

A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics

Michael Kohlhase

Computer Science, Jacobs University Bremen
<http://kwarc.info/kohlhase>

Abstract. To understand mathematical language we have to understand the words of mathematics. In particular, for machine-supported knowledge management and digital libraries, we need machine-actionable terminology databases (termbases). However, terminologies for Mathematics and related subjects differ from vocabularies for general natural languages in many ways. In this paper we analyze these and develop a data model for SMGloM the Semantic, Multilingual Glossary of Mathematics and show how it can be encoded in the OMDoc/MMT theory graph model. This structured representation naturally accounts for many of the terminological and ontological relations of a semantic terminology (aka. glossary). We also demonstrate how we can account for multilinguality in this setting.

1 Introduction

Text-based information systems for mathematics and the linguistics of mathematics are still in their infancy due to the inherent complexity of mathematical documents, domains, and knowledge. One issue of particular importance is the problem of dealing with mathematical vocabularies, since they are intimately linked with both the underlying domain of mathematical knowledge and the linguistic structures that make up the particular documents. In general natural language processing, the establishment of machine-actionable terminology databases has kick-started so many applications and systems that the field is unthinkable without such resources. The SMGloM (Semantic Multilingual Glossary for Mathematics) is an attempt to jump-start similar applications.

The SMGloM system [SMG] builds on the MMT API [Rab13] and MathHub.info [CICM1414] for archiving and editing support. It supplies glossary-oriented web services that answer termbase queries, e.g. for terminological relations, definitions, or translations and generates glossaries for sub-corpora. The current glossary contains

- ca. 150 glossary entries from elementary mathematics, to provide a basis for further development and
- ca. 350 are special concepts from number theory to explore the suitability of the SMGloM for more advanced areas of mathematics.

In this paper we analyze the special needs of terminologies for Mathematics and related subjects and develop a data model for the SMGloM. This structured

representation naturally accounts for many of the terminological and ontological relations of a semantic terminology (aka. glossary).

Let us briefly recap the relevant linguistic and epistemological issues involved in terminological databases to ground our discussion of the special case of mathematical terminologies.

Glossaries Traditionally, a glossary consists of a list of technical/non-standard terms with short definitions ordered alphabetically or in the chronology of the document it illustrates. Figure 1 shows an example from Mathematics.

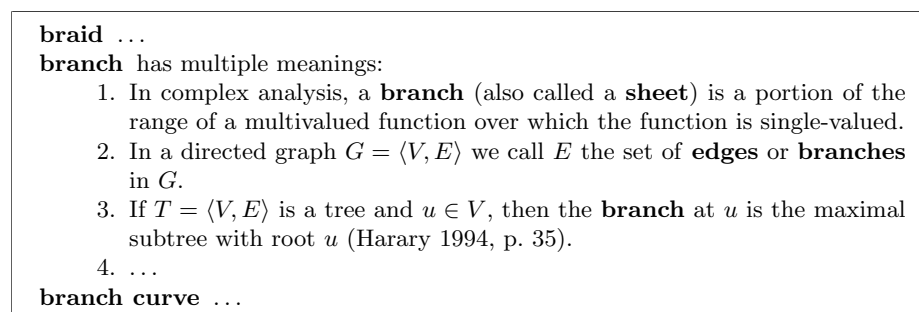


Fig. 1. A Glossary Entry for Mathematics

Terminologies Modern glossaries are usually generated from **terminologies** or **termbases** – i.e. special ontologies that organize terms and their definitions by terminological relations and/or the inherent structure of the underlying domain.

Terms are words and compound words that in specific contexts are given specific meanings. These may deviate from the meaning the same words have in other contexts and in everyday language. More specifically, we consider terms as **lexemes** which summarize the various inflectional variants of a word or compound word into a single unit of lexical meaning. Lexemes are usually referenced by their **lemma** (or **citation form**) – a particular form of a lexeme that is chosen by convention to represent a canonical form of a lexeme. Grammatical information about a lexeme is represented in a **lexicon** – a listing of the lexemes of a language or sub-language organized by lemmata.

Terminological relations are semantic relations between terms¹. The ones commonly used in terminologies are the following:

synonymy two terms are synonymous, if they have the same meaning, i.e. they are interchangeable in a context without changing the truth value of the proposition in which they are embedded.

¹ In linguistics, these relations are usually called “semantic relations”, but in the context of this note, the term “semantic” is so convoluted that we will highlight the fact that they are relations between terms.

hypernymy term Y is a hypernym of term X if every X is a (kind of) Y .
hyponymy the converse relation of hypernymy
meronymy term Y is a meronym of term X if Y is a part of X
holonymy the converse relation of meronymy
homonymy two terms are homonyms if they have the same pronunciation and spelling (but different meanings).
antonymy two terms are antonyms, if they have opposite meanings: one is the antithesis of the other.

