Mathematical Formulae in Wikimedia Projects 2020

Moritz Schubotz[1,2], André Greiner-Petter[1], Norman Meuschke[1,3], Olaf Teschke[2], and Bela Gipp[1,3]

[1]University of Wuppertal, Germany (andre.greiner-petter@zbmath.org, {last}@uni-wuppertal.de)
[2]FIZ-Karlsruhe, Germany ({first.last}@fiz-karlsruhe.de)
[3]University of Konstanz, Germany ({first.last}@uni-konstanz.de)

May 6, 2020

Abstract

This poster summarizes our contributions to Wikimedia's processing pipeline for mathematical formulae. We describe how we have supported the transition from rendering formulae as course-grained PNG images in 2001 to providing modern semantically enriched language-independent MathML formulae in 2020. Additionally, we describe our plans to improve the accessibility and discoverability of mathematical knowledge in Wikimedia projects further.

1 Introduction

Mathematical formulae are an integral part of Wikipedia and other projects of the Wikimedia foundation[1]. The MediaWiki software is the technical backbone of many Wikimedia projects, including Wikipedia. Since 2003, wikitext – the markup language of MediaWiki – supports mathematical content [9]. For example, MediaWiki converts the wikitext code `<math>E=mc^2</math>` to the formula $E = mc^2$. While the markup for mathematical formulae has remained stable since 2003, MediaWiki's pipeline for processing wikitext to render formulae has changed significantly.

Initially, MediaWiki used LaTeX to convert math tags in wikitext to PNG images. The rendering process was slow, the images did not integrate well into the text, were inaccessible to screen readers for visually impaired users, and scaled poorly for both small and high-resolution screens. To alleviate these problems, we started developing a new JavaScript-based rendering backend called

[1] List of Wikimedia projects: https://meta.wikimedia.org/wiki/Wikimedia_projects

Mathoid in 2013 [9]. Mathoid invokes MathJax on the server-side to convert the LaTeX code to MathML and SVG output. The new rendering pipeline became available in production in 2016 [8].

Improving the rendering of mathematical formulae was only a first step towards our ultimate goal of enhancing the discoverability of mathematical knowledge. Working towards that goal, we developed a first math search engine prototype for Wikipedia [7] in 2012. However, we found that classic, lexical search functionality for mathematical content has little practical benefit. The NTCIR MathIR competitions [1, 10], which will continue at the CLEF conference 2020, have confirmed our experience. The competitions use Wikipedia as a dataset to evaluate mathematical information retrieval systems. The NTCIR results indicate that systems employing established information retrieval technology fail to add significant value for the average Wikipedia user [11]. Deploying math search to Wikipedia requires a semantic understanding of formulae, which in turn necessitates semantic augmentation of formulae.

To increase the availability of semantic information on mathematical formulae, we implemented the rendering of formulae in Wikidata – the central structured knowledge base for Wikimedia projects. The new functionality greatly facilitates the donation of semantically annotated mathematical formulae for volunteers.

The availability of semantic formula data in Wikidata has thus far enabled several research projects, e.g., on math question answering [12], semantic augmentation of mathematical content [5], and mathematical information retrieval [6]. However, the connection between formulae in Wikidata and Wikipedia had no immediate benefit for the average Wikipedia user until January 2020.

2 Enhanced formulae in Wikipedia

In January 2020, we deployed a feature that enables enhancing mathematical formulae in Wikipedia with semantics from Wikidata. For instance, the wikitext code `<math qid=Q35875>E=mc^2</math>` now connects the formula $E = mc^2$ to the corresponding Wikidata item by creating a hyperlink from the formula to the special page shown in Figure 1. The special page displays the formulae together with its name, description, and type, which the page fetches from Wikidata. This information is available for most formulae in all languages. Moreover, the page displays elements of the formula modeled as `has part` annotations of the Wikidata item.

The `has part` annotation is not limited to individual identifiers but also applicable to complex terms, such as $\frac{1}{2}m_0v^2$, i.e., the kinetic energy approximation for slow velocities. For example, we demonstrated using the annotation for the Grothendieck–Riemann–Roch theorem $\mathrm{ch}(f_!\mathcal{F}^\bullet)\mathrm{td}(Y) = f_*(\mathrm{ch}(\mathcal{F}^\bullet)\mathrm{td}(X))$. The smooth quasi-projective schemes X and Y in the theorem lack Wikipedia articles. However, dedicated articles on *quasi-projective variet*y and *smooth scheme* exist. We proposed modeling this situation by creating the new Wikidata item *smooth quasi-projective scheme*, which links to the existing articles as subclasses. To create a clickable link from the Wikidata item to Wikipedia, we could create a new Wikipedia article on *smooth quasi-projective scheme.* Alternatively, we could add a new section on *smooth quasi-projective scheme* to the article on *quasi-projective variety* and create a redirect from the Wikidata item to the new section.

