The SMGloM project or why we need a semantic glossary of mathematics

Michael Kohlhase and Wolfram Sperber

Abstract In this article, we describe a new terminological and notational base for mathematics: The Semantic Multilingual Glossary of Mathematics (in the following SMGloM) is an ontology for mathematical concepts, objects or models. The terminological and notational data can be applied, e.g., in a more semantic text and formula search and the disambiguation of symbols and formulae in mathematical publications or the translation of mathematical terms. The paper is focused to present the intention, the needs, the framework, and user scenarios of the SMGloM concept, but not to the technical details of the data model and its implementation.

1 Introduction and Motivation

Mathematics works with abstract Concepts, Objects or Models (COMs) analyzing their structures and the possible mappings COMs. Mathematics is known for its strength based on a well-defined vocabulary, and any mathematical endeavor requires a clear definition of COMs or in other words a well-defined "terminological and notational base", which is special in that COMs are given

Michael Kohlhase
Jacobs-University Bremen, Computer Science & Electrical Engineering, Campus Ring 1 28759 Bremen Germany,
✉ m.kohlhase@jacobs-university.de

Wolfram Sperber
FIZ Karlsruhe Leibniz Institut für Informationsinfrastruktur, zbMATH, Franklinstr. 11, 10587 Berlin
✉ wolfram@zentralblatt-math.org

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. -, No. -, -
ISSN 2363-9881
CC BY-SA

by abstract definitions, not by objects of the perceived world. The mathematical language is a natural language of its own but it is different from common languages in some features: The mathematical vocabulary consists of (compound) words and notations (symbols[1] and formulae).

Mathematical notations mark a degree of formalization of mathematics, they allow the presentation of mathematical statements and theories in a condensed and highly compressed way. Notations of a COM can be single characters or glyphs but can also be combinations of some characters or notations, e.g., the general group of degree n over a field F: $GL(n,F)$. Notations define a special alphabet for mathematics. Simple notations can be concatenated to form complex notations ("formulae") which allow for a precise identification of COMs. The notations have a well-defined semantic meaning and properties. Notations may take priority over a verbal name, e.g., special partial differential equations later were given the name of mathematicians, for example Schrödinger equation.

Mathematical language is the base for scholarly communication and discussion in mathematics. One of the first layout languages for mathematics was TEX, developed in the late 70s. TEX was designed as a layout language which allows mathematicians and other scientists to write their manuscripts in their own style without the help of publishing experts.

A terminological base is of central importance for the mathematical community. In the digital age, a lot of activities are developing Web services for the mathematical terminology, e.g., Wikipedia, (Online) Encyclopedia of Mathematics, Planet Math, etc. These services provide a considerable part of mathematical terminology and are a relevant source for humans for referencing. But, the existing services do not exhaust the full potential of Semantic Web methods and technologies, especially in mathematics. In particular, they don't allow automatic processing and reasoning.

OpenMath [**BusCapCar:2oms04**] was the first universal semantic markup language for mathematics. MathML [**CarlisleEd:MathML3**] is another XML dialect for mathematics which can be used for both the markup of layout (Presentation MathML) and semantics (Content MathML) which is is fully compatible (convertible) to OpenMath. Finally OMDoc [**Kohlhase:OMDoc1.2**] extends these with elements of the mathematical meta-structure such as theorems, axioms, proofs, etc. Thus OMDoc supports the semantification (flexible formalization: "flexiformalization") of mathematical publications which can

[1] In the following, we use "notation" instead of "symbols and formulae" to avoid incompatibilities with the meaning of "symbol" in the SMGloM data model which is described in detail in [**Kohlhase:dmesmgm14**].

be used for a machine-based content analysis as it represents a partial formalization. It would allow to extend the access and the retrieval of mathematical literature from the document level of today to knowledge level in the future .

But MathML and OMDoc are not well-suited for human input: too elongated, formal, and expensive. To fill the gap between input and semantic Web languages, a semantic extension of TEX named STEX [**sTeX:github:on**] was developed in the last years. STEX also allows for a semantic markup (by annotation) of mathematical notations and a conversion of LATEX to MathML (Presentation and Content MathML) and OMDoc. For more detail, we refer to the Web sites of STEX, MathML, and OMDoc. The most advanced tool for converting LATEX to MathML is LaTeXML [**Miller:latexml**].

