Alexander M. Elizarov, Alexander V. Kirillovich, Evgeny K. Lipachev, Olga A. Nevzorova, Valery D. Solovyev, Nikita G. Zhiltsov

MATHEMATICAL KNOWLEDGE REPRESENTATION: SEMANTIC MODELS AND FORMALISMS

Kazan Federal University, N. I. Lobachevskii Institute of Mathematics and Mechanics, Higher Institute of Information Technologies and Information Systems, Kremlevskaya Str. 18, 420008, Kazan, Russia;
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Baumana Str. 20, 420111, Kazan, Russia

E-mail address: {amelizarov, alik.kirillovich, elipachev, onevzoro}@gmail.com, maki.solovyev@mail.ru, nikita.zhiltsov@gmail.com

Abstract. The paper provides a survey of semantic methods for solution of fundamental tasks in mathematical knowledge management. Ontological models and formalisms are discussed. We propose an ontology of mathematical knowledge, covering wide range of fields of mathematics. We demonstrate applications of this representation in mathematical formula search, and learning.

1. Introduction

The rapid growth of the modern science requires effective purpose-built information systems. Since inception of the first scientific information systems, mathematicians have been involved in the full cycle of software product development, from idea to implementation. A well-known example is TEX, an open source typesetting system designed and mostly written by Donald Knuth [1]. TEX has a solid community of developers, researchers, and enthusiasts, who contribute new packages [2]. The reader is likely aware of Mathematica [3] and WolframAlpha [4] commercial systems, led by a mathematician and physicist Stephen Wolfram according to his principles of computational knowledge

---

2000 Mathematical Subject Classification. 68T30, 68P20.

Key words and phrases. Ontology engineering, mathematical knowledge, metadata extraction, information retrieval, math formula search.

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities (project 3056).

theory (see e.g. [5]). Tools for mathematical content management are developed with the help of communities of mathematicians, e.g. MathJax [6, 7] by American Mathematical Society, as well as independent researchers, e.g. ASCIIMathML [8]. Math-Net.Ru [9], a collection of publications from refereed journals, and arXiv.org, a collection of publicly available pre-prints, are information systems that benefit from contributions of the mathematical community. The similar situation can be seen in other natural sciences. For example, there are examples of information systems developed by chemists [10, 11]. However, the contemporary science community clearly lacks information systems, covering all its needs.

Main challenges in mathematical knowledge management (MKM) are discussed in [12] – [19]. Further, we frame the most urgent tasks:

- modeling representations of mathematical knowledge, i.e., techniques for representing MKM include data structures, logics, formal theories, diagrams;
- presentation formats, i.e., formats, programming languages etc.;
- authoring languages and tools;
- creating repositories of formalized mathematics, and mathematical digital libraries;
- mathematical search and retrieval, i.e., querying collections of mathematical documents;
- implementing math assistants, tutoring and assessment systems;
- developing collaboration tools for mathematics;
- creating new tools for detecting repurposing material, including plagiarism of others' work and self-plagiarism;
- creation of "live documents" [20];
- creation of interactive documents, e.g. efforts of the Liber Mathematicae community [21, 22] and Computable Document Format (CDF) [23] by Wolfram;
- developing deduction systems, i.e., theorem provers and computer algebra systems (e.g. [24, 25]). The solution of this task requires rigid formalization of mathematical statements and proofs.

While mathematics is full of formalisms, there is currently no a single widely accepted formalism for computer mathematics. To tackle this issue, we describe an approach that is based on Semantic Web models and technologies [26]. At the core of integration of mathematical resources, there is building structured representation of the scientific content. World Wide Web Consortium (W3C) (www.w3.org) is an international community to develop standard and technologies of Semantic Web, including special purpose markup languages for many domains.

In this paper, we elaborate semantic-based approaches to solve some of the tasks described above. In Section 2, we outline existing semantic models for

mathematical documents. In Section 3, we present $OntoMath^{PRO}$, a novel ontological model for mathematics that was developed by the authors together with mathematicians from Kazan Federal University. Section 4 contains concrete applications in search as well as education powered by the ontology.