We will call a termbase **semantic**, if it contains terminological relations and/or a representation of the domain relations.

The paradigmatic example of a termbase organized along terminological relations is WordNet [Fel98; WN]. In WordNet the synonymy relation is treated specially: the set of synonyms – called a **synset** – is taken to represent a specific entity in the world – a **semantic object** – and forms the basic representational unit of digital vocabularies. Indeed, all other terminological relations are inherited between synonyms, so it is sensible to quotient out the synonymy relation and use synsets.

Semantic terminologies are very useful linguistic resources: WordNet been used as the basis for many different services and components in information systems, including word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, and machine translation. Note that WordNet and related lexical resources do not model the relations of the objects the terms describe other than via the terminological relations above. For instance, WordNet is ignorant of the fact that a “son” is “male”² and a “child” of another “man”. In particular, definitions are not first-class citizens in WordNet-like resources, they are included into the data set for the purposes documentation, primarily so that human lexicographers can delineate the synsets. But to fully “understand” terms in their contexts – e.g. to automate processing of documents that involve such terms, and drawing inferences from them – domain relations like the ones above are crucial.

Domain Relations Semantic glossaries and digital vocabularies usually make some relations between entries explicit, so that they can be used for reasoning and applications. Linguistically, the domain relations – i.e. the relations between the (classes of) objects denoted by words – come into play in the form of **semantic roles** – the thematic relations that express the role that a noun phrase plays with respect to the action or state described by a sentence’s verb. The basic idea is that one cannot understand the meaning of a single word without access to all the essential knowledge that relates to that word.

Prominent examples of termbases with semantic roles include FrameNet [FN10; FN] and PropBank [PKG05; PB]. The former collects the semantic roles into **frames** like **Being_born** with a role **Child**, and additional roles like **Time**, **Place**, **Relatives**, etc. Such resources allow additional natural language pro-

² Do not confuse that with the grammatical gender of the word “son” is masculine or the fact that “man” is a hypernym of “son”.

cessing steps like “semantic role labeling”, which in turn allow the extraction of facts from texts, e.g. in the form of RDF triples which can then be used for textual entailment queries, question answering, etc.; see e.g. [Leh+13] for applications and references.

A Semantic, Multilingual Termbase for Mathematics For the SMGloM data model we will essentially start with the intuitions from term bases above, but adapt them to the special situation of **mathematical vernacular**, the everyday language used in writing mathematics in textbooks, articles, and to blackboards. This is a mixture of natural language, formulae, and diagrams³ all of which utilize special, domain-dependent, and dynamically extensible vocabularies. SMGloM differs from resources like FrameNet in the domain representation: we will reuse the OMDoc/MMT format for representing mathematical domains.

2 A Data Model for SMGloM

The data model of SMGloM is organized as a semantic term base with strong terminological relations and an explicit and expressive domain ontology. The terms are used as “named mathematical entities” in the sense that they are rigid designators in Kripke’s sense, rather than univalent descriptions.

2.1 Components of Terminology in Mathematics

Whereas in general natural language word meanings are grounded in the perceived world, the special vocabularies used in mathematics are usually grounded by (more or less rigorous) definitions of the mathematical objects and concepts they denote: We have learned to reliably and precisely recognize an object as a “chair” even though we have a hard time when asked to give a precise definition⁴ of what constitutes a “chair”, but we cannot directly experience a “symplectic group” and are left only with its definition to determine its meaning. In both cases, the word references an object or a set of objects that are uniform in some way so they can be subsumed under a concept; we will consider both as **semantic objects**. As mathematical objects can still have multiple “names” with which designate them, we will use the definitions themselves as the representatives of the respective semantic objects. Every definition will have an identifier which we call the **symbol** and use it for identifying the semantic object.

Note that even though the symbol name will in practice usually be derived from (the lemma of) the definiens of the definition, they are not (conceptually) the same. The technical terms normally found in glossaries arise as “verbalizations” (see Section 2.6) of symbols in diverse languages. In general there is a

³ Even though diagrams and their structural and lexical components are very interesting subject of study, we leave them to future work.

⁴ Arguably such definitions exist – take for instance Wikipedia’s page on chairs, but they are usually post-hoc and have little to do with our day-to-day use of the word and its meaning derived from this practice.

many-to-many relationship between terms and symbols: several terms pointing to the same definition, as well as several definitions communicated via the same term. In this way, symbols roughly correspond to synsets in WordNet.