Powerful mathematical encyclopediae are not only important for the communication but also for a lot of use cases. In the last years, FIZ Karlsruhe (FIZ) was involved in two projects, the [**delivermath2015**] and the [**mathsearch2015**] project, which also gave impulses for the development of a semantified new glossary of mathematics.

The DeLiVerMATH project aimed at the development of concepts and tools for the automatic content analysis of mathematical publications, especially keyword extraction and classification. For keyword extraction, natural language processing methods are used. In a first step, the types of the words (more generally tokens) are determined by the well-known "Part of Speech Tagging". Then, noun phrases which are the most relevant ones for the characterization of the content are identified. In a third step, the relevant noun phrases have to be selected. This can be done by different methods, e.g., neural networks. A semantic glossary base would be very helpful here.

The joint MathSearch project of FIZ and Jacobs-University Bremen (JUB), developed and realized a formula search for the zbMATH database. The zbMATH data are TEX-encoded and therefore contain only marginal semantic information. As a consequence, the meaning of the notations is ambiguous. A glossary which contains also semantified notifications would be very important for a disambiguation of notations and for improving the retrieval facilities.

In summary, mathematical encyclopediae will be a substantial part of mathematical knowledge. The Web and the development of Semantic Web technologies in mathematics allow for new advanced versions of mathematics encyclopediae which are useful for a lot of applications.

2 SMGloM: The terminology and the data model

The terms "controlled vocabularies", "thesauri", "glossaries", "terminological bases", "ontologies", "symbols", "notations", etc. are used differently in various communities and address different aspects. For a semantic mathematical glossary the terminology must be unified and clarified.

2.1 SMGloM: The terminology and design principles of SMGloM

In the pre-digital era, *mathematical glossaries* (or *encyclopediae*) consisted of alphabetically ordered lists of mathematical terms which were given by short articles containing the definitions of the COMs, and also references to other terms which are relevant for a term. Two prominent examples are [**naas2015**] and [**hazewinkel2015**]. They are useful reference bases of mathematical COMs.

Terminological bases are collections of terms. Mathematical terms are words or compound words of COMs which have a well-defined meaning in a certain context, their meaning may deviate from the meaning of the same words or phrases in other contexts and in everyday language. COMs can be represented by terms and/or notations, e.g., "Laplace operator" can be substituted with "Δ". COMs can have multiple terms or notations, e.g., "differential quotient" is synonym for "derivative", the same is valid for notations, e.g., "$f'(x)$" is synonym for "$\frac{df}{dx}$". The same terms and notations can describe different mathematical COMs, e.g., Δ can be the notation for a difference or the Laplace operator, as a parameter in different mathematical notations, etc.

Words or phrases can have various inflectional variants which linguistics calls "lexemes". Lexemes are usually referenced by their "lemma" (or citation form). Also notations can have variants. Parameters are one degree of freedom of notations. The design of SMGloM allows parameter versions as well as parameter-free ones.

Often, COMs can be defined in more than one way, e.g., "Euclidean geometry" can be described by different axioms. Equivalence of definitions must be proved which can be hard work. New research results can lead to new definitions. They are a first kind of semantic relations. Terminological relations are a general type of semantic relations. [**WordNet:on**] has marked some prominent semantic relations between terms: synonymy, hypernymy and hyponymy, meronomy and holonymy, homonymy and antonomy. We also address another

type of relations between COMs in SMGloM: dependency relations consisting of the relations to other COMS which are necessary to define a COM.

Semantic terminological bases usually make some relations between terms explicit, so that they can be used for automatic reasoning and customized presentations. In practise, the terminology and relations between terms, COMs, notations are much more complex. The paper of [**Kohlhase:dmesmgm14**] is – to the best of our knowledge - the first systematic discussion about the subject. SMGloM is not only a terminological terminal base but also a notational base for mathematics with a broad range of semantic relations.

2.2 SMGloM: The data model

SMGloM is designed as a building block for mathematical knowledge management. SMGloM should be a useful service for both the mathematical community but also for the automatic analysis and processing of information.