2. Semantic Models of Mathematical Documents

In this section, we give an overview of state-of-the-art semantic models of mathematical documents.

2.1. **Semantic markup for formulas.** Semantic markup enables automatic intelligent information processing. For representation of mathematical formulas, there has been developed Mathematical Markup Language (MathML) [27]. MathML was designed by W3C as a machine-readable language to both present and consume mathematical content in WWW. The increasing role of this language in mathematical content management is discussed in [28].

Widely used tools for authoring mathematical articles include LaTeX-based integrated development environments and office packages with mathematical formula support, such as MS Word+MathType. MathML Word2TeX [29] and LaTeXML [30] can be leveraged to convert documents from popular formats to XHTML+MathML for publication in Web.

2.2. **High-level models.** Open Mathematical Documents (OMDoc) [31], an XML-based language, is integrated with MathML/OpenMath and adds support of statements, theories, and rhetorical structures to formalize mathematical documents. OMDoc has been used for interaction between structured specification systems and automated theorem provers [32]. The OMDoc OWL Ontology (available at `http://kwarc.info/projects/docOnto/omdoc.html`) is based on the notion of statements. Sub-statement structures include definitions, theorems, lemmas, corollaries, proof steps. The relation set comprises of partonomic (whole-part), logical dependency, and verbalizing properties. The paper [33] presents an OMDoc-based approach to author mathematical lecture notes using STeX macro package [33, 34, 35] in LaTeX and expose them as Linked Data accessible in Web. STeX offers macros for introducing new mathematical symbols and using arbitrary metadata vocabularies. STeX is integrated with OMDoc ontology, providing definitions of OpenMath symbols and elements of the logical structure of mathematical documents, such as theorems and proofs. This model also makes such documents directly available from the Web converting them to XHTML/ RDFa format and offers different types of services like notation explanation, versioning and semantic search.

The MathLang Document Rhetorical (DRa) Ontology [36] characterizes document structure elements according to their mathematical rhetorical roles that are similar to the ones defined in the statement level of OMDoc. This

semantics focuses on formalizing proof skeletons for generation proof checker templates.

The Mocassin Ontology [37] encompasses many structural elements of the models mentioned above. However, this model is more oriented on representing structural elements and relations between them, e.g. logical dependency or referencing, occuring frequently in published scholarly papers in mathematics. In [37] we demonstrate its utility in the information extraction scenario.

2.3. **Terminological resources.** Terminological resources, such as vocabularies, datasets, thesauri, and ontologies include descriptions of mathematical knowledge objects.

The general-purpose DBpedia dataset [38] contains, according to our estimates, about 7,800 concepts (including 1,500 concepts with labels in Russian) from algebra, 46,000 (9,200) concepts from geometry, 30,000 (4,300) concepts from mathematical logic, 150,000 (28,000) from mathematical analysis, and 165,000 (39,000) concepts on theory of probability and statistics.

The ScienceWISE project [39] gives over 2,500 mathematical definitions, including concepts from mathematical physics, connected with subclass-of, whole-part, associative, and importance relationships.

The Online Encyclopedia of Integer Sequences [40] is a knowledge base of facts about numbers. Given a sequence of integers, this service (`http://oeis.org`) displays the information about its name, general formula, implementation in programming languages, successive numbers, references, and other relevant information.

Cambridge Mathematical Thesaurus [41] contains a taxonomy of about 4,500 entities in 9 languages from the undergraduate level mathematics, connected with logical dependency and associative relationships.

3. Ontologies as Formalisms for Mathematical Knowledge Representation

We introduce ontology-based formalisms for knowledge representation as well as our novel ontological model for mathematics.

3.1. **Basic terms.** Both knowledge representation and knowledge interchange between information agents, such as researchers and information systems, rely on a conceptualization [42]. Each communication agent has its own vocabulary to refer to elements of the conceptualization. Therefore, discrepancy between agent protocols can occurr for two reasons: i) agents may have different conceptualizations; ii) they may have incompatible models of languages, i.e., meanings of terms. Effective communication requires a single conceptualization as well as the sharable vocabulary. Ontologies suffice this requirement.

