# The SMGloM Project and System Towards a Terminology and Ontology for Mathematics

Deyan Ginev[1], Mihnea Iancu[1], Constantin Jucovshi[1], Andrea Kohlhase[1], Michael Kohlhase[1], Akbar Oripov[1], Jürgen Schefter[2], Wolfram Sperber[2], Olaf Teschke[2], and Tom Wiesing[2]

[1] Computer Science, Jacobs University Bremen; `http://kwarc.info`
[2] Zentralblatt Math, Berlin; `http://zbmath.org`

**Abstract.** Mathematical vernacular – the everyday language we use to communicate about mathematics is characterized by a special vocabulary. If we want to support humans with mathematical documents, we need to extract their semantics and for that we need a resource that captures the terminological, linguistic, and ontological aspects of the mathematical vocabulary. In the SMGloM project and system, we aim to do just this. We present the glossary system prototype, the content organization, and the envisioned community aspects.

## 1 Introduction

One of the challenging aspects of mathematical language is its special terminology of technical terms that are defined in various mathematical documents. To alleviate this, mathematicians use special glossaries, traditionally lists of terms in a particular domain of knowledge with the definitions for those terms. Originally, glossaries appeared as alphabetical lists of new/introduced terms with short definitions in the back of books to help readers understand the contents. Another kind of resource that deals with terminology of mathematics are "dictionaries", which align mathematical terms in different languages by their meaning – originally without giving a definition.

In the last decades the term "glossary" has also been applied to digital vocabularies (online encyclopedias, thesauri, dictionaries, etc.), which have become important resources in knowledge-based systems. This is especially true for vocabularies that have a *i*) semantic aspect – i.e. some of the relations between the terms and the concepts, objects, and models they denote are made explicit and machine-actionable, they are also called "ontologies" – or *ii*) that are multilingual. Digital vocabularies can be hand-curated, or machine-generated/collected; an example of the former is the WordNet lexical database for English, [WN] an example of the latter is DBPedia, [DBP13] but they can also be hybrid, e.g. the UWN/Menta project [YAGO] generates a multilingual WordNet by automatically adding other languages by crawling Wikipedia.

We present the SMGloM project, which aims to create a semantic, multilingual glossary for mathematics. This resource combines the characteristics of dictionaries and glossaries, with those of ontologies, but restricts the content to definitions and the relations to the lexical ones to keep the task manageable. Here we give a high-level overview over the data model, the SMGloM system, organizational and legal issues, possible applications, and the state of the effort of seeding the glossary.

## 2 The SMGloM System

**Data Model and Encoding** We build the data model of SMGloM on top of the one of OMDoc/Mmt [Koh06; RK13], which provides views, statements, and theories. In a nutshell – see [Koh14] for details, a **glossary entry** consists of one **symbol**, its **definition**, and a set of **verbalizations** and **notations**. A symbol is a formal identifier of a mathematical object/concept (i.e a formal object). The verbalizations relate it to lexical entries (identified by the stem of the head), which we call **glossary terms**. The definitions could be written down in a formal logic, but in the SMGloM, we write them down in mathematical vernacular (common mathematical language) and encode them in sTeX, a variant of LaTeXwith semantic annotations [sTeX]). Figure 1 shows the concrete form of a glossary entry. It consists of a **module signature**, which specifies the formal part: here the module name graphconnected in **mhmodsig** environment and the symbol connected via the **symi** macro. The module signature also specifies the conceptual dependency on paths in graphs by importing the grpahpath module via **gimport**.

```
\begin{mhmodsig}[creators=hwk]{graphconnected}
\gimport{graphpath}
\symi{connected}
\end{mhmodsig}

\begin{mhmodnl}[creators=hwk]{graphconnected}{en}
\begin{definition}
  A non−empty \trefi[graph]{graph} $G$ is said to be \defi{connected} if any two
  of its \trefis[graph]{node} are linked by a \trefi[graphpath]{path} in $G$.
\end{definition}
\end{mhmodnl}

\begin{mhmodnl}[creators=hwk]{graphconnected}{de}
\begin{definition}
  Ein \mtrefi[graph?graph]{Graph} $G$ hei"st \defi[connected]{zusammenh"angend}
  wenn je zwei seiner \mtrefi[graph?node]{Knoten} durch einen
  \mtrefi[graphpath?path]{Weg} verbunden sind.
\end{definition}
\end{mhmodnl}
```

**Fig. 1.** A Glossary Entry for "onnected" graphs encoded in sTeX

The informal – and language-specific – part is given in the **language binding** given in the **mhmodnl** environment. In Figure 1 this consists of a single definition, where the definiendum is marked up via the **defi** macro and the concepts of a "graph", a "node", and a "path" are via the **trefi**, **trefi** (for plurals), and **mtrefi** (for multilinguality) macros. Their optional argument specifies the glossary module they are imported from and – after the ? the symbol name. In the German langauge binding we can appreciate the difference between the symbol name connected and its **verbalization** – the word "zusammenängend" used to denote it in German (this is the content of the **defi** macro). We will use the fact that symbols coordinate verbalizations in our multilingual glossary and mathematical dictionary (see Section 3 below).