As was said above, COMs can often be defined in various ways. Each definition of a COM will be assigned to a SMGloM item. Different equivalent definitions of a COM are characterized by the semantic relation for equivalence. SMGloM consists of entries which are named "modules". Figure 1 gives an overview of the data model

- **Modules**: Modules are the atoms of SMGloM, they consist of the module signature and the language bindings. Modules and COMs are closely related, but it's only a 1:1 relation in the simplest case: Modules can cover groups of aggregated COMs (groups of COMs which are defined simultaneously). On the other hand, COMs can be defined in more than one way and thus in more than one module.
- **Module signatures**: A module has exactly one signature which contains the module identifier, the semantic relations to other modules and the notations of a COM (which are typically language-independent). All known variants of the notations have to be declared in the module signature. This covers different notations for (ordinary or partial) derivations of a function, the notations can contain parameters which can be written in different form, etc.
- **Language bindings**: The concept of language bindings allows for a multilingual approach of SMGloM. A module can have any number of language bindings (one for each language), these contain the definitions of the terms

of a COM in that language. As English is the lingua franca in mathematics, this definition can be used as base and origin for adding further language bindings in other languages.

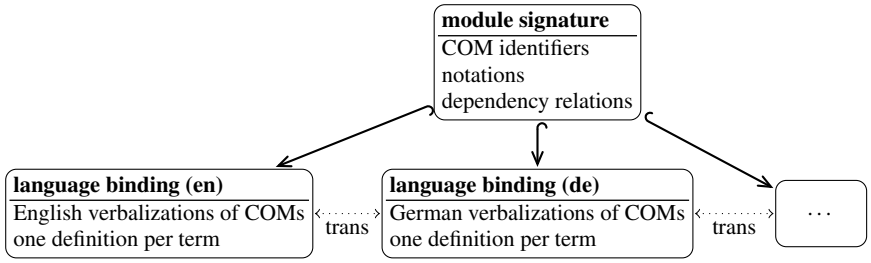

Fig. 1 The SMGloM Data Model

SMGloM models the semantic relations as relations between modules:

- **Dependency relations**: The modules of the COMs which are necessary for creating a new module must be declared in the module signature. The information can be used to create a dependency graph between modules (and COMs).
- **Terminological relations**: First terminological relations are used in the language bindings. The definitions in a language binding should list all possible terms of a COM: These sets are the terminological synonym sets (synsets) of Wordnet.

The SMGloM data model inherits linguistic, library-oriented, and mathematical aspects. It is different from other terminological bases, e.g., Wikipedia, which have a much simpler structure: The SMGloM structure provides a 'microscopic' insight into the mathematical terminology: Special definitions can be addressed unambiguously. This is relevant, e.g., for proving theorems but also for embedding of different definitions in its mathematical context, and for maintaining the language bindings in several languages. Also, other definitions of a COM, can be simply added by introducing a new module in SMGloM. For the first time, notations are in a unique relation with a COM which is relevant for a machine-based content analysis of mathematical publications. Last but not least essential advantages result from the semantic relations which are modeled in the modules and between the modules.

Semantic relations between modules are separate elements of the SMGloM concept wich are called "views". A first example for views are equivalence

relations between different definitions of a COM. These relations will allow - analogously to synsets for terms - to provide a view to the "synset" of all equivalent definitions of a COM. The dependency relations and views embed the terms which are defined in a module in their mathematical context in a human and machine-understandable way.

The development of the data model is not finished. Currently, the refined specification and typing of views, which were introduced in the MMT concept, see [**Kohlhase:dmesmgm14**], is a hotspot in the discussion of the SMGloM concept.

3 Implementation & State of the System

Some problems must be solved for the realization of SMGloM besides the further development of the architecture: the technical implementation and the organizational realization of the glossary. A semantic mathematical glossary requires semantic languages for the input. sTEX has the potential for such a semantified encoding. A distributed system for input and quality control seems to be a convenient organizational frame for SMGloM.

3.1 Technical implementation

sTEX is used for encoding SMGloM. The use of TEX has a long tradition in the mathematical community, it is the de facto standard for mathematical publishing. sTEX additionally allows semantic markup of the mathematical content (terms, their definitions, and notations), module structure, and relations without forgoing high-quality presentation. Up to now, tools for an automatic assistance of the input are missing, but are high-priority development goals.