Improving the classical definition by T.Gruber [43], the authors of [44] define an ontology as "a formal, explicit specification of a shared conceptualization".

An ontology defines basic concepts and relations between them of a given domain and includes:

- classes
- properties
- restrictions.

Hence, we accept the formal approach to ontology definition given by N. Guarino according to formal semantics [45].

Definition 1. *An extensional relational structure is a tuple* $S = (D, R)$ *where*

- D *is a set called the universe of discourse*
- R *is a set of relations on* D*.*

Let W the set of world states (also called worlds, or possible worlds) for an area of interest.

Definition 2. *A conceptual relation (or intensional relation)* ρ^n *of arity* n *on* $< D, W >$ *is a total function* $\rho^n : W \to 2^{D^n}$ *from the set* W *into the set of all* n*-ary (extensional) relations on* D*.*

From Definition 2, we can provide a formal definition of conceptualization.

Definition 3. *A conceptualization (or intensional relational structure) is a triple* $C = (D, W, \mathfrak{R})$ *with*

- D *a universe of discourse;*
- W *a set of world states;*
- $\mathfrak{R}$ *a set of conceptual relations on the domain space* $< D, W >$*.*

Ontological commitment establishes the proper meanings of vocabulary elements. Let $\mathbf{L}$ be a first-order logical language with vocabulary $\mathbf{V}$ and $\mathbf{C} = (D, W, \mathfrak{R})$, a conceptualization.

Definition 4. *An ontological commitment (or intensional first order structure) for* $\mathbf{L}$ *is a tuple* $\mathbf{K} = (\mathbf{C}, \mathfrak{I})$*, where* $\mathfrak{I}$ *(called intensional interpretation function) is a total function* $\mathfrak{I} : V \to D \cup \mathfrak{R}$ *that maps each vocabulary symbol of* $\mathbf{V}$ *to either an element of* D *or an intensional relation belonging to the set* $\mathfrak{R}$*.*

Let $I : V \to D \cup R$ be any function that maps vocabulary to the union of elements and relations of the universe of discourse (called extensional interpretation function), and S is from Definition 1. An intended model is a model that conforms the chosen ontological commitment, or formally

Definition 5. *A model* $M = (S, I)$ *is called an intended model of* $\mathbf{L}$ *according to* $\mathbf{K}$ *if*

(1) *for all constant symbols* $c \in \mathbf{V}$ *we have* $I(c) = \mathfrak{I}(c)$;

(2) *there exists a world state* $w \in W$ *such that, for each predicate symbol* $v \in \mathbf{V}$ *there exists an intensional relation* $\rho \in \mathfrak{R}$ *such that* $\mathfrak{I}(v) = \rho$ *and* $I(v) = \rho(w)$.

Finally, the ontology is defined as follows:

Definition 6 (Ontology)**.** *An ontology* $\mathbf{O_K}$ *for ontological commitment* $\mathbf{K}$ *is a logical theory consisting of a set of formulas of* $\mathbf{L}$*, constructed so that the set of its models matches as close as possible the set of intended models of* $\mathbf{L}$ *according to* $\mathbf{K}$.

The ontology can be expressed in various formalisms. The most ubiquitous languages are F-logic [46], and, particularly, description logics languages [47]. In practice, Web Ontology Language (OWL) [48], a knowledge representation language founded on a description logic SHIQ, is the most used in the Semantic Web community.

3.2. **OntoMath^{PRO}.** *OntoMath^{PRO}* [49] is the first attempt to build an ontology of mathematical knowledge objects according to principles described above.

Hence, we apply formalisms from the previous section to mathematics. We assume that, in our case, the universe of discourse is mathematical objects from scientific refereed publications. The conceptualization for mathematics is principles for classication of objects according to their characteristics. The vocabulary represents the mathematical terminology. The ontological commitment is meanings of mathematical terms widely accepted in the contemporary mathematical community. Then, the ontology captures the accepted conceptualization and the ontological commitment.