But mathematical vernacular also contains formulae as special phrasal structures. We observe that formulae are complex expressions that describe mathematical objects in terms of symbols. In fact, they can be “read out” into equivalent verbal phrases, e.g. for visually impaired recipients. In this transformation, specific and characteristic parts of the formulae correspond to the symbols involved. We call these their **notations**, they act as an additional lexical component. Finally, we have the terminological and domain relations as above, only that we have to re-interpret them to the more rigorous and structured domain of mathematical knowledge.

For the purposes of SMGloM a glossary entry consists of five kinds of information, which we will describe in the rest of this section.

1. a *symbol* identified by a definition (see Section 2.2)
2. its *verbalizations* (common names; see Section 2.6)
3. its various *notations* (formula representations; see 2.5)
4. *terminological relations* to other glossary entries. (see 3.5)
5. *domain relations* to other glossary entries. (see 3.6)

2.2 Symbols and their Definitions

A **definition** consists of a **definiendum** – the term introduced in the definition – and a **definiens** – a text fragment that gives the definiendum its meaning. In the simplest of all cases, the definiens is an expression or formula that does not contain the definiendum and we can directly associate a symbol for the definiendum with the definition as an identifier. We call this case a **simple definition**.

Definition: A **directed graph** (or **digraph**) is a pair $\langle V, E \rangle$ such that V is a set of **vertices** (or **nodes**) and $E \subseteq V \times V$ is the set of its **edges**.

Fig. 2. A Definition for multiple concepts

We will rely on the reader’s mathematical experience and forego a classification of definitional forms here, but note that definitions of structured mathematical objects often naturally define more than one term. Take, for instance, the definition of a graph in Figure 2. This introduces three concepts: “**directed graph**”, “**vertex**”, and “**edge**”, which we take as symbols and the synonyms “**digraph**” for “**directed graph**” and “**node**” for “**vertex**”. We can allow such definitions in SMGloM without losing the principal one-definition-one-symbol invariant if we understand them as aggregated forms. The one in Figure 2 is an aggregation of the three definitions (one per symbol) in Figure 3. But the separation of the definitions in Figure 3 is awkward and artificial, and arguably readers would prefer to see the single definition in Figure 2 in a glossary over one of the ones in Figure 3.

<p><i>Def:</i> A directed graph (or digraph) is a pair $\langle V, E \rangle$ of sets, such that $E \subseteq V \times V$.</p> <p><i>Def:</i> Let $G = \langle V, E \rangle$ be a digraph, then V is the set of vertices (or nodes) of G.</p> <p><i>Def:</i> Let $G = \langle V, E \rangle$ be a digraph, then we call E the set of edges of G.</p>
--

Fig. 3. The Definition from Figure 2 separated into Simple Definitions

2.3 Glossary Modules

To further support grouping symbols into semantic fields, SMGloM provides **modules**: groups of definitions that belong together conceptually. SMGloM modules are conceptually similar to OPENMATH content dictionaries [Bus+04] (CDs), and we follow the lead of OPENMATH and identify symbols by their module name (c) and their symbol name s (and their CD base g , the base URI of the CDs) and write this as $g?c?s$ following MMT conventions [RK13].

Note that there is a non-trivial design decision in taking the definitions as representatives of mathematical semantic objects in SMGloM as there are often multiple, equivalent ways of defining the “same” mathematical objects. For instance, a group can be defined as a base set with a binary i) associative operation \circ that admits a unit and inverses or ii) cancellative operation $/$. These two definitions are logically equivalent, since we can define a/b as $a \circ b^{-1}$ and $a \circ b$ as $a/(b/(b/b))$. As this example already shows, logical equivalence can be non-trivial, and in many cases is only discovered a long time after the definition of the mathematical objects themselves. Therefore different definitions receive different glossary entries in SMGloM with different symbols.

In our example the two definitions give rise to two symbols `group1` and `group2`, and we do not consider them synonyms (they are in different synsets), but **homonyms** words that have different “meanings” (which are logically equivalent in this case). In a sense, the two symbols model how an objects appears to the observer, similarly to the “[evening star](#)” and the “[morning star](#)” which both refer to the planet Venus. It seems reasonable to conserve this level of modeling in a linguistic/semantic resource like SMGloM.

2.4 Symbols and Multilinguality

Another SMGloM design decision is that we allow mathematical vernacular for definitions. As written/spoken mathematical language is tied to a particular natural language, we abstract from this arbitrary choice by allowing **translations** of the definition in different languages, which we consider “indistinguishable” for a SMGloM module.

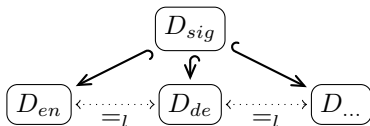


Fig. 4. Language Equality

In Figure 4 we see a situation where the content of a glossary entry D_{sig} is characterized as the equivalence class of definitions in specific languages D_* that are translations of each other – we call the translation relation **language equality** and we depict it by $=_l$; see [KK06] for an in-depth discussion on language-equality and related issues.