In general Glossary entries are grouped into a **glossary module**, which is represented as $n + 1$ OMDoc/Mmt theories: the module signature and $n$ language bindings. Figure 2 shows graphconnected glossary module in the center and its 1-neighborhood wrt. the imports relation.

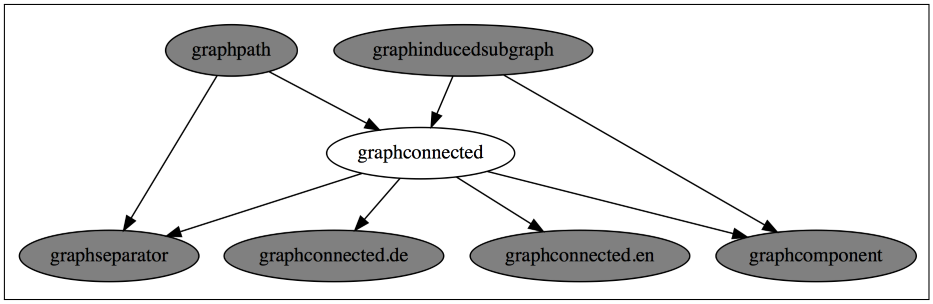


**Fig. 2.** The module Graph around "Connected Graph"

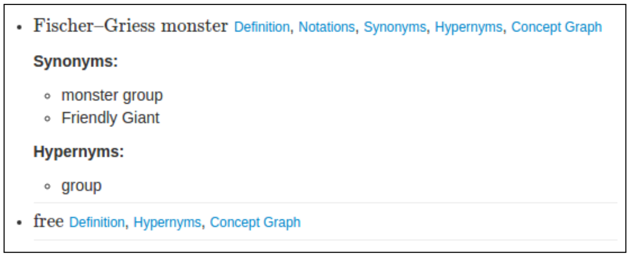
**A Terminology of Mathematics** In fact, we have all the information we need for a mathematical terminology:

*i*) we can identify **semantic fields** – here mathematical concepts, objects, and models. Oher terminologies like WordNet use "synsets": sets of synonyms for this,

*ii*) we can link semantic fields to technical terms,

*iii*) and we can relate concepts to each other via **terminological relations** like synonymy, hyper/hyponymy, and meronymy.

The last point needs some elaboration. In SMGloM we identify certain symbols as **primary** in their module, e.g. graph is primary in the grpahs module while node and edge that are defined in the same module are not. The terminlological relations can be read off from the imports relation (see [Koh14] for details): For instance, if $t$ verbalizes a primary symbol $s$ in module $m$, which imports a module $m'$ with primary symbol $s'$, which has verbalization $t'$, then $t'$ is a hyponym of $t$. For instance, "tree" is a hyponym of "graph". Figure 3 shows computed terminological relations in SMGloM.



- Fischer–Griess monster Definition, Notations, Synonyms, Hypernyms, Concept Graph

  **Synonyms:**

  - monster group
  - Friendly Giant

  **Hypernyms:**

  - group
- free Definition, Hypernyms, Concept Graph

**Fig. 3.** Terminological Relations in the Glossary

**An Ontology of Mathematics** As Figure 2 already suggests, the SMGloM data induces not only a mathematical terminology, but an ontology of mathematics as well. But note that again, the concepts of the ontology are not the glossary modules themselves, but the mathematical concepts, objects, and models, i.e. the symbols. But again, the taxonomic information can be gleaned from the module graph structure: primary symbols become ontological concepts, whereas non-primary ones become roles – detalis are mainly due to their linguistic forms. For instance the definition of a "graph" as a pair $\langle V, E\rangle$ of "vertices" $V$ and "edges" $E \subseteq V^2$ leads to the concept graph with two functional roles nodes: graph $\to$ set and vertices: graph $\to$ set. Together with the definition of the adjective "connected" on graphs in Figure 1, we get the sub-concept connected−graph which inherits these two roles from graph.