The current version of *OntoMath^{PRO}* contains concepts from the preselected fields of mathematics, such as number theory, set theory, algebra, analysis, geometry, mathematical logic, discrete mathematics, theory of computation, differential equations, numerical analysis, probability theory, and statistics. The ontology defines six relations, such as taxonomic relation, logical dependency, associative relation between objects, belongingness of objects to fields of mathematics, and associative relation between problems and tasks.

Each mathematical concept is represented as a class in the ontology. The class has definitions both in Russian and English, relations with other classes, and links to verified Semantic Web resources [38, 39].

The current version of ontology has 3,449 classes, 3,627 taxonomic and 1,139 non-taxonomic relations. We distinguish two hierarchies of classes: a taxonomy of the fields of mathematics and a taxonomy of mathematical knowledge objects. In the taxonomy of fields, most fundamental fields, such as geometry and analysis, have been elaborated thoroughly. For example, there have been defined specific sub-fields of geometry: analytic geometry, differential geometry, fractal geometry and others. There are three types of top level concepts in

the taxonomy of mathematical knowledge objects: i) basic metamathematical concepts, e.g. Set, Operator, Map, Function, Predicate etc; ii) root elements of the concepts related to the particular fields of mathematics, e.g. Element of Logics; iii) common scientific concepts: Problem, Method, Statement, and Formula. Concrete theoretical results, e.g. Arslanov's completeness criterion.

4. Applications

We present applications of the proposed semantic models for mathematical formula search and learning.

4.1. **Mathematical formula search.** We have implemented two applications for mathematical formula search: syntactical search of formulas in MathML, and semantic ontology-based search.

The syntactical search leverages formula parts from documents formatted in TEX. Our algorithm [50] transforms formulas in TEX to MathML. We set up an information retrieval system prototype for a collection of articles in Lobachevskii Journal of Mathematics (LJM, `http://ljm.ksu.ru`). For the end-user, the query input interface supports a convenient TEX syntax. The search hit description includes hightlighted occurrences of formulas as well as document metadata.

In our previous work [51], we have developed a semantic publishing platform for scientific collections in mathematics that analyzes the underlying semantics in mathematical scholarly papers and effectively builds their consolidated ontology-based representation. The current data set contains a semantic representation of articles of "Proceedings of Higher Education Institutions: Mathematics journal".

Our demo application (`http://cll.niimm.ksu.ru/mathsearch`) features a use case of querying mathematical formulas in the published dataset that are relevant to a given mathematical concept. The supported user input is close to a keyword search: our system is agnostic to a particular symbolic notation used to express mathematical concepts, and the user is able to select query suggestions by keywords. Our search interface also supports filtering by the document structure context, i.e., a particular segment of the document (e.g. a theorem or a definition) that contains the relevant formula.

4.2. **Learning.** For a practicing mathematician, an ability of solving problems is crucial. The proficient solver must realize relationships between particular methods, tasks, and proof techniques to make the transition from solving problems to proving theorems [52]. We describe our experiments on ontology-based assessment of the competence of students, who attended a course on numerical analysis.

For our experiments, we extracted a small fragment of $OntoMath^{PRO}$ ontology. It contains taxonomies of tasks and solving methods for systems of linear equations (numerical analysis) as well as relationships between them.

The experiment participants were students who attended the course and had high overall grades. Each participant is given a list of classes and asked to link them using only two relationships: taxonomic relation and *solves*. Therefore, we treat this task as a classification task. We use standard performance measures for classification tasks, such as precision (P), recall (R), and F-score $= 2 \cdot \frac{P*R}{P+R}$.

According to our results, reconstruction of concept properties is the hardest task (35% F-score on average) for most students comparing to reconstruction of taxonomies (83%). It means that the ontology could be used by students to conceive the correct conceptualization of a field of mathematics. The detailed analysis of the experiments is provided in [49].

5. Conclusion

The paper summarizes the key tasks in mathematical knowledge representation. We give an overview of state-of-the-art semantic models of mathematical documents. We introduce ontology-based formalisms for knowledge representation as well as our novel ontological model, $OntoMath^{PRO}$, for mathematics. We present applications of the proposed semantic models for mathematical formula search and learning.