Concretely, a glossary module is represented as $n + 1$ **glossary components**:

- one for the language-independent part (called the **module signature**, it introduces the symbols, their dependencies, and notations, since they are largely independent of the natural language), and
- n **language bindings**, which introduce the definitions – they are written in a particular mathematical vernacular – and the language-specific verbalizations of symbols. We could imagine “language bindings” for different logical systems as a possible future extension of the SMGloM, which adds formalizations. These would behave just like the regular language bindings, only that they are fully formal.

The reason for this construction is that the vocabulary of mathematics is language-independent, because it is grounded in definitions, which can be translated – unlike general natural language vocabularies where semantic fields do not necessarily coincide.

2.5 Notations

Many mathematical objects have special symbols or formula fragments that identify them. For instance, Euler’s number is written as e and the imaginary unit of complex numbers is written as i (in mathematics, in electrical engineering it is written as j ; “standard” notations vary with the community). Parameterized or functional mathematical objects, often have complex notations, e.g. the n -th Bernoulli number is written as B_n and the special linear group of degree n over a field F is traditionally written as $SL(n, F)$. In SMGloM, we treat notations as mathematical objects themselves and reify them into notation definitions, since we want to model them in glossary components. Notation definitions are pairs $\langle \mathbb{C}, \mathbb{P} \rangle$, where \mathbb{C} is a content schema (a representation of a formula with metavariables – here indicated by $?x$) paired with a presentation \mathbb{P} of the same schema. For instance, the notation for a functional symbols like the special linear group above, the head is a pattern of the form $@(\text{slg}; ?n, ?f)$ ⁵ and the body is the formula $SL(?n, ?f)$. Notation definitions are useful in two ways: used left-to-right (i.e. given a content representation) they can be used for styling, i.e. transforming content representations (here Content MathML) to presentations (here Presentation MathML). In the other direction, they can be used for notation-based parsing – i.e. context-sensitive parsing with a dynamic (formula) lexicon.

2.6 Verbalizations

Abstract mathematical concepts (named mathematical entities; NMEs) may have multiple names – at least one per language, e.g. the English nouns “vertex” and “node” in the example in Figure 2 and the corresponding German noun “Knoten”. We specify this symbol-to-phrase relation via **verbalization definitions**, which are symbol-phrase pairs. As the NMEs are often not part of the regular lexicon of a language, we often need to specify syntactic/semantic

⁵ We will use $@(a; l)$ to denote a content MathML application of a function a to an argument list l .

information about the phrases. We do this in the form of verbalization definitions. Similarly to a notation definition, a **verbalization definition** is a pair $\langle \mathbb{C}, \mathbb{N} \rangle$, where the **head** \mathbb{C} is a content schema and the **body** \mathbb{N} is a natural language phrase schema, i.e. a phrase with metavariables. For simple cases like the verbalization “node” for the symbol **vertex** the verbalization definition is rather simple, it is just the pair $\langle \text{vertex}, \text{node} \rangle$. For functional symbols like the special linear group above, the head is a pattern of the form $\text{@}(\text{slg}; ?n, ?f)$ and the body is the text schema

[special linear group][of degree ?n][over the field ?f]

where phrases are delimited by square brackets. Note that verbalization definitions can be used in both directions like notation definitions. We use them as a linguistic resource for parsing, but also for the generation of standard glossaries or wikifiers. We abstract from grammatical information here and reduce terms and phrases to their lemmata, assuming a suitable lexicon component that manages information about inflection and aggregation schemata. For instance, with suitable notation and verbalization definitions we can generate or parse aggregated declarations like “ $\text{SL}(n, \mathbb{R})$ and $\text{SL}(m, \mathbb{C})$ are the special linear groups of orders n and m over the fields \mathbb{R} and \mathbb{C} ”.

3 Implementing the Data Model in OMDoc/MMT

We (re)-interpret the data model introduced in the last section in terms of the OMDoc/MMT theory graph (see [RK13] for a discussion of MMT theory graphs, the formal core of OMDoc). A **theory graph** is a graph, where the nodes are theories and the edges are theory morphisms: truth-preserving mappings from expressions in the source theory to expressions in the target theory. OMDoc/MMT theories are essentially collections of

- **concept declarations**, together with
- **axioms** (in particular definitions) that state what properties the concepts have, and
- **notation definitions** that specify the presentation of symbols.

Theory morphisms come in four forms:

- **structures** which define their target theory to be an extension of the source theory; **inclusions** are those structures whose mapping is the identity,
- **views** which interpret the mathematical objects of the source theory as such of the target theory (for instance, the natural numbers with addition can be interpreted as a monoid if we interpret 0 as the unit element).
- **metatheory-relations** which import the symbols of the meta-language into a theory.