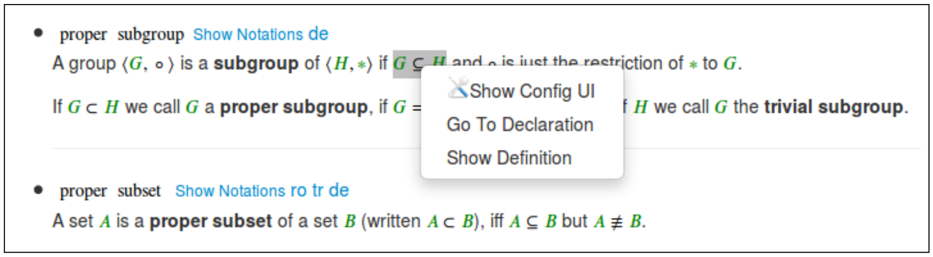
**Induced Terminological and Ontological Relations via Views** The imports relation in the module graph gives rise to direct terminological and ontological relations. But experienced mathematicians recognize more relations, for instance, that the set $E$ of "edges" of a "graph" $G = \langle V, E$ form a "relation on" $V$. We call this ability **mathematical literacy** in **??**; they can be modeled by a new form of edge in the module graph in OMDoc/Mmt: views. Epistemically, these behave like the imports relation, but their truth-preserving nature has to be proven by proof obligations. If we generalize the computation of terminological and ontological relations to allow views, then we obtain **induced** relations that are recognized – and utilized by mathematically literate users.

**Organizing a Communal Resource** The ultimate cause of the SMGloM project and system is to facilitate the establishment of a knowledge resource for mathematics. We need to take appropriate organizational measures to support this. We are currently establishing a wiki-like archive submission system for glossary modules on MathHub [MH] and thinking of a quality assurance system that is based on a community/karma-driven approval system. Openness and semantic stability are ensured by a special licensing and publication regime: The SMGloM license [SPL] protects symbols against non-conservative changes while allowing derived works.

## 3 Applications of the SMGloM

The main advantage of SMGloM over existing terminological resources for mathematics is that it makes important linguistic and ontological relations explicit that these do not. This extension makes a large variety of applications feasible without requiring full formalization, the cost of which would be prohibitive. We will sketch some of the applications here.

**Glossary of Mathematical Terms** An interface that presents SMGloM like a traditional glossary, i.e. as a (sorted) list of glossary entries. In addition, the semantic information in SMGloM can be used to adequately mark up references to as well as relations with (e.g. "synonym of", or "translation of") other entries. See Figure 4 for the current interface. There can be sub-glossaries, for certain areas of mathematics, for certain languages, etc.



- proper subgroup Show Notations de

  A group $\langle G, \circ \rangle$ is a **subgroup** of $\langle H, * \rangle$ if $G \subseteq H$ and $\circ$ is just the restriction of $*$ to $G$.

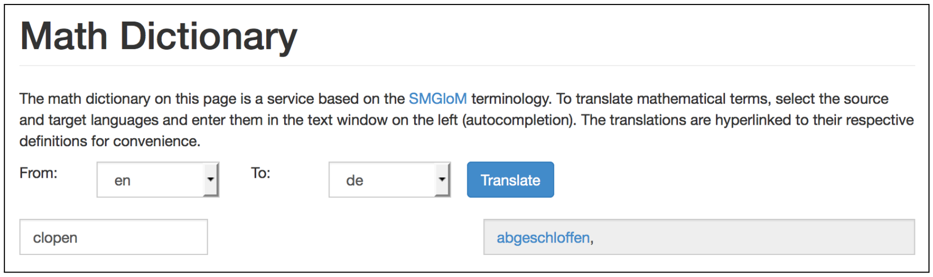
  If $G \subset H$ we call $G$ a **proper subgroup**, if $G =$ f $H$ we call $G$ the **trivial subgroup**.

- proper subset Show Notations ro tr de

  A set $A$ is a **proper subset** of a set $B$ (written $A \subset B$), iff $A \subseteq B$ but $A \not\equiv B$.

**Fig. 4.** The glossary interface at `https://mathhub.info/mh/glossary`

**Mathematical Dictionaries** The mathematical terminology is synchronized by content symbols in SMGloM, therefore a mathematical dictionary is simply an interface problem; see Figure 5. Again, all terms are hyperlinked to their definitions.



Math Dictionary

The math dictionary on this page is a service based on the SMGloM terminology. To translate mathematical terms, select the source and target languages and enter them in the text window on the left (autocompletion). The translations are hyperlinked to their respective definitions for convenience.

From: en To: de

clopen

abgeschloffen,

**Fig. 5.** The dictionary interface at `https://mathhub.info/mh/dictionary`

**Flexible Styling/Presentation** If we have formulae in content markup (i.e. in content MathML e.g. in OMDoc or sTeX), then we can adapt the rendering of formulae with symbols that having multiple notations in SMGloM to the user's preferences. Then, each user can state their notational preferences (in terms of

SMGloM notation definitions), and the formulae in SMGloM will be rendered using these, adapting to the preferences of the reader.

**Notation-Based-Parsing** The notation definitions from SMGloM can be seen as user-contributed grammar rules. Therefore, they can be used for parsing formulae from presentation to content markup [Tol16]. This will lead to a context-sensitive formula parser, where "context" is defined by the SMGloM glossary modules currently in focus – here the data model in term of OMDoc/Mmt theories directly contributes to the applications of the SMGloM.

