# Entity linking for zbMATH Open

Marcel Fuhrmann, Philipp Scharpf and Moritz Schubotz

## 1 Introduction

As scientific knowledge expands, specialization has become increasingly prominent, with researchers focusing on narrower fields to deepen understanding and foster innovation. While specialization enables experts to tackle complex problems precisely, it also challenges interdisciplinary collaboration, as breakthroughs often require integrating insights across diverse fields.

Even in mathematics alone, no one can be an expert in all specialized fields, each with its own ecosystem of terminologies. Every once in a while, a reader stumbles across a term or phrase in a document or abstract unfamiliar to them, or at least the specific details are not clearly stated.

zbMATH Open is an information service for mathematicians in research and teaching, with networked information on mathematical topics, authors, publications, references, and software. It provides detailed information on mathematical publications dating back to 1868. It provides access to around five million bibliographic entries with reviews or abstracts from more than five thousand journals and book series and some 200,000 books.

Therefore, providing context to all mathematical terms and phrases, proofs, lemmas, theorems, and concepts in these texts cannot be accomplished by augmenting every abstract on zbMATH Open manually. Thus, we decided on an unsupervised machine-learning algorithm.

To further support the disambiguation of mathematical terms, we are currently developing a new service to augment any abstract of a document published on zbMATH Open so that it includes links to the context of the corresponding mathematical entity.

## 2 What is the goal?

Generally speaking, entity linking means associating the main entity types, i.e., persons, organizations, locations, dates, and times, to a representation in a knowledge base or knowledge graph [1].

Entity linking can be challenging due to entity name variations or ambiguity. Several types exist, e.g., a single concept that can be called by multiple names (synonymous) or a single name that can mean multiple concepts (polysemy) [2]. This requires entity linking to utilize the entity's context or additional information to disambiguate.

Mathematical entity linking can be employed to ground mathematical entities in documents semantically. Wikidata[1] can help to achieve this by storing and linking both the concept name (with a persistent identifier called QID) and the corresponding Wikipedia page (or, if applicable, a mathematical formula) [5].

QIDs are unique and persistent identifiers used in Wikidata with corresponding concept item pages (often also linked to a set of Wikipedia articles in different languages) to refer to a specific, unique concept.

An example of how an unsupervised machine learning algorithm would augment the abstract of a document shown on zbMATH Open with mathematical entity linking can be viewed in Figure 1. On the left side, the zbMATH document page is displayed, and links to the corresponding Wikipedia article are displayed on the right side.

## 3 Applied methodology for implementing mathematical entity linking

In this section, we explain how the algorithm of our machine-learning algorithm works.

Following the ideas of [3, 4], two steps are required for formula concept discovery and recognition: First, we need to define a list of entities, and second, we need to identify mentions of those entities. For the list of entities, we used mathematical concepts from Wikidata. We collected QIDs related to topics in mathematics from the Wikidata API.[2] We created our own database consisting of a list of tuples, each containing a mathematical term and its corresponding Wikidata page.

However, the content of the items stored in Wikidata may change over time. Entries get corrected, and others may be removed to minimize redundancies or other reasons. Thus, it is

[1] https://www.wikidata.org

[2] https://www.wikidata.org/w/api.php

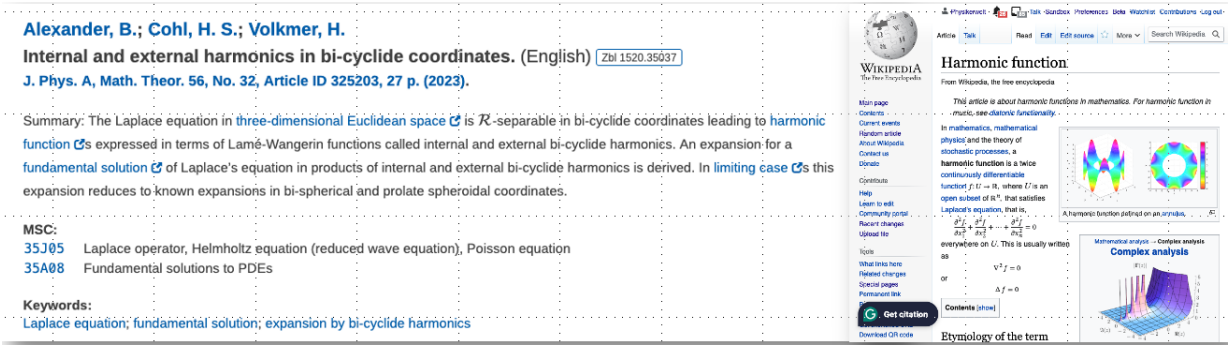


*Figure 1.* Example preview of the new entity linking function in zbMATH Open

important to update this augmentation database regularly to ensure its sustainability. For this purpose, an automated update system was created.