Note that the notion theory morphism is rather strong in OMDoc/MMT, as it allows renaming of concepts. Structures and the meta theory relation are truth/meaning-preserving by virtue of the extension property, essentially the target theory is defined so that they are: all symbols and axioms are in the target after translation. To establish a view, we need to prove all the source axioms (after translation) in the target theory.

3.1 Glossary Components as OMDoc/MMT Theories

We can implement the SMGloM data model directly in OMDoc/MMT theory graphs – indeed the MMT API drives the SMGloM system. Note that the setup in Figure 4 can directly be represented by giving theories for the module signature and its language bindings and interpreting the dependencies as OMDoc/MMT inclusions. There is however one problem we still need to solve: the module signatures introduce the symbols of the glossary module, but their meaning is specified in the language bindings, which include them. Therefore we need to extend MMT with a new feature: **adoptions**, i.e. views from the language bindings to the module signatures that are definitional – they establish the meaning of the symbols in the module signature by postulating that definitions in the language bindings hold there. Note that in this sense, adoptions are similar to structures – only that meaning travels in the reverse direction. Like these, adoptions do not induce proof obligations. The lower half of Figure 5 shows the situation, the double squiggly arrows are the adoptions. Note that the adoptions can only work, if the definitions in the various language bindings are translations – indeed the adoptions postulate them, but we cannot check them in the SMGloM system.

But we can use the theory graph to even more advantage in SMGloM, if we take the MMT meta-level into account. We can model the fact that e.g. the language binding D_{en} is written in English by specifying the theory MV_{en} (English mathematical vernacular) as its meta-theory. In Figure 5, we find the module/bindings construction of Figure 4 at the bottom layer, and their vernaculars in the layer above. These inherit from generic language theories L_* and a module signature MV for mathematical vernacular⁶. Note that the mathematical vernacular meta-level (the middle layer in Figure 5) is structurally isomorphic to the domain level. In particular, we can think of MV as a signature of mathematical vernacular: it contains symbols for meta-mathematical concepts like quantification, connectives, definitional equality, etc. In future extensions of the SMGloM by formal content, this is the spaces, where the logics would live – see [Cod+11].

The third level in Figure 5 contains the generic (i.e. non-mathematical) vocabularies of the respective natural languages. They are just stubs in SMGloM that can be coupled with non-math-specific linguistic/lexical resources in the future.

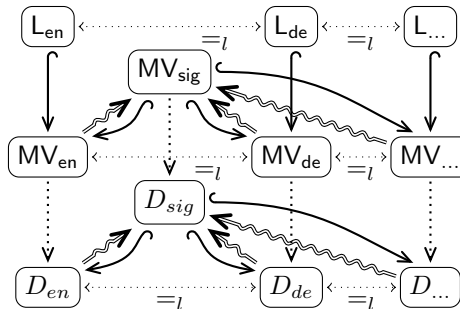


Fig. 5. The Language Metalevel

⁶ Actually, what we have depicted as a single theory here is a whole theory graph of inter-dependent theories.

3.2 Multilingual Theory Morphisms

In the SMGloM, where glossary items are structured, multilingual modules (see Figure 4), theory morphisms are similarity structured. Consider the situation on the right, where we have a module MS for metric spaces, and another (NVS) for metric vector spaces.

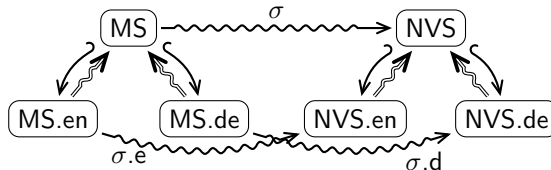


Fig. 6. A multilingual view

It is well-known that a normed vector space $\langle V, \|\cdot\| \rangle$ induces a distance function $d(x, y) := \|x - y\|$ and thus a metric space $\langle V, d \rangle$. The OMDoc/MMT views that make up this structured relation between glossary modules is represented by the three wavy arrows in Figure 6. Here σ is the translation that assigns the base set V to itself and $d(x, y)$ to $\|x - y\|$. The two OMDoc/MMT views $\sigma.e$ and $\sigma.d$ include σ and add the proofs (in English and German respectively) for the proof obligations induced by the metric space axioms.

3.3 Notations & Verbalizations

We employ OMDoc notation definitions which directly implement the content/presentation pairs in XML syntax (see [Koh10] for details). It turns out that for the structurally similar verbalization definitions introduced in Section 2.6, we re-use the OMDoc/MMT notation definitions mechanism, only that the “presentation” component is not presentation MathML, but in natural language phrase structures (in the respective languages).

