math.wikipedia.org: A vision for a collaborative, semi-formal, language independent math(s) encyclopedia

J. Corneli and M. Schubotz

[1] Goldsmiths, University of London, London, UK
j.corneli@gold.ac.uk
[2] Universität Konstanz, Konstanz, Germany
moritz.schubotz@uni.kn

Introduction

It has been argued that the right representation is the key to successful reasoning [Bun13, p. 16]. We would broaden this observation to point out that while both representation and reasoning are crucial, they are not in themselves sufficient for most applied cognitive and computational workflows. In particular, the processes by which representations are created are logically prior. With large-scale projects, human and HCI factors must be considered, and suitable management strategies devised. Below, we consider the representation and workflow issues surrounding the creation of a large machine-readable mathematics knowledge base. Such a knowledge base would differ from a repository of human-readable texts, in that it would support a range of automated reasoning tasks. For example, one such task is automated tutoring. An ongoing master's thesis [Dud16] on a related simpler task, automated question answering, has helped to motivate the current inquiry.

1 What has been done?

As far as human-readable mathematics goes, one of the most successful and visible projects is Wikipedia, the world's largest encyclopedia. The Mathematics WikiProject indexes 31444 articles at the time of writing.[1] For some sub-fields of STEM, Wikipedia also contains a significant amount of related domain knowledge. Until recently, the main – and still most familiar – way of representing this knowledge was in encyclopedic human readable articles in different languages, which were to some degree linked to each other. With the implementation of Wikidata this has begun to change. Now, there is a centralized database that has unique IDs for each semantic concept. From those concepts, there are links to the individual language versions of the Wikipedia articles. In addition, Wikidata contains links between concepts, and properties of the items of various data types, including mathematical formulae [SS16]. In this way, Wikidata forms the backbone of a knowledge graph. The project, then, exists somewhere between two domains: *digital libraries* and *artificial intelligence*.

2 Our plan for the future

We propose to rely on Wikidata, and build "math.wikipedia.org" as a frontend to that. This service will allow users to conveniently store and retrieve formulas, terms, and other mathematical objects. Building a knowledge base with machine-readable formulas and links between

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Mathematics/Count

concepts will produce a richer domain for reasoning than we get from existing projects or other mathematical digital library proposals that lack semantic models.

Working with an existing popular platform should help bootstrap the social side, and allow us to focus our development effort on domain-specific issues. State of the art technology now exists for mining relevant knowledge from existing knowledge resources. Wikidata and a handful of other projects have proposed knowledge representation formats for mathematics. One current example, *Gamma function* on Wikidata is pictured in Figure 1. The page is quite terse. More details would be needed to bring the Wikidata page closer to feature parity with the Wikipedia pages in English or German. Indeed, better-articulated frameworks would be required to adequately represent more complex mathematical concepts, e.g., *adjunction in a category*.[2] Whatever the concept, the definition and any formulas should be accompanied by framing information (e.g., who invented it, where has the concept been discussed?).

3 How will the public benefit?

Improvements to Wikipedia: If all of this information is available within Wikidata, we and others can use it to make a better Wikipedia using semi-automated methods. For example, article placeholders, autolinking of entries, automatic provision of references, and so on, could all be provided. By using the knowledge graph, such services can be delivered with contextual sensitivity; e.g., to point out that "this article is missing a basic explanation of $X/Y/Z$.")

Reasoning and computational methods available by default: Having a computational "frame" over the material, we would be able to infer, e.g, "Here is an example; here's how you compute with it." An implementation of the key Pólya heuristic "If you can't solve a problem, there is an easier problem you can solve: find it" would be valuable for tutoring applications.

Integration with other knowledge bases and AI systems: Thinking of the system as a data consumer, we should be able to use it to index and centrally store concepts from the broader literature. Thinking of the system as a data provider, if the Wikidata representations are done right, we should be able to use this data in AI applications, for example, as a resource to use in more general question-answering systems that make use of mathematical reasoning.