The SMGloM approach was evaluated in spring 2015. The result is a prototype for SMGloM. Experiences of the test phase were used for tuning and adapting the data model and encoding.

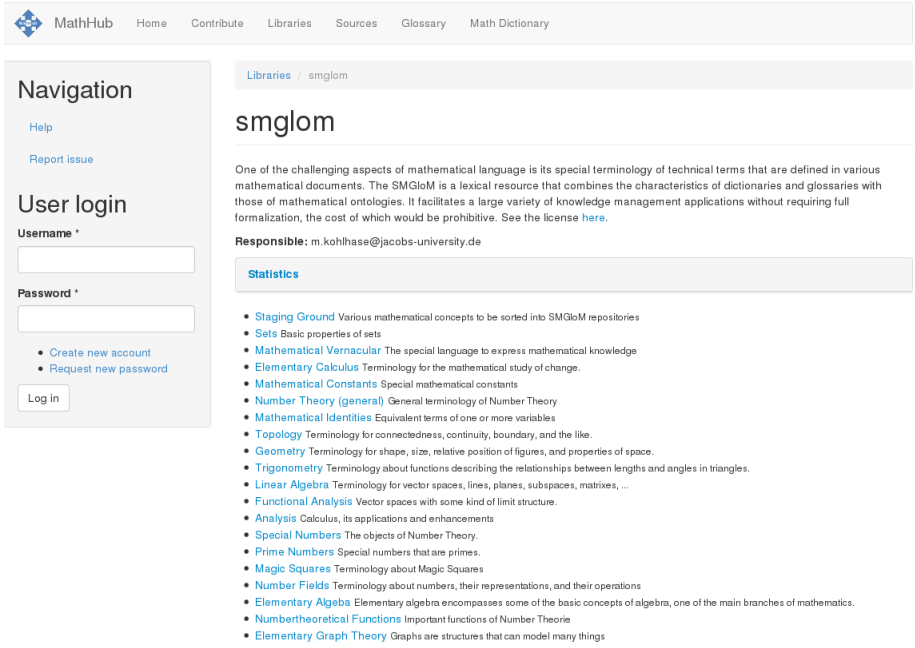

Fig. 2 The homepage of SMGloM

3.2 State of the art

Within the MathSearch project, the concept of SMGloM has been checked and a first prototype has been installed at [**SMGloM:on**]. The further development of SMGloM is a dynamic process. The feedback of the input has led to some modifications of the original SMGloM data model and the implementation rules. Also, first tools have been developed which support a technically correct input, control, and maintenance of SMGloM, e.g., for adaption or correction of existing entries and their semantic relations. New or modified entries can lead to redundant relations which must be handled by the system.

In the meantime, the prototype of SMGloM contains nearly 500 items and more than 1,500 language bindings (English, German, Romanian, Turkish, Chinese) which have been created manually by FIZ and JUB. The prototype provides, analogously to other encyclopedias, an alphabetically sorted list of the glossary terms, the definitions and the so-called concept graphs which are a

visualization of the dependency relations of a certain term with other SMGloM terms. Also a translator for the SMGloM terms has been realized in the prototype.

The WebSite of SMGloM will be introduced by some screenshoots in the following. Figure 2 shows the homepage of SMGloM which gives an overview about the mathematical subjects currently realized in SMGloM. Note that this is organized by "Math Archives" which is chosen mainly for management (rights management) reasons. Figure 3 shows the (English) term list of the glossary generated from the content. For each term, a definition (see Figure 4) and a dependency graph (see Figure 5) can be accessed. Alternative languages can be accessed by the links on the right.

