Final Report for the DFG-Project „MathIR“

Methods and Tools to Advance the Retrieval of Mathematical Knowledge from Digital Libraries for Search-, Recommendation- and Assistance-Systems

Bela Gipp, André Greiner-Petter, Moritz Schubotz, Norman Meuschke

March 2023

Reference Number: GI 1259-1

Project Number: 350192710

Applicant

Prof. Dr. Bela Gipp
University of Göttingen
Dept. of Computer Science
Scientific Information Analytics Group
Papendiek 14, 37073 Göttingen

German Project Title

Methoden und Werkzeuge zur Verbesserung des Zugriffs auf
mathematisches Wissen in Digitalen Bibliotheken für
Such-, Empfehlungs- und Assistenzsysteme

Reporting Period

July 1st, 2018 – December 31st, 2022

Abstract

This project investigated new approaches and technologies to enhance the accessibility of mathematical content and its semantic information for a broad range of information retrieval applications. To achieve this goal, the project addressed three main research challenges: (1) syntactic analysis of mathematical expressions, (2) semantic enrichment of mathematical expressions, and (3) evaluation using quality metrics and demonstrators. To make our research useful for the research community, we published tools that enable researchers to process mathematical expressions more effectively and efficiently.

The project has made significant research contributions to various Mathematical Information Retrieval (MathIR) tasks and systems, including plagiarism detection and recommendation systems, search engines, the first mathematical type assistance system, math question answering and tutoring systems, automatic plausibility checks for mathematical expressions in Wikipedia, automatic computability of mathematical content via Computer Algebra Systems (CAS), and others. Although our project focused on MathIR tasks, its impact on other natural language research was significant, leading to a more extensive range of demonstrators than originally expected. Many of these demonstrators introduced novel applications, such as the tutoring system PhysWikiQuiz [26] or LaCASt [2], which automatically verifies the correctness of math formulae in Wikipedia or the Digital Library of Mathematical Functions (DLMF) via commercial CAS.

During the project, we published 29 peer-reviewed articles in international venues, including prestigious conferences like the *Joint Conference on Digital Libraries (JCDL)* [8, 3, 6, 29, 10, 30] and *The Web Conference (WWW)* [4, 22] (CORE rank A*), as well as journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* [1] (IF: 24.314) and *Scientometrics* [5] (IF: 3.801). Our Wikipedia demonstrator was also featured in public media. Furthermore, we actively presented our contributions, especially demonstrators, to the research community in multiple workshops [14, 20, 21, 24, 23, 30, 11, 26].

This project has strengthened our international collaborations, particularly with colleagues at the National Institute of Standards and Technology (NIST) in the US and the National Institute of Informatics (NII) in Japan. Several sub-projects were partially developed in course projects and theses at the Universities of Konstanz, Wuppertal, and Göttingen, exposing junior researchers to cutting-edge technologies and sensitizing students and researchers to the outstanding issues in MathIR technologies. We firmly believe that this project has a lasting effect on following MathIR technologies. Several of the sub-projects initiated as part of this grant are ongoing and motivating follow-up DFG projects, such as *Analyzing Mathematics to Detect Disguised Academic Plagiarism* (project no. 437179652).

Most Influential Project Publications

Publications with Scientific Quality Assurance

[1] A. Greiner-Petter, M. Schubotz, C. Breitinger, P. Scharpf, A. Aizawa, and B. Gipp. “Do the Math: Making Mathematics in Wikipedia Computable”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (Aug. 2022), pp. 4384–4395. ISSN: 0162-8828. DOI: `10.1109/TPAMI.2022.3195261`.

[2] A. Greiner-Petter, H. S. Cohl, A. Youssef, et al. “Comparative Verification of the Digital Library of Mathematical Functions and Computer Algebra Systems”. In: *Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, (TACAS)*. Springer, Apr. 2022, pp. 87–105. DOI: `10.1007/978-3-030-99524-9_5`.

[3] P. Scharpf, M. Schubotz, and B. Gipp. “Mining Mathematical Documents for Question Answering via Unsupervised Formula Labeling”. In: *Proc. ACM/IEEE JCDL*. ACM, June 2022, pp. 1–11. DOI: `10.1145/3529372.3530925`.

[4] A. Greiner-Petter, M. Schubotz, F. Müller, et al. “Discovering Mathematical Objects of Interest — A Study of Mathematical Notations”. In: *Proc. WWW*. ACM, Apr. 2020, pp. 1445–1456. DOI: `10.1145/3366423.3380218`.

[5] A. Greiner-Petter, A. Youssef, T. Ruas, et al. “Math-Word Embedding in Math Search and Semantic Extraction”. In: *Scientometrics* 125.3 (Dec. 2020), pp. 3017–3046. ISSN: 0138-9130. DOI: `10.1007/s11192-020-03502-9`.

[6] P. Scharpf, M. Schubotz, A. Youssef, F. Hamborg, N. Meuschke, and B. Gipp. “Classification and Clustering of arXiv Documents, Sections, and Abstracts, Comparing Encodings of Natural and Mathematical Language”. In: *Proc. ACM/IEEE JCDL*. ACM, Aug. 2020, pp. 137–146. DOI: `10.1145/3383583.3398529`.

[7] A. Greiner-Petter, M. Schubotz, H. S. Cohl, and B. Gipp. “Semantic Preserving Bijective Mappings for Expressions Involving Special Functions between Computer Algebra Systems and Document Preparation Systems”. In: *Aslib Journal of Information Management* 71.3 (May 2019), pp. 415–439. ISSN: 2050-3806. DOI: `10.1108/AJIM-08-2018-0185`.