4 Related work

The NIST Digital Library of Mathematical Functions [Mil13] and the derived Digital Repository of Mathematical Formulae [CMS+14, CSM+15] are inspiring examples of formula-centric storage. Contentmine is an interesting current project that has had success mining knowledge out of research papers (for now, mostly in chemistry and bioscience). It would be interesting to apply the methods to the mathematics and physics content stored in arXiv. Babar is another project for extracting knowledge out of Wikipedia, specifically [dL12]. The authors of the current paper have been highly involved with representing mathematics on the web in a format suited for human readers, notably through work with the MathML Association and PlanetMath. Representing mathematical knowledge in a format suitable for both human and machine interaction requires a somewhat different approach. Michael Kohlhase's recent work on a Semantic Multilingual Glossary for Mathematics is relevant prior art, as are the earlier OpenMath Content Dictionaries. A range of more formal representations exists in the context of projects like HOL [GK15], Isabelle, Coq [CFGW04], and Mizar, where contributions to the Mizar Mathematical Library correspond to articles in the journal *Formalized Mathematics*.

[2] https://en.wikipedia.org/wiki/Adjoint_functors

5 Conclusions

If we compare the few high-level properties of the gamma function that are expressed in Figure 1 with those that are captured in connection with the function's formalisation in HOL [SH14], it becomes even more starkly clear that the current Wikidata treatment lacks detail. In addition to listing more properties, the formal treatment naturally provides the proofs of these properties. At the moment, Wikidata even lacks the property '`demonstrates`' that could be used to connect the statement of a theorem with a proof.[3] In principle, one could be added – which would then beg the question as to how proofs would be represented in the system. However, proofs are just one of many challenging topics to address. A simple example: how should Wikidata answer the question "What is the area of a circle?" – given that 'disk' would be correct in mathematical English, but *Kreisfläche* (lit. *circle area*) is the relevant concept in German.[4]

More broadly: what role will semantic representation and AI software play in the development of the online mathematics ecosystem? What workflows will support the use and further development of the constituent online resources? This is the upstream part of the "collaborative development of open services based on representations of mathematical knowledge" [Ion16]. The fruits of repository- and KB-building efforts depend on due attention to these roots.

Acknowledgment Corneli's work was supported by an EPSRC-funded project studying "The Integration and Interaction of Multiple Mathematical Reasoning Processes" (EP/N014758/1).

References

[Bun13] Alan Bundy. The interaction of representation and reasoning. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 469(2157), 2013.

[CFGW04] Luís Cruz-Filipe, Herman Geuvers, and Freek Wiedijk. C-CoRN, the constructive Coq repository at Nijmegen. In *International Conference on Mathematical Knowledge Management*, pages 88–103. Springer, 2004.

[CMS+14] H. S. Cohl, A. McClain M, B. V. Saunders, M. Schubotz, and J. C. Williams. Digital Repository of Mathematical Formulae. In *CICM*, LNCS, 2014.

[CSM+15] H. S. Cohl, M. Schubotz, M. A. McClain, B. V. Saunders, C. Y. Zou, A. S. Mohammed, and A. A. Danoff. Growing the DRMF with Generic LaTeX Sources. In *CICM*, 2015.

[dL12] Pierre Raymond de Lacaze. BABAR: Wikipedia Knowledge Extraction, 2012.

[Dud16] Kaushal Dudhat. Teaching Mathematics to Question Answering System. Master's thesis, Universität Konstanz, 2016.

[GK15] Thibault Gauthier and Cezary Kaliszyk. Sharing HOL4 and HOL Light proof knowledge. In *Logic for Programming, Artificial Intelligence, and Reasoning*, pages 372–386. Springer, 2015.

[Ion16] Patrick Ion. The Effort to Realize a Global Digital Mathematics Library. In *International Congress on Mathematical Software*, pages 458–466. Springer, 2016.

[Mil13] Bruce R. Miller. Three years of DLMF: web, math and search. In J. Carette and D. Aspinall, editors, *CICM*, volume 7961 of *LNCS*, 2013.

[SH14] Umair Siddique and Osman Hasan. On the formalization of gamma function in HOL. *Journal of automated reasoning*, 53(4):407–429, 2014.

[SS16] M. Schubotz and A. P. Sexton. A Smooth Transition to Modern mathoid-based Math Rendering in Wikipedia with Automatic Visual Regression Testing. In M. Kohlhase, editor, *WiP at CICM*, Aachen, 2016.