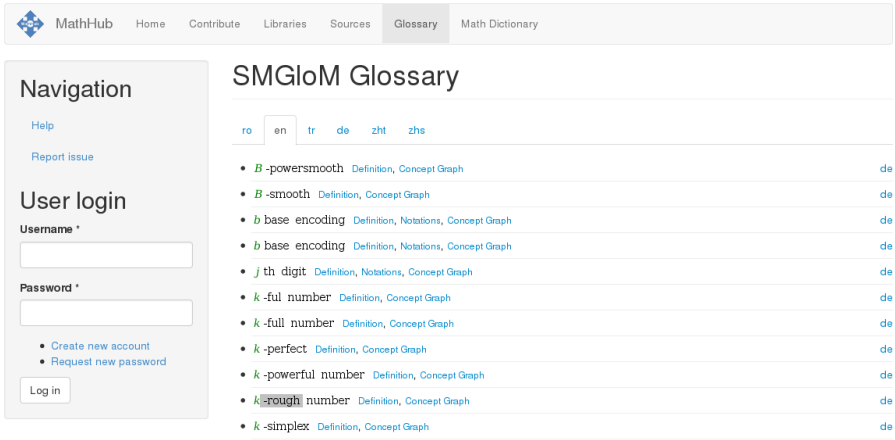

Fig. 3 The term list of SMGloM (snippet)

Fig. 4 The English language binding for the term "prime number"

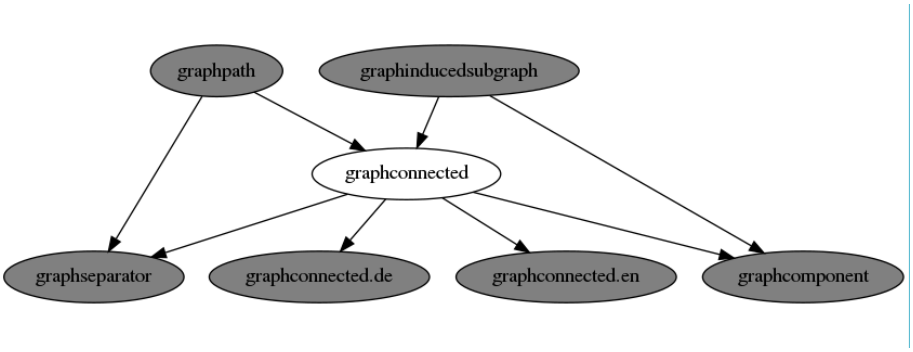

Fig. 5 The dependency graph for the term "Connected graph"

The dependency relations in the module signature allow for creating graphs which contain the terms which were used in the definition of a certain term and the terms which use this term for their definition. The overall graph is very complex (too many nodes and semantic relations). But local graphs which present a node and its environment could be very helpful for the user. Currently, the graphs contain only dependency relations of first order (parents and childs) but can be extended by views.

SMGloM has achieved a first stage of development, but not the maturity and robustness which is necessary for a widely distributed production system and service. Also, the input of the prototype requires some familiarity with the SMGloM concept and experiences with TEX.

As a next step, the existing SMGloM prototype should be developed to a working system and useful service. Therefore, a project proposal is under preparation. A Todo list and some concepts for the organization, input and control will be discussed below.

3.3 Use cases

In the following, we list some possible use cases which are partially implemented in the prototype.

- **Visualization**: A first application of the SMGloM model is the graphical presentation of the dependency relations. The modules are the vertices of the graph, the dependency relations are the edges. This is already realized

in the prototype. The graph embeds a COM in its mathematical context by visualizing the dependency relation between COMs.
- **Dictionary**: SMGloM contains a mathematical dictionary. If an SMGloM entry covers more than one language binding in a module, SMGloM can be used immediately for translation. This has been also realized in the prototype.
- **Controlled vocabulary for mathematics**: Authors of mathematical information can refer SMGloM terms and notations which support a standardized use of the mathematical vocabulary. The practical use of this feature requires a sufficient coverage of the mathematical vocabulary.
- **Semantified publishing**: Especially, SMGloM defines identifiers for the notations and terms of COMs. Authors can refer to the semantically annotated notations and terms in their publications. This would simplify the reading but also the machine-based content analysis of publications.
- **Disambiguation**: SMGloM can identify various meanings of notations or terms. This can be used for the disambiguation of terms and notations, intellectually or by automatic means.
- **Retrieval**: Terms and notations of COMs are connected by the data model. This allows a common search for terms and notations.
- **Customizing**: SMGloM allows to customize the presentation style (using of preferred terms and notations) in mathematical publications. Publications can be transformed to the preferred style of the user. This could be a useful application, e.g., for reading of historic documents.