[8] N. Meuschke, V. Stange, M. Schubotz, M. Kramer, and B. Gipp. “Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations”. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, June 2019, pp. 120–129. DOI: `10.1109/jcdl.2019.00026`.

[9] P. Scharpf, I. Mackerracher, M. Schubotz, J. Beel, C. Breitinger, and B. Gipp. “AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Sept. 2019, pp. 532–533. DOI: `10.1145/3298689.3347042`.

[10] M. Schubotz, A. Greiner-Petter, P. Scharpf, N. Meuschke, H. S. Cohl, and B. Gipp. “Improving the Representation and Conversion of Mathematical Formulae by Considering Their Textual Context”. In: *Proc. ACM/IEEE JCDL*. ACM, May 2018, pp. 233–242. DOI: `10.1145/3197026.3197058`.

A complete list of publications resulting from this DFG project is available at
`gipp.com/pub/#mathir`.

Contents

1 Progress Report

1.1 Background and Project Objectives

Information Retrieval (IR) systems have become indispensable for finding relevant information despite today's ubiquitous information overload. Web search engines like Google and Bing, and the recommender systems embedded within Amazon and Netflix are prominent examples of the many domain-specific systems for indexing and accessing content we use daily. In academia, finding relevant literature is a vital task in all research disciplines but the exponential growth in the number of publications makes it increasingly challenging. Consequently, IR systems also play a crucial role in facilitating access to scholarly literature.

Despite IR systems' importance in accessing scholarly literature and their dependence on accessible information, much essential data remains unused, particularly non-textual data, such as images, audio and video data, and mathematical content. Recent research has focused on improving the accessibility of information in non-textual data, but mathematical content has been largely neglected. However, researchers, especially in STEM[1] fields, often communicate critical information via mathematical expressions. At best, ignoring mathematical content can be confusing. More likely, it renders the content useless.

Existing IR systems for mathematical content rely on visual resemblance and rarely access semantic information on mathematical expressions. Therefore, the goal of this DFG project was to research new approaches and technologies to automatically make semantic information in mathematical expressions accessible for a wide range of IR applications.

To achieve this goal, we addressed three major issues: (1) the syntactic analysis of mathematical expressions, (2) the semantic enrichment of mathematical expressions, and (3) the evaluation using quality metrics and demonstrators. To make our research accessible and useful for the research community, we published tools that enable researchers to process mathematical expressions more effectively and efficiently. This approach was inspired by well-established tools for natural language processing tasks, such as Stanford's Natural Language Processing Toolkit. We aimed to provide similarly flexible processing engines for mathematical expressions.

1.2 Project Results

Hereafter, we explain the projects we completed as part of the grant. As most of these projects have contributed towards multiple research objectives, we have organized the report into project-centered descriptions to provide a better understanding of the overall progress and results. Each section of the report briefly summarizes how the project has contributed towards the grant's overall goal, followed by a summary of the project and its results. For a more detailed overview of the project's contribution towards the research objectives, please refer to Appendix 1.7. Most of the projects have implemented their own demonstrators. The following list shows all the demonstrators that have been developed as part of this grant:

- PhysWikiQuiz [26]: A question-answering system for physics-related questions.

[1] Science, Technology, Engineering, and Mathematics.

- MathQA [3, 30]: A question-answering system for math-related questions (laid the foundation for PhysWikiQuiz).
- AnnoMathTeX [22, 9]: A tool to annotate mathematical LaTeX expressions with mathematical concepts.
- MathMLben [10]: A benchmark for generated MathML data.
- MathMLTools [16]: A development toolset that improves handling MathML data in Java and provides numerous interfaces for typical tasks, such as conversions, similarity calculations, and data compliance tests.
- LaCASt [12, 7, 2, 1]: A conversion tool that translates mathematical LaTeX into the syntax of Computer Algebra Systems.
- Mediawiki Extensions [29]: Tools that enhance the semantics of mathematical expressions in Mediawiki applications, such as Wikipedia.
- DLMF/DRMF [2]: The translations of the LaCASt project shall be added to the DLMF/DRMF; the associated evaluations helped to detect mathematical errors in the DLMF.

1.2.1 Fundamental MathIR Contributions

This section focuses on projects that addressed fundamental problems in MathIR systems. An issue that most mathematical data handling systems exhibit is that they either consider individual symbols or the entire expression but neglect significant and meaningful subexpressions. Compound mathematical expressions contain vital semantic information that is lost when logical components, such as function calls, arguments, parameter structures, and arithmetic logic, are ignored. The composition of mathematical formulae has received little attention from the NLP community, mainly because identifying meaningful subexpressions is context-dependent.

Another fundamental issue is the lack of a ground truth [11] for components of mathematical expressions which hampers the development of machine learning applications in the MathIR community. A major issue for creating such a large annotated dataset is the lack of a unified standard and the open question if such a ground truth can actually exist. Our work addressed both fundamental issues by studying mathematical objects of interest and developing the AnnoMathTex and MathMLben tools to further develop standardized annotations.

Mathematical Objects of Interest (MOI) Mathematical Information Retrieval (MathIR) systems face a common issue—they cannot differentiate between important mathematical expressions, such as functions, and less important or replaceable information, such as variables. In contrast, natural language processing approaches use stop word removal to focus on semantically essential text parts. An equivalent method does not exist for mathematical expressions. To discover the mathematical