[3] https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Mathematics#Modelling_proofs_and_properties_of_some_objects

[4] https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Mathematics#Consistent_naming_convention

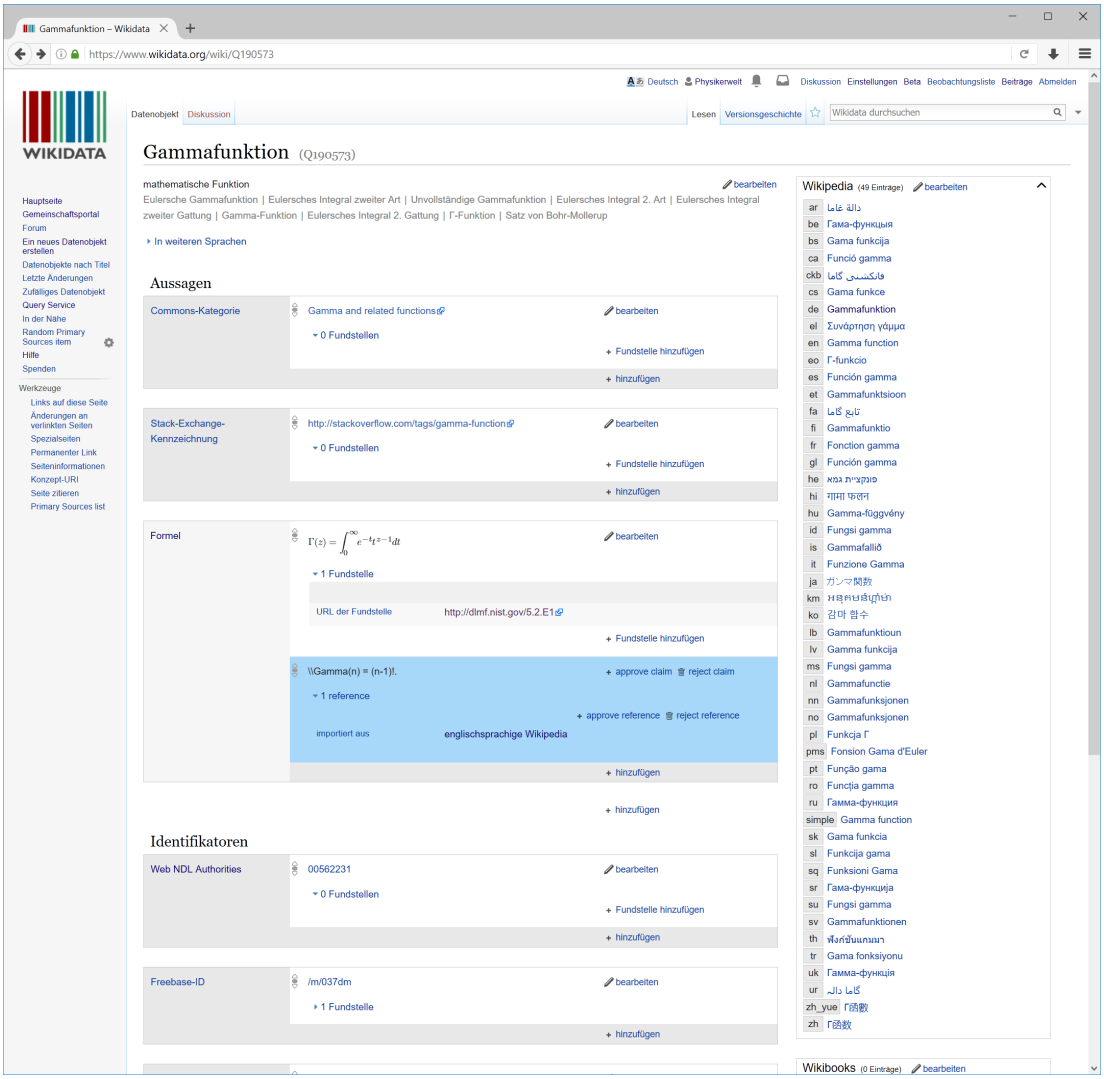

Figure 1: Screenshot of the Wikidata page for Euler's gamma function https://www.wikidata.org/wiki/Q190573, with German language localization. On the top of the page is the title and the id. Below there is a natural language description and a list of aliases. In the block on the right there are links to the various Wikipedia pages about the gamma function. One of the blocks on the left contains a definition that was manually entered, and originates from the NIST Digital Library of Mathematical Functions. The formula in the blue box, $\Gamma(n) = (n-1)!$, was automatically extracted from the English Wikipedia by Yash Nager (a master's student at Universität Konstanz), and currently waits for human verification. Links to external identifiers and other related resources are provided.