The **SMGloM** Project and System

Deyan Ginev[1], Mihnea Iancu[1], Constantin Jucovshi[1], Andrea Kohlhase[1],
Michael Kohlhase[1], Heinz Kröger[2], Jürgen Schefter[2], and Wolfram Sperber[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 vocabu-
lary. If we want to support humans with mathematical documents, we
need a resource that captures the terminological, linguistic, and ontolog-
ical 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 terminol-
ogy 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. Orig-
inally, 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 "dictio-
naries", 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 vo-
cabularies (online encyclopedias, thesauri, dictionaries, etc.), which have be-
come important resources in knowledge-based systems. This is especially true
for vocabularies that have a   *i*) semantic aspect – i.e. some of the relations
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, multi-
lingual 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, 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; in **SMGloM** natural language with $\mathrm{S\!T\!E\!X}$ annotations). Thus we consider "the definition" of a symbol to be given by a set of vernacular definitions, which are assumed to be translations of each other – an important structural invariant of the **SMGloM** that needs to be maintained.

Glossary entries are often grouped into a **glossary module**, which is represented as $n + 1$ OMDoc/Mmt theories: one for the language-independent part (called the **module signature**, it introduces the symbols, their dependencies, and notations), and n for the **language bindings** (which introduce the definitions and verbalizations of symbols).

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 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 1 for an example. There can be sub-glossaries, for certain areas of mathematics, for certain languages, etc.

branch

    1. Module `graph`

| syno- | hyper- | hypo- | mero- |nyms | de | ro | zh |

        A **graph** is a <u>structure</u> $\langle V, E \rangle$ such that V is a <u>set</u> and $E \subseteq V \times V$ is a subset of the set of <u>pairs</u> from V. We call V the <u>vertices</u> (or **node**s, **point**s, **junction**s) and E the **edge**s (or **line**s, **branch**es, **arc**s) of G.

    2. Module `inverse function`

| syno- | hyper- | hypo- | mero- |nyms | de | ro | zh |

        A **branch** of a <u>multivalued function</u> f is a <u>univalent</u> sub-relation $b \subseteq f$.

    3. ...

branch curve

Fig. 1. Sample glossary interface for SMGloM

Flexible Styling/Presentation If we have formulae in content markup (i.e. in content MathML e.g. in OMDoc or SТEX), 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 in the longer run. 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 Because 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 give a glossary of terms create definitional links in documents. A math-specific example is the NNexus system [GC13], 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 `http://mathhub.info/smglom/smglom` The SMGloM system partially automates editing, management, refactoring, quality control, etc.

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 500 entries – most with two language bindings. The current glossary contains *i)* ca. 150 glossary entries from elementary mathematics, to provide a basis for further development *ii)* ca. 350 are special concepts from number theory to explore the suitability of the SMGloM for more advanced areas of mathematics. Once the authoring and quality assurance systems have progressed sufficiently (ETA summer 2014), we want to open the system to the mathematical community. By that time, we hope to be able to show some of the applications sketched in Section 3.

References

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

[GC13] Deyan Ginev and Joseph Corneli. "NNexus Reloaded". 2013. URL: `https://svn.kwarc.info/repos/dginev/papers/cicm13-nnexus-reloaded.pdf`.

[Koh14] Michael Kohlhase. *A Data Model and Encoding for a Semantic, Multilingual Glossary of Mathematics.* submitted. 2014.

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

[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).