3.4 Synsets: Direct Synonymy

We have two forms of “synonyms” in SMGloM: **direct synonyms** that are directly given in definitions, and induced ones (see below). For example, the definition in Figure 2 introduces the terms “**vertex**” and “**node**” as direct synonyms. Indeed, the definiendum markup gives rise to the verbalization definitions $\langle \text{dgraph?vertex}, \text{vertex} \rangle$ and $\langle \text{dgraph?vertex}, \text{node} \rangle$ respectively, i.e. the lemmata “**vertex**” and “**node**” refer to the symbol **vertex** in the theory **dgraph**. In essence we use symbol-synchronization for the representation of direct synonyms, and thus we can use the symbols as representations of synsets of the SMGloM term base. Note that this interpretation also sees translations as special cases of synonyms, as they also refer to the same (language-independent) symbol. In SMGloM we identify synsets with symbols and thus model terminological relations as relations between symbols. This allows us to model them as theory morphisms and use the OMDoc/MMT machinery to explain their contributions and properties. For the moment we restrict ourselves to inclusions and leave structures and views to Section 3.6.

3.5 Direct Terminological Relations

In OMDoc/MMT theory graphs, we often have a systematic dualism between the theory T as a structured object and the mathematical structure⁷ it introduces, we call it the **primary object** and denote it with \bar{T} , all other symbols are called **secondary**. Consider for instance the case of directed graphs above, where the theory has secondary symbols for vertices and edges; and incidentally, the primary object of the glossary module in Figure 2 is the concept a digraph, i.e. the structure $\langle V, E \rangle$ which consists of (sets of) vertices and edges (both secondary concepts). Similarly, the theory of groups has a primary object made up of its secondary objects: the base set, the operation, the unit, and the inverse operation.

In our experience, secondary symbols mostly (all?) seem to be functional objects whose first argument is the primary symbol. For instance the “edges of” a graph. This makes the setup of SMGloM modules very similar to classes in object-oriented classes, where the secondary objects correspond to methods, and (more importantly for a linguistic resource like SMGloM) to frames in FrameNet, where the secondary symbols correspond to the semantic roles. We will conduct a survey on this on the SMGloM corpus once its more mature.

Hyper/Hyponymy For the hyponymy and hyperonymy relations, we employ the notion of theory morphisms from OMDoc/MMT. If there is an import from S to T , then \bar{T} is a hypernym of \bar{S} and that a hyponym of S . Consider for instance, the notion of a “tree” as a digraph with special properties (a unique initial node and in-degree 1 on all others). Extending the digraph glossary module to one for trees naturally gives rise to an inclusion morphism that maps the principal symbol **digraph** to the new principal symbol **tree**. Thus the term “tree” is a hypernym of “digraph” (and “directed graph”, since that is a direct synonym).

For the secondary symbols we have a related effect. They are usually inherited along theory morphisms together with the primary symbols, but they keep their meaning, only that their domain is restricted to the more specialized primary symbol. This relation which we tentatively call **domain restriction** is related to the notion of **selectional restriction** in lexical semantics – cf. [Ash14] for a recent contribution that seems compatible with the SMGloM data model.

Meronymy Note that the inclusion relation we have encountered above is very naturally a theory morphism by construction: all objects and their properties of the source theory are imported into the target theory. As the imports relation is invoked whenever a mathematical object is referenced (used) in the definiens of another, we interpret the inclusion relation as the SMGloM counterpart of the meronymy relation: if there is an import from theory S to theory T , then \bar{S} is part of \bar{T} . Take for instance a definition of a ring $\bar{\text{ring}} = \langle R, +, 0, -, *, 1 \rangle$ via an inclusion from a commutative group $\bar{\text{grp}} = \langle R, +, 0, - \rangle$ and a monoid

⁷ We have an unfortunate name clash with MMT “structures” here we mean the mathematical object, e.g. the pair $\langle V, E \rangle$ in Figure 2.

$\overline{\text{mon}} = \langle R, *, 1 \rangle$. This directly gives us two meronymy relations: The monoid $\overline{\text{mon}}$ and the commutative group $\overline{\text{grp}}$ are both “parts of” the $\overline{\text{ring}}$. As a consequence, inclusions where the primary symbol of the source is not mapped to the primary symbol of the target theory give rise to meronymy relations between the primary symbols.

3.6 Induced Terminological Relations

We now turn to the other kind of theory morphisms: structures and views and their contribution to terminological relations. We first observe that structures and views bridge a greater conceptual distance than inclusions and adoptions, since the induced mapping is not the identity. Note that the distinction made here between inclusions and structures is a gradual one based on the complexity of the mapping. In particular, structures with injective symbol mappings may seem closer to inclusions than to structures that map to complex terms. Moreover, while inclusions and structures are definitional (their targets are defined in terms of them), views carry proof obligations that show their truth-preserving nature; this translates into an even greater cognitive distance of the **induced terminological relations**.