We emphasize that while the ontology has achieved maturity, it is the result of ongoing work. The ontology is publicly available on `ontomathpro.org`. On this webpage, we encourage our colleagues to take part in collaborative editing, including correction and contributing new classes, relations, and definitions. We also organize a discussion to prospect novel applications.

Acknowledgments: A. Kirillovich would like to thank Evelina Khakimova (University of Virginia), Claudia Acevedo (Lemoine Editores), and Maria Isabel Duarte (EAFIT University) for the assistance in the work with bibliographic sources.

References

[1] D. E. Knuth, *The TEX book* (Addison-Wesley Publishing Company, 1986).
[2] CTAN. Comprehensive TEX Archive Network. URL: `http://www.ctan.org/`
[3] Wolfram Mathematica. URL: `http://www.wolfram.com/mathematica/`
[4] WolframAlpha computational knowledge engine. URL: `http://www.wolframalpha.com/`
[5] S. Wolfram, *A New Kind of Science* (Wolfram Media, Inc., 2002).
[6] MathJax. Beautiful math in all browsers. URL: `http://www.mathjax.org/`
[7] D. Cervone, Notices of the American Mathematical Society **59** (2), 312-316 (2012).

[8] ASCIIMathML.js (ver 2.0): Translating ASCII math notation to MathML and graphics. URL: http://www1.chapman.edu/~jipsen/mathml/asciimath.html
[9] A. B. Zhizhchenko, A. D. Izaak, Russian Math. Surveys **62** (5), 943-966 (2007).
[10] S. Kuhn, T. Helmus, R.J. Lancashire, P. Murray-Rust, H.S. Rzepa, C. Steinbeck, and E.L. Willighagen, J. Chem. Inf. Mod. **47** (6), 2015-2034 (2007).
[11] P. Murray-Rust, Journal of Cheminformatics, **3**: 48 (2011).
[12] MKM-IG. Mathematical Knowledge Management. URL: http://www.mkm-ig.org/
[13] W. Sperber, Search engines and bibliographic databases. In *A Focus on Mathematics*, Ed. by B. Wegner and Staff Unit Communications (FIZ Karlsruhe, 26-30, 2008).
[14] J. Carette, W. M. Farmer, A Review of Mathematical Knowledge Management. In *Intelligent Computer Mathematics*. Lecture Notes in Computer Science **5625**. 233-246 (2009).
[15] P. D. F. Ion, Mathematics and the World Wide Web. In *Intelligent Computer Mathematics*. Lecture Notes in Computer Science **7961**. 230-245 (2013).
[16] C. Lange, Semantic Web. **4** (2). 119-158 (2013).
[17] H. Barendregt, F. Wiedijk, Transactions A of the Royal Society, **363** (1835). 2351-2375 (2005).
[18] A. M. Elizarov, E. K. Lipachev, O. A. Nevzorova, V. D. Solov'ev. Doklady Mathematics. **90** (1). 521-524 (2014).
[19] E. V. Biryal'tsev, A. M. Elizarov, N. G. Zhil'tsov, E .K. Lipachev, O. A. Nevzorova, V.D. Solov'ev. Automatic Documentation and Mathematical Linguistics. **48** (2). 81-85 (2014).
[20] S. Parinov, M. Kogalovsky, Applied Informatics, **6** (2009). URL: http://www.ipr-ras.ru/articles/koga-pari09-2.pdf
[21] Liber Mathematicae, URL: http://math.colorado.edu/libermath/
[22] M. J. Pflaum, J. Tuley, arXiv:1102.5720
[23] Computable Document Format (CDF) for Interactive Content. URL: http://wolfram.com/cdf.
[24] Mizar Project. URL: http://mizar.org/project/.
[25] The Coq Proof Assistant. URL: http://coq.inria.fr/.
[26] T. Berners-Lee, J. Hendler, O. Lassila, Scientific american, **284** (5). 28-37 (2001).
[27] R. Ausbrooks et al., Mathematical Markup Language (MathML) Version 3.0. W3C Candidate Recommendation of 15 December 2009. World Wide Web Consortium. **13**. (2009).
[28] R. Miner, Notices of the AMS. **52**, 532-538 (2005).
[29] Word to LaTeX, LaTeX to Word Converters. URL: http://www.tex2word.com/
[30] A LaTeX to XML/HTML/MathML Converter. URL: http://dlmf.nist.gov/LaTeXML/.
[31] M. Kohlhase, *OMDoc – an open markup format for mathematical documents [Version 1.2]* (Berlin: Springer, 2006).
[32] M. Iancu, M. Kohlhase, F. Rabe, J. Urban, Journal of Automated Reasoning. **50** (2). 191-202 (2013).