Aside from implementing the new feature, defining a decision-making process for the integration of math rendering features into Wikipedia was equally important. For this purpose, we founded the Wikimedia Community Group Math as an international steering committee with authority to decide on future features of the math rendering component of Wikipedia.

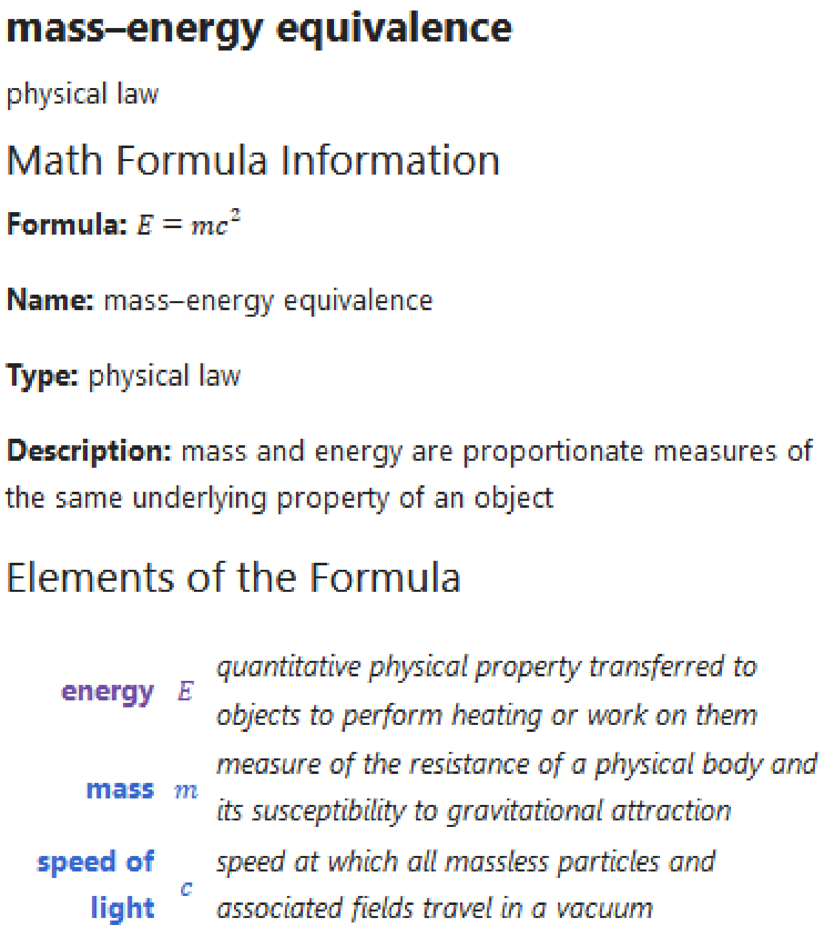
3 Conclusion & Future Work

After working on Wikipedia's math support for several years, we have deployed the first feature that goes beyond improving the display of formulae. Realizing the feature became possible through the inauguration of the Wikimedia Community Group Math.

The new feature helps Wikipedia users to better understand the meaning of mathematical formulae by providing details on the elements of formulae. Because the new feature is available in all language editions of Wikipedia, all users benefit from the improvement. Rolling out the feature for all languages was important to us because using Wikipedia for more in-depth investigations is significantly more prevalent in languages other than English [3]. Nevertheless, also in the English Wikipedia, fewer than one percent of the articles have a quality rating of good or higher [4]. Providing better tool support to editors can help in raising the quality of articles. In that regard, our semantic enhancements of mathematical formulae will flank other semi-automated methods, such as recommending sections [4] and related articles [14].

To stimulate the wide-spread adoption of semantic annotations for mathematical formulae, we are currently working on tools that support editors in creating the annotations. With `AnnoMathTex` [6], we are developing a tool that facilitates annotating mathematical formulae by providing a graphical user interface that includes machine learning assisted suggestions [2] for annotations. Moreover, we will integrate a field into the visual wikitext editor that will suggest Wikipedia authors to link the Wikidata id of a formula if the formula is in the Wikidata database. Improved tool support will particularly enable smaller language editions of Wikipedia to benefit from the new feature because the annotations performed in any language will be available in all languages automatically.

