for publication in mathematics and the natural sciences.

- Standard internal data formats: All data is represented in standard data formats (TEI, MathML, XML, RDF/OWL).
- Control of the syntactic and semantic analysis: All rules governing the syntactic and semantic analysis should be modifiable by non-programmers to facilitate future extensions.

## 6.2 Preliminary Analysis

mArachna uses TEI (Text Encoding Initiative [?]) as its intermediate internal text storage format, with all mathematical formulae and symbols being encoded in Presentation MathML [?]. Currently, mArachna provides a LaTeX-to-TEI-converter. The TEI representation of the text is analyzed to extract the entities and their relations to each other, as well as the information provided on their internal structure level. Following this preliminary analysis, the system segments the entities into single sentences, annotating them with additional information concerning their parent entity and their role within the internal structure of that entity. Mathematical formulae are generally separated from the surrounding natural language text for further processing (simple formulae are replaced with corresponding natural language text). This preliminary analysis is implemented in Java using rule-based string comparison.

## 6.3 Syntactical Analysis

In the next step, each extracted natural language sentence is analyzed using computer-linguistic methods. The natural language analysis is implemented using the TRALE-System [?] based on a Head-Driven Phrase Structure Grammar. TRALE is a PROLOG adaptation of ALE for the German language [?]. TRALE, as used in mArachna, has been extended by expanding the underlying dictionary and grammar to include the specifics of mathematical language. Complex formulae have to be processed separately from the natural language analysis. This process is, however, not yet implemented. The natural language analysis performed by TRALE provides detailed syntactic and even some limited semantic information about each sentence. The output of the TRALE system is converted into an abstract syntax tree representing the structure of the analyzed sentence. At this point, the separately processed formulae will have to be reintegrated with the natural language text for further semantic analysis.

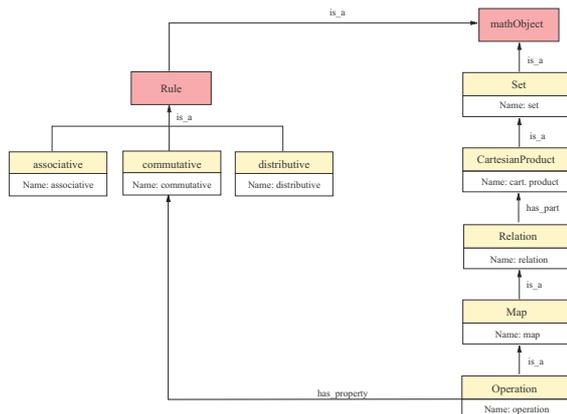


Figure 3. Basic Knowledge Base

## 6.4 Semantic Analysis

The semantic analysis is implemented in the form of an embedded JavaScript interpreter (Rhino TODO REF). The syntax trees are categorized according to typical structures characteristic for specific mathematical entities and semantic constructs. Each category is transformed into a specific triple structure defined by external JavaScript rules. These rules map typical mathematical language constructs onto the corresponding basic mathematical concepts (e.g. proposition, assumption, definition of a term etc). The resulting triples are annotated with additional information. This includes both the context within the original document as well as their classification within the context of the final OWL documents generated by mArachna. For each element of the triples it has to be decided if they represent OWL classes or individuals of OWL classes. This distinction poses a significant challenge for the automated analysis, currently forcing mArachna to use OWL Full.

## 6.5 Knowledge Base

The conversion of the generated, annotated triples into OWL documents for storage within the knowledge base is performed using the Java-based Jena framework [?]. Jena offers the advantage that it provides APIs for persistent storage and an integrated SPARQL (TODO REF) engine for RDF-based queries. The knowledge extracted from each analyzed text generates a separate, independent knowledge base. This separation is necessary to retain the idiosyncrasies and preferences of each author. Initially, these knowledge bases contain basic mathematical knowledge (see Fig 3) comprising axiomatic set theory and first order logic. To avoid inconsistencies within

the knowledge base, new information is added using a semi-automated approach: information is integrated into the knowledge base if and only if there are no conflicts with existing entries. This means that new nodes have to be connected to existing ones; duplications or contradictions are inadmissible. In case of conflicts, the user may provide additional information or delete the conflicting entries manually. This user-intervention is performed through Jena. This dual model of information management is based on well-known models of human knowledge processing: humans will be able to integrate new knowledge into their world view only if it can be linked to existing knowledge. Insufficient or incorrect prior knowledge may lead to misinterpretations of new information. As a consequence, incorrect knowledge may be deleted or corrected under certain conditions (see ch. 1). It should be noted that, given the nature of the sources, human intervention into the automated integration is not the rule; most of the knowledge presented in a textbook follows the rules of consistency required for the automated integration.

## 7 Linguistic Analysis: An Example

Take as an example (see Fig. 4) the definition of a group [?]