4 SMGloM – a Community-Based Approach and Quality Control

Wikipedia has impressively demonstrated the merits and strengths of a community-based distributed enterprise for the development of terminological bases in the Web. Wikipedia has shown what is needed to be successful: a smart technical system for an intuitive and distributed input and a rigorous quality control policy. Also the free use of Wikipedia is essential for the success of Wikipedia.

SMGloM must alos be a community-based distributed activity:

- **Input and quality control**: Only the mathematical community has the expertise to create such a glossary and ensure a high level of quality.
- **Usability**: The mathematical community is strongly interested in such a service because it is a useful service for the daily work.

- **Use**: SMGloM is provided by the community for the community as a free service.

SMGloM needs broad support from the mathematical community for

- **Input**: A substantial coverage of the mathematical vocabulary is essential for its acceptance. Each mathematician can be a contributor for SMGloM by creating input corresponding to the data model.
- **Quality control**: Only mathematicians who are experts on a mathematical subject can control the SMGloM input and ensure its quality.
- **Conceptual development and technical infrastructure**: This contains the advancement of the SMGloM concepts and their implementation, the development of an editorial system for the workflow and quality control in SMGloM and interfaces for applications.
- **Steering**: Publishing, quality control, and the technical and conceptual development need coordination by a steering committee of experts.

This means, various groups of the mathematical community will be involved in the SMGloM initiative. For the conceptual development and technical infrastructure a special group of mathematicians with expertise in Semantic Web technologies is indispensable. SMGloM needs a stable personnel base to guarantee the quality and sustainability of SMGloM. All activities should be steered by an "'Executive Board" (EB). The EB coordinates the work of authors, editors, and developers und is responsible for the communication with the mathematical community (especially with the International Mathematical Union and other learned societies). The EB develops a business plan for the SMGloM initiative. The development of this infrastructure is one of the most important steps for the next time.

5 Conclusion

SMGloM is a concept and a new type of a mathematical glossary. Currently, SMGloM in the stage of a prototype. But we hope, that the concept and its potential for applications are attractive for the community and will be developed as a useful service in the near future.

SMGloM can be an important step to a sophisticated mathematical knowledge management on the Web and a starting point for realizing a Semantic Web for mathematics.

Even though the system is operational at a basic level, thre are still many conceptual and technical challenges to further develop an active community of contributors.

- **Input**: Guideline have to be developed showing what to do and how to do it. Moreover, the SMGloM input system should provide as much as possible technical support for authors, e.g. by
  - a graphical user interface conducting the user through the input process
  - visualization tools especially for the semantic relations
  - tools for automatic proving the syntactic correctness of the input
  - filters for detecting redundant or controverse input in SMGloM, e.g., avoiding of non-unique identifiers
  - tools for processing and maintaining of SMGloM data (e.g., the dependency graph)
- **Quality control**: SMGloM input and control needs both the content and formal correctness.
  - *correctness of the content*: The origin of a definition is a strong indicator for the correctness of a definition. Therefore, the source of the definition – typically a publication – should be declared. Also, the semantic relations allowing to embed a term in its mathematical context must be correct and nearly complete.
  - *Formal correctness* (modules, views, encoding of notations, links, terms, etc.): Without a correct encoding, SMGloM could not provide the enhanced features of SMGloM, especially automatic processing and reasoning of information. Hence, a strict control of formal correctness is necessary.

  In particular, SMGloM needs a special editorial system. The existing editorial systems for quality control of mathematical publications are focused on the mathematical content. The control of formal correctness requires additional methods, e.g., the use of sophisticated concepts of software engineering. The editorial system should assist authors and editors to create and publish high-quality SMGlom entries.
- **Conceptual development and technical infrastructure**: Besides the tasks listed above (input support, editorial system) the following tasks are addressed

- *Maintenance and Web presentation of the SMGloM service*: This covers the presentation of SMGloM, interfaces and application, embedding of SMGloM in the mathematical knowledge management activities, but also the day-to-day business.
- *Machine-based input generation*: In principle, input generation can be done also by machines. Concepts and methods for a machine-based input generation should be developed and discussed. This would help to increase the number of SMGloM items rapidly.