Acknowledgments We thank the Wikimedia Foundation, particularly Marco Obrovac, for their continuous support. Our research was supported by the German Research Foundation DFG GI 1259/1.



mass–energy equivalence

physical law

Math Formula Information

Formula: $E = mc^2$

Name: mass–energy equivalence

Type: physical law

Description: mass and energy are proportionate measures of the same underlying property of an object

Elements of the Formula

energy	E	*quantitative physical property transferred to objects to perform heating or work on them*
mass	m	*measure of the resistance of a physical body and its susceptibility to gravitational attraction*
speed of light	c	*speed at which all massless particles and associated fields travel in a vacuum*

Figure 1: Semantic enhancement of the formula $E = mc^2$. `https://en.wikipedia.org/wiki/Special:MathWikibase?qid=Q35875`

References

[1] A. Aizawa et al. “NTCIR-11 Math-2 Task Overview”. In: *Proc. NTCIR.* 2014, pp. 88–98.

[2] A. Greiner-Petter et al. “Discovering Mathematical Objects of Interest—A Study of Mathematical Notations”. In: *Proc. WWW.* 2020, pp. 1445–1456. DOI: 10.1145/3366423.3380218.

[3] F. Lemmerich et al. “Why the World Reads Wikipedia: Beyond English Speakers”. In: *Proc. ACM WSDM.* 2019, pp. 618–626. DOI: 10.1145/3289600.3291021.

[4] T. Piccardi et al. “Structuring Wikipedia Articles with Section Recommendations”. In: *Proc. ACM SIGIR.* 2018, pp. 665–674. DOI: 10.1145/3209978.3209984.

[5] P. Scharpf, M. Schubotz, and B. Gipp. “Representing Mathematical Formulae in Content MathML using Wikidata”. In: *Proc. ACM SIGIR.* Vol. 2132. 2018, pp. 46–59.

[6] P. Scharpf et al. “*AnnoMathTeX* - A Formula Identifier Annotation Recommender System for STEM Documents”. In: *Proc. ACM RecSys.* 2019, pp. 532–533. DOI: 10.1145/3298689.3347042.

[7] M. Schubotz. “Making Math Searchable in Wikipedia”. In: *CoRR* abs/1304.5475 (July 30, 2012). DOI: 10.14279/depositonce-5034. arXiv: 1304.5475.

[8] M. Schubotz and A. P. Sexton. “A Smooth Transition to Modern mathoid-based Math Rendering in Wikipedia with Automatic Visual Regression Testing”. In: *WiP at CICM.* Vol. 1785. 2016, pp. 132–145.

[9] M. Schubotz and G. Wicke. “Mathoid: Robust, Scalable, Fast and Accessible Math Rendering for Wikipedia”. In: *Proc. CICM.* Vol. 8543. 2014, pp. 224–235. DOI: 10.1007/978-3-319-08434-3_17.

[10] M. Schubotz et al. “Challenges of Mathematical Information Retrieval in the NTCIR-11 Math Wikipedia Task”. In: *Proc. ACM SIGIR.* 2015, pp. 951–954. DOI: 10.1145/2766462.2767787.

[11] M. Schubotz et al. “Exploring the One-Brain-Barrier: a Manual Contribution to the NTCIR -12 Math Task”. In: *Exploring the One-Brain-Barrier. Proc. NTCIR.* 2016.

[12] M. Schubotz et al. “Introducing MathQA - A Math-Aware Question Answering System”. In: *Proc. ACM/IEEE JCDL.* June 2018. DOI: 10.1108/IDD-06-2018-0022.

[14] M. Schwarzer et al. “Evaluating Link-based Recommendations for Wikipedia”. In: *Proc. ACM/IEEE JCDL.* 2016, pp. 191–200. DOI: 10.1145/2910896.2910908.

Listing 1: Use the following `BibTeX` code to cite this article

```
@inproceedings{Schubotz2020,
  author    = {Schubotz, Moritz and Greiner-Petter, Andr\'{e
      } and Meuschke, Norman and Teschke, Olaf and Gipp, Bela
      },
  booktitle = {Proceedings of the ACM/IEEE Joint Conference
      on Digital Libraries in 2020 (JCDL '20), August 1--5,
      2020, Virtual Event, China},
  date      = {2020},
  doi       = {10.1145/3383583.3398557},
  preprint  = {https://arxiv.org/pdf/2003.09417},
  title     = {Mathematical Formulae in Wikimedia Projects
      2020},
}
```