[33] C. David, M. Kohlhase, C. Lange, F. Rabe, N. Zhiltsov, V. Zholudev, Proc. 7th Extended Semantic Web Conference (ESWC). 370-375 (2010). URL: http://arxiv.org/pdf/1004.3390.pdf.

[34] M. Kohlhase, STEX: Semantic Markup in TEX/LATEX (2005). URL: https://svn.kwarc.info/repos/stex/trunk/sty/stex.pdf.

[35] M. Kohlhase, Math. Comput. Sci. **2**. 279-304 (2008).

[36] F. Kamareddine, J. B. Wells, Electr. Notes Theor. Comput. Sci. **205** (C), 5-30 (2008). URL: http://www.sciencedirect.com/science/article/pii/S1571066108001680.

[37] V. Solovyev, N. Zhiltsov, Proc. of the Int. Conf. on Web Intelligence, Mining and Semantics (WIMS'11). ACM, 21:1-21:9 (2011).

[38] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A Nucleus for a Web of Open Data. In *The semantic web*. Springer Berlin Heidelberg. 722-735 (2007).

[39] K. Aberer, A. Boyarsky, P. Cudré-Mauroux, G. Demartini, O. Ruchayskiy, ScienceWISE: A Web-based Interactive Semantic Platform for Scientific Collaboration. In *10th International Semantic Web Conference* (ISWC 2011 - Demo, 2011).

[40] N. Sloane, Notices of the AMS. **50** (8). 912-915 (2003).

[41] R. Thomas, MSOR Connections. **4** (3) (2004).

[42] M. R. Genesereth, N. J. Nilsson, *Logical Foundations of Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1987).

[43] T. R. Gruber. Knowledge Acquisition. **5** (2). 199-220 (1993).

[44] R. Studer, R. Benjamins, D. Fensel, Data & Knowledge Engineering. **25** (1-2). 161-198 (1998).

[45] N. Guarino, D. Oberle, S. Staab, *What Is an Ontology*. In *Handbook on Ontologies*. Springer, 2th edition. 1-17 (2009).

[46] J. Angele, M. Kifer, G. Lausen, Ontologies in F-Logic. In *Handbook on Ontologies*. Springer, 2th edition. 45-68 (2009).

[47] F. Baader, I. Horrocks, U. Sattler, *Description Logics*. In *Handbook on Ontologies*. Springer, 2th edition. 21-43 (2009).

[48] P. Hitzler, M. Krötzsch, B. Parsia, P.F. Patel-Schneider, S. Rudolph (eds.), OWL 2 Web Ontology Language Primer, http://www.w3.org/TR/owl2-primer/.

[49] O. Nevzorova, N. Zhiltsov, A. Kirillovich, E. Lipachev. KESW 2014, CCIS 468. 105-119 (2014). URL: http://arxiv.org/abs/1407.4833.

[50] A.M. Elizarov, E.K. Lipachev, M.A. Malakhaltsev, *Web Technologies for Mathematicians: The Basics of MathML. A Practical Guide* (Moscow: Fizmatlit, 2010) (In Russian).

[51] O. Nevzorova, N. Zhiltsov, D. Zaikin, O. Zhibrik, A. Kirillovich, V. Nevzorov, E. Birialtsev, 12th Int. Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013. Proceedings, Part I. **8218**. Springer Berlin Heidelberg. 379-394 (2013).

[52] D. Velleman, *How to Prove It: A Structured Approach*, 2 ed. (Cambridge University Press, 2006).