Homonymy Logical equivalence of glossary modules – i.e. homonymy of the terms that verbalize the primary symbols – is just a case of theory isomorphism. In the example with the two groups from Section 2.3 we have two SMGloM modules which are represented OMDoc/MMT theories. Their equivalence can be encoded by a theory isomorphism: two views which compose to the identity. As any logical equivalence can be expressed as theory isomorphisms (given suitable glossary modules), homonymy is conservative over OMDoc/MMT theory graphs.

View-Induced Hyponymy (aka. Examples) We have already seen that theory inclusions induce hyponymy (the “isa relation”) between the principal symbols, e.g. a group “is a” monoid. The “induced hyponymy relation” – e.g. $\langle \mathbb{N}, + \rangle$ “is a” monoid if we interpret 0 as the unit element is very salient in mathematics: we consider $\langle \mathbb{N}, +, 0 \rangle$ as an example of a monoid. The proof obligations of the underlying view verify that this is indeed true. Giving examples – and counter-examples – from other mathematical areas is an important mathematical practice necessary for fully understanding mathematical concepts and fostering intuitions about applications. Regular hyponyms are usually not considered good examples, since they are too direct.

Induced Synonymy In a graph definition in [Har69] we find the terms “**0-simplex**” for the nodes and “**1-simplex**” for the edges of a graph. We interpret such “synonyms” as metaphoric. Given a definition

Definition: A **k -simplex** is a k -dimensional polytope which is the convex hull of $k + 1$ affinely independent points in k -space.

Harary’s definition makes sense if we map nodes to 0-simplices and edges to 1-simplices. In SMGloM we would model this via a glossary module for simplices and a view from the graph module to that. Then we can understand Harary’s names as synonyms induced by this view. Note that in order for these to be “synonyms” in the sense of this paper, we also need a (partial) view back from simplices to edges (that is defined on them), but that is also easy to do. The view directly accounts for the metaphoric character. We “borrow” terms for graphs from a related (via the view) field of simplices.

As the conceptual gap covered by views can vary greatly – the identity endomorphism covers none – the distinction between direct- and view-induced synonyms is flexible (and in the mind of the beholder). A first delineation could be whether the analogy mappings that give rise to the (originally metaphoric) names are inner-mathematical or extra-mathematical. If they are inner-mathematical then we should state the views, if they are not, then we cannot really. An example of synonyms introduced by an extra-mathematical (from plant anatomy) view is the junction/branch metaphor for vertices/edges in graphs. Given these criteria, it becomes debatable whether to interpret the synonyms point/line for vertex/edge via a view into point/line geometry.

A very positive effect of interpreting synonyms via views is that this also gives an account of the coordination of synonyms. We observe that verbalizations are coordinated in “conceptual systems”. In particular, we will seldom find “mixed metaphors” in Math, where people use the word “point” for the concept of a vertex and “branch” for an edge in the same situation. Requiring the existence of a view that maps the whole situation into a coherent glossary module explains this observation nicely. Similar consideration should hold for notations, but we will leave their study to future work.

4 Conclusion

We have presented a data model model for a mathematical termbase. As mathematical terminologies are based in definitions not in perceptions of the physical world, modeling the mathematical domain becomes as important as modeling the terminological relations for a machine-actionable resource. We integrate both aspects by modeling glossary terms by OMDoc/MMT symbols, glossary modules (semantic fields of terms correlated by their meaning) as theories, and terminological relations by theory morphisms, so that we can make use of the OMDoc/MMT machinery – and even implementation – for a glossary system.

Our treatment of multilinguality has some similarity to the GF Mathematical Library (MGF [SX11]), but that concentrates on grammatical aspects where we focus on terminological relations and definitions. An extension of SMGloM by GF-based grammatical information seems an attractive avenue of future research as the OMDoc/MMT domain model is an extension of the OpenMath CDs used in MGF and the GF framework is based on a type theory that is similar to the meta-level of OMDoc/MMT.

Currently, the SMGloM only contains a handful of views to establish the concept and serve as examples. As we have seen above, views give rise to interesting semantic/linguistic phenomena, so this is where we have to invest most of the curation efforts.

Eventually, we will support multiple surface syntaxes for OMDoc, but initially, we use sTeX, a semantical variant of L^AT_EX; see [Koh08; sTeX].