Let  $G$  be a set,  $G \neq \emptyset$ ,  $a, b \in G$

$$\begin{aligned} * : G \times G &\mapsto G \\ (a, b) &\mapsto a * b. \end{aligned}$$

$(G, *)$  is called a *Group* if the following conditions hold:

1. Associativity:  $\forall a, b$  and  $c \in G, a * (b * c) = (a * b) * c$
2. Neutral Element:  $\exists e \in G$  such that  $\forall a \in G, a * e = e * a = a$
3. Inverse Element:  $\exists a^{-1} \in G$  such that  $\forall a \in G, a * a^{-1} = a^{-1} * a = e$  where  $e$  is the neutral element.

Based on the knowledge of the use of certain keywords in mathematical texts, the linguistic analysis has identified “Let  $M$  ? a map” as the assumption, “The tuple ? if the following holds:” as the proposition and the following three points as the properties of a definition. The morphological, syntactical and semantic analysis results in the representation as triples of predicate, relation and object. For example, the phrase “Let  $M$  be a set” is mapped to the triple (is\_a, set,  $M$ ). Finally, identifying identical predicates and objects of different triples with each other generates the network of relations. The relations between entities

are described using OWL. The representation in OWL provides an overview of the mathematical knowledge that is generated by extracting relevant information from the knowledge base, with special attention given to the underlying field ontology.

## 8 Conclusions and Evaluation of the Basic Concept

At the moment mArachna exists as a prototypical implementation to analyze text written in German (it will be extended to analyze texts written in English). The prototype is capable of analyzing short passages of text, such as single mathematical entities as described in this paper. The semantic extraction leads to information snippets of the analyzed mathematical text. This information is successfully integrated into the discussed knowledge base. As such, the prototype serves as a proof-of-principle, validating the approach of mArachna. The system is capable of automatically generating a low-level ontology of mathematics from natural language texts based solely on natural language processing techniques. Future evaluation will include analyzing complete textbooks on Linear Algebra and merging the resulting knowledge bases to test the validity of the ontology mapping approach. The results will be compared with a ”standard” mathematical encyclopedia, particularly concerning the generated (or, in the case of the standard encyclopedia, used) ontologies for this field of mathematics.

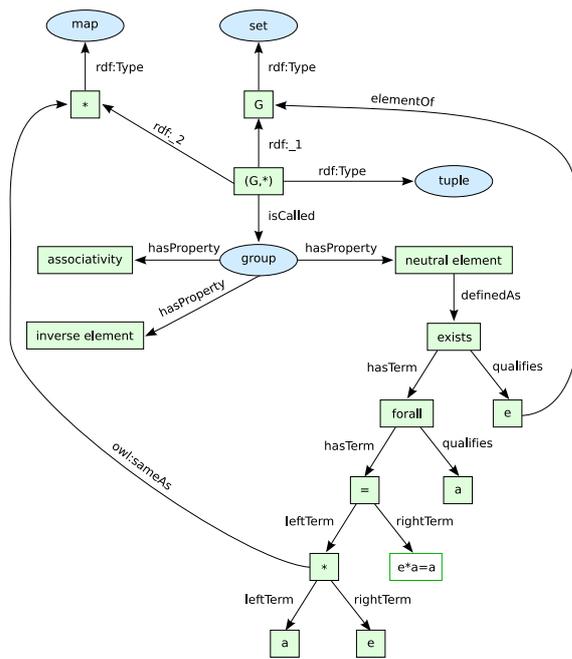
## 9 Future Work: Ontology Engineering in mArachna

### 9.1 Summary

Mathematical language, in particular within the entities, places a strict emphasis on the transfer of knowledge. Thus, the knowledge base consists of mathematical knowledge without any need of interpretation. The acquisition process (semantic extraction) follows strict rules based on the structure of entities. The same is true for the organization and storage of mathematical content in the knowledge base. But still, some problems remain with the knowledge base with regard to merging or extracting knowledge.

### 9.2 Knowledge Management and Ontology Engineering

The notations and phrasing used can vary significantly between different authors, based on



**Figure 4. Knowledge Base: Definition of a Group**

personal preferences, didactical goals etc. As a result, each text is stored in its own corresponding knowledge base. Future developments of mArachna will have to implement a sensible ontology mapping to create one, unified knowledge base from the separate smaller bases associated with one text each (see Fig. 4). In reverse, strategies and tools for the modularization of this unified knowledge base into smaller bases, specifically adapted to the needs and preferences of specific target user groups, have to be conceived and implemented. These “modularized” knowledge bases in turn could serve as the foundation for an intelligent retrieval system based on the idea of PIAs (personal information agents). For example, in education it is advisable to use different versions of an entity for elementary school and high school. Modularized knowledge bases in different notations would support the harmonization of the teaching material within a series of courses.

### 9.3 Retrieval

The aim of the future design of a retrieval interface will be the support of users in learning and understanding mathematics. The interface should provide different tools for selecting information based on personal preferences (more axiomatic oriented, more example oriented). In addition it would be desirable to integrate the administration of different roles into a generic in-

terface (students, teacher and administrator).