Another challenge was the creation of a robust algorithm for concept recognition that searches any given text for phrases that match the respective terms and replaces the simple text with underlying links to the corresponding Wikidata webpage or Wikipedia link if existing. This algorithm first removes any punctuation marks and removes latex code if existing. Then, natural language processing searches the text in question and identifies mathematical expressions in their variations and ambiguities, then predicts the most probable candidate based on the context.

This is going to be applied to any abstracts that are available to zbMATH Open, provided their respective licence agreements allow for it. The new service is expected to be available in Q4 2024.

## 4 Outlook and conclusion

To expand the usage of this technology, we plan on introducing a similar approach for mathematical formulas appearing in abstracts that correspond to a Wikidata entry.

This project is quite more complex, as each formula needs to be analysed and extracted from the text using open-source code libraries for symbolic mathematics to classify the identifiers and operators in their mathematical context correctly.

Currently, we are investigating different approaches. Each of those methods consists of analysing variable symbols and operators. They differ in how the extracted symbols are interpreted:

- we construct a knowledge graph query to find similar formulas and retrieve the best fitting candidate,
- extracts and categorizes individual parts from a formula string into identifiers and operators,
- prompting the formula retrieval via an open source large language model.

In addition, the concepts are currently used for display only on zbMATH Open. In the future, we plan to use the concepts for navigation and search. For example, they can be used to classify articles that are more fine-grained than currently possible with the mathematical subject classification. In addition, they could be used to filter search results by concept.

To summarize, the new natural language processing will be useful for the automated augmentation of the abstract texts stored within the zbMATH Open database. It will extend the readability and findability to all publications, and therefore, open access for zbMATH Open data will continue to expand.

*References*


[1] A. Goyal, V. Gupta and M. Kumar, Recent named entity recognition and classification techniques: A systematic review. *Comput. Sci. Rev.* **29** 21–43 (2018)

[2] Z.-Y. Ming and T. S. Chua, Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling. *Inf. Sci.* **307**, 18–38 (2015)

[3] Ph. Scharpf, M. Schubotz, H. S. Cohl, C. Breitinger and B. Gipp, Discovery and recognition of formula concepts using machine learning. *Scientometrics* **128**, 4971–5025 (2023)

[4] Ph. Scharpf, M. Schubotz and B. Gipp, Towards explaining STEM document classification using mathematical entity linking. arXiv: 2109.00954v1 (2021)

[5] Ph. Scharpf, M. Schubotz, A. Spitz, N. Meuschke, A. Greiner-Petter and B. Gipp, Entity linking with Wikidata: A systematic literature review. *ACM Comput. Surv.*, submitted (2024)



Marcel Fuhrmann studied physics at Humboldt-Universität zu Berlin. He completed his PhD in astrophysics and worked as a postdoctoral researcher in applied mathematics at the University of Potsdam. From 2017 to 2021 he worked at the German Federal Institute for Risk Assessment. In 2021, he moved to FIZ Karlsruhe and worked on the zbMATH Open REST API and other related projects.

marcel.fuhrmann@fiz-karlsruhe.de

Philipp Scharpf studied physics at the ETH Zurich, the University of Zurich, and the University of Konstanz and completed his doctorate in computer science in the fields of information retrieval and machine learning at the University of Konstanz and the University of Göttingen. Since 2022, he has been a freelance consultant for data and AI solutions and a lecturer at the University of Stuttgart for big data and learning analytics.

scharpf@gipplab.org

Moritz Schubotz studied physics and computer science at the University of Wuppertal and TU Berlin. After his PhD, he was a postdoc at the University of Konstanz and the National Institute of Informatics in Tokyo before joining FIZ Karlsruhe in 2019. At FIZ Karlsruhe, he researches methods to improve zbMATH Open and related research infrastructure.

moritz.schubotz@fiz-karlsruhe.de