**More Semantic Search** As SMGloM declares symbols together with notations, definitions and verbalizations it provides an unique opportunity for applying semantic search services based on it in a variety of settings:

1. notation-based parsing in the input phase could make formula entry into an interactive disambiguation process. For instance, a user enters e^?x, and the system ask her: "with e, do you mean Euler's number?", and also: "Is $e^{?x}$ a power operation?". The answers will then help refine the search.
2. Alternatively, search could use disambiguation as a facet in the search to refine the results or for clustering the results.
3. Furthermore, the SMGloM information could be used for query expansion (both visible or automatic): if the user searches for e, then the query could be expanded e.g. by *i*) the string Euler's Number (there is an interesting question about what to do with the language dependency here) and even *ii*) the formula $\lim_{?n\to\infty}(1+\frac{1}{?n})^{?n}$ ($?n$ is a query variable).

**Verbalization-Based Translation** One of the most tedious parts of translating mathematical documents is the correct use of technical terms. A semantically preloaded text (i.e. one that has all formulae in content markup and many semantic objects explicitly marked up) can be term-translated automatically using the translation relation induced by SMGloM. Of course, synonyms must be resolved consistently (there has to be an interface for this). This (and related semantic tasks) are for domain specialists. The intervening text can be done by lesser trained individuals (or even a variant of google translate). This will make translations much cheaper and will make math available in more languages.

**Wikifiers like NNexus** Wikifiers are systems that given a glossary of terms create definitional links in documents. A math-specific example is the NNexus system [GC14], it can already use the SMGloM glossary.

## 4 Conclusion & State

We have described a project to establish a public, semantic, and multilingual termbase for mathematics. We have a first prototype that supports authoring of glossary entries and glossary management at `https://mathhub.info/smglom`. The SMGloM system partially automates editing, management, refactoring, quality control, etc; for more information see `https://mathhub.info/help/main.html`.

To make public contributions to SMGloM feasible, it must already contain a nucleus of (basic) entries that can be referenced in other glossary components.

The SMGloM project is currently working towards a basic inventory of glossary entries, and has almost arrived at the first milestone of 700 entries – most with two language bindings (English and German), some with 6 (+ Romanian, Chinese, Turkish, and Bulgarian). The current glossary contains

*i)* ca. 300 glossary entries from elementary mathematics, to provide a basis for further development

*ii)* ca. 400 are special concepts from number theory to explore the suitability of the SMGloM for more advanced areas of mathematics.

*iii)* ca. 30 views that generate induced terminological and ontological relations.

**Acknowledgements** Work on the SMGloM system has been partially supported by the Leibniz association under grant SAW-2012-FIZ_KA-2 and the German Research Foundation (DFG) under grant KO 2428/13-1.

## References

[DBP13] *DBpedia.* Sept. 17, 2013. URL: http://dbpedia.org (visited on 02/21/2014).

[GC14] Deyan Ginev and Joseph Corneli. "NNexus Reloaded". In: *Intelligent Computer Mathematics 2014.* Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 423–426. ISBN: 978-3-319-08433-6. URL: http://arxiv.org/abs/1404.6548.

[Koh06] Michael Kohlhase. OMDOC – *An open markup format for mathematical documents [Version 1.2].* LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[Koh14] Michael Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics". In: *Intelligent Computer Mathematics 2014.* Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 169–183. ISBN: 978-3-319-08433-6. URL: http://kwarc.info/kohlhase/papers/cicm14-smglom.pdf.

[MH] *MathHub.info: Active Mathematics.* URL: http://mathhub.info (visited on 01/28/2014).

[RK13] Florian Rabe and Michael Kohlhase. "A Scalable Module System". In: *Information & Computation* 0.230 (2013), pp. 1–54. URL: http://kwarc.info/frabe/Research/mmt.pdf.

[SPL] Michael Kohlhase. *The SMGloM Public License (SPL) Version 0.1.* URL: https://mathhub.info/help/spl0.1.html.

[sTeX] *KWARC/sTeX.* URL: https://github.com/KWARC/sTeX (visited on 05/15/2015).

[Tol16] Ion Toloaca. "MathSemantifier – A Notation-Based Semantification Study". B. Sc. Thesis. Jacobs University Bremen, 2016.

[Wat+14] Stephan Watt et al., eds. *Intelligent Computer Mathematics.* LNCS 8543. Springer, 2014. ISBN: 978-3-319-08433-6.

[WN] *WordNet: A lexical database for English.* URL: https://wordnet.princeton.edu/ (visited on 05/26/2013).

[YAGO] *Towards a Universal Multilingual Wordnet*. URL: `http://www.mpi-inf.mpg.de/yago-naga/uwn/` (visited on 05/26/2013).