### 9.4 Processing of Mathematical Formulae

As formulae form a major portion of mathematical texts and constitute a primary source of information in these texts, it is desirable to be able to include their content in the analysis and representation created by mArachna. Currently, mArachna is not capable of this important feature yet. However, we are investigating an approach to rectify this deficiency. We propose using a syntactical analysis similar to those used in computer algebra systems in combination with contextual grammars (e.g. Montague grammars) to correlate the information given in a formula with information already provided in the surrounding natural language text. Using this approach should enable mArachna to integrate formulae and their informational content in the network created by the analysis of the natural language text. It should be pointed out that we do not aim for machine-based understanding of the formulae, as automatic reasoning systems would require. Instead, formulae are to be treated as a different representation of mathematical knowledge, to be integrated into the knowledge base in a similar manner to that used for the natural language text. However, the analysis proposed here can be used as a first step in a further process leading to viable input for such reasoning systems, providing additional assistance in building the knowledge base.

### References

- [1] Hilbert, D.: Die Grundlagen der Mathematik. Abhandlungen aus dem mathematischen Seminar der Hamburgischen Universität 6 (1928) 65–85
- [2] Bourbaki, N.: Die Architektur der Mathematik. Mathematiker über die Mathematik. Springer, Berlin, Heidelberg, New York (1974)
- [3] Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandte Systeme I., Monatsheft f. Mathematik und Physik (1931-1932) 147f
- [4] Gruber, T., Olsen, G.: An Ontology for Engineering Mathematics. Technical Report KSL-94-18, Stanford University (1994)
- [5] Baur, J.: Syntax und Semantik mathematischer Texte. Master’s thesis, Universität at des Saarlandes, Fachbereich Computerlinguistik (November 1999)

- [6] S.Müller: TRALE. <http://www.cl.uni-bremen.de/Software/Trale/index.html>
- [7] Wolfram Research: MathWorld. <http://mathworld.wolfram.com>
- [8] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, London (1998)
- [9] Fellbaum: WordNet. <http://wordnet.princeton.edu>
- [10] GermaNet Team: GermaNet. <http://www.sfs.uni-tuebingen.de/lsd/english.html>
- [11] Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin, Heidelberg, New York (2006)
- [12] Urban, J.: MoMM – Fast Interreduction and Retrieval in Large Libraries of Formalized Mathematics. *International Journal on Artificial Intelligence Tools* (15(1)) (2006) 109–130
- [13] Urban, J.: MizarMode – An Integrated Proof Assistance Tool for the Mizar Way of Formalizing Mathematics. *Journal of Applied Logic* (2005)
- [14] Urban, J.: XML-izing Mizar: Making Semantic Processing and Presentation of MML Easy, MKM2005 (2002)
- [15] M. Pinkall and J. Siekmann and C. Benzmüller and I. Kruijff-Korbayova: DIALOG. <http://www.ags.uni-sb.de/~dialog/>
- [16] A. Asperti and I. Padovani and C. Sacerdoti Coen and I. Schena: HELM and the Semantic Web. In Boulton, R.J., Jackson, P.B., eds.: *Theorem Proving in Higher Order Logics*, 14th International Conference, TPHOLs 2001, Edinburgh, Scotland, UK, September 3-6, 2001, Proceedings. Volume 2152 of *Lecture Notes in Computer Science.*, Springer (2001)
- [17] Asperti, A., Zacchiroli, S.: Searching Mathematics on the Web: State of the Art and Future Developments. In: *Joint Proceedings of the ECM4 Satellite Conference on Electronic Publishing at KTH Stockholm, AMS - SM M Special Session, Houston /* (2004)
- [18] Albayrak, S., Wollny, S., Varone, N., Lommatzsch, A., Milosevic, D.: Agent Technology for Personalized Information Filtering: The PIA-System. *ACM Symposium on Applied Computing* (2005)
- [19] Collier, N., K, T.: PIA-Core: Semantic Annotation through Example-based Learning, *Third International Conference on Language Resources and Evaluation* (May 2002) 1611–1614
- [20] The PIA Project: PIA. <http://www.pia-services.de/>
- [21] Jeschke, S.: *Mathematik in Virtuellen Wissensräumen - InK-Strukturen und IT-Technologien in Lehre und Forschung*. PhD thesis, Technische Universität Berlin (2004)
- [22] Natho, N.: *mArachna: Eine semantische Analyse der mathematischen Sprache für ein computergestütztes Information Retrieval*. PhD thesis, Technische Universität Berlin (2005)
- [23] Grottko, S., Jeschke, S., Natho, N., Seiler, R.: *mArachna: A Classification Scheme for Semantic Retrieval in eLearning Environments in Mathematics*. *Proceedings of the 3rd International Conference on Multimedia and ICTs in Education*, June 7-10, 2005, Caceres/Spain (2005)
- [24] W3C: Web Ontology Language. <http://www.w3c.org/2004/OWL>
- [25] The TEI Consortium: Text Encoding Initiative. <http://www.tei-c.org>
- [26] W3C: MATHML. <http://www.w3.org/Math>
- [27] Müller, S.: *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*. In: *Linguistische Arbeiten*, No. 394. Max Niemeyer Verlag, Tübingen (2005)
- [28] Jena: A Semantic Web Framework for Java. <http://jena.sourceforge.net>
- [29] Wüst: *Mathematik für Physiker und Mathematiker*, Bd.1. Wiley-VCH (2005)