An feature of mathematical domain modeling which we have not included in the SMGloM is the assignment of sorts/types to mathematical concepts. This is probably the most immediate next step after consolidating the initial corpus to the data model described in this paper: Sortal and type-restrictions are important cognitive devices in semantic domains and representing them significantly enhances the expressivity and adequacy of lexical/linguistic as well as logical modeling. But the integration of linguistic and logical constraints – in particular selectional restrictions of verbs and adjectives – into a universal sort/type system for mathematics is no small feat, therefore we leave it to future work. But we conjecture that the SMGloM data model of a mathematical term base with a theory graph structure is the right setting to investigate **selectional restriction** in lexical semantics. We plan to use all “unary predicate symbols” in SMGloM as possible types and study what this means for the selection restrictions taking [Ash14] into account as a departure for this work.

Acknowledgements Work on the concepts presented here 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. The development of the data model has profited from discussions with Deyan Ginev, Wolfram Sperber, and Mihnea Iancu.

References

- [Ash14] Nicholas Asher. “Selectional Restrictions, Types, and Categories”. In: *Journal of Applied Logic* 12.2 (2014), pp. 75–87.
- [Bus+04] Stephen Buswell et al. *The Open Math Standard, Version 2.0*. Tech. rep. The OpenMath Society, 2004. URL: <http://www.openmath.org/standard/om20>.
- [CICM1414] Mihnea Iancu et al. “System Description: MathHub.info”. In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. Lecture Notes in Computer Science. accepted. Springer, 2014. URL: <http://kwarc.info/kohlhase/submit/cicm14-mathhub.pdf>. Forthcoming.
- [Cod+11] Mihai Codescu et al. “Project Abstract: Logic Atlas and Integrator (LATIN)”. In: *Intelligent Computer Mathematics*. Ed. by James Davenport et al. LNAI 6824. Springer Verlag, 2011, pp. 289–291.
- [Fel98] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

- [FN] *Frame Net*. URL: <https://framenet.icsi.berkeley.edu> (visited on 02/06/2014).
- [FN10] Josef Ruppenhofer et al. *FrameNet II: Extended Theory and Practice*. 2010. URL: <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>.
- [Har69] Frank Harary. *Graph Theory*. Addison Wesley, 1969.
- [KK06] Andrea Kohlhase and Michael Kohlhase. “An Exploration in the Space of Mathematical Knowledge”. In: *Mathematical Knowledge Management, MKM’05*. Ed. by Michael Kohlhase. LNAI 3863. Springer Verlag, 2006, pp. 17–32. URL: <http://kwarc.info/kohlhase/papers/mkm05.pdf>.
- [Koh08] Michael Kohlhase. “Using L^AT_EX as a Semantic Markup Format”. In: *Mathematics in Computer Science 2.2* (2008), pp. 279–304. URL: <https://svn.kwarc.info/repos/stex/doc/mcs08/stex.pdf>.
- [Koh10] Michael Kohlhase. “An Open Markup Format for Mathematical Documents OMDoc [Version 1.3]”. Draft Specification. 2010. URL: <https://svn.omdoc.org/repos/omdoc/branches/omdoc-1.3/doc/spec/main.pdf>.
- [Leh+13] Jens Lehmann et al. “DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia”. In: *Semantic Web Journal* (2013), p. 29. URL: <http://www.semantic-web-journal.net/system/files/swj558.pdf>.
- [PB] *Proposition Bank*. URL: <http://verbs.colorado.edu/~mpalmer/projects/ace.html> (visited on 02/06/2014).
- [PKG05] Martha Palmer, Paul Kingsbury, and Daniel Gildea. “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: *Computational Linguistics* 31.1 (2005), pp. 71–106. DOI: 10.1162/0891201053630264.
- [Rab13] Florian. Rabe. “The MMT API: A Generic MKM System”. In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics. (Bath, UK, July 8–12, 2013). Ed. by Jacques Carette et al. Lecture Notes in Computer Science 7961. Springer, 2013, pp. 339–343. DOI: 10.1007/978-3-642-39320-4.
- [RK13] Florian Rabe and Michael Kohlhase. In: *Information & Computation* 0.230 (2013), pp. 1–54. URL: <http://kwarc.info/frabe/Research/mmt.pdf>.
- [SMG] *SMGloM Glossary*. URL: <http://mathhub.info/mh/glossary> (visited on 04/21/2014).
- [sTeX] *Semantic Markup for L^AT_EX*. Project Homepage. URL: <http://trac.kwarc.info/sTeX/> (visited on 02/22/2011).
- [SX11] Jordi Saludes and Sebastian Xambó. “The GF Mathematics Library”. In: *THedu*. Ed. by Pedro Quaresma and Ralph-Johan Back. Vol. 79. EPTCS. 2011, pp. 102–110. DOI: 10.4204/EPTCS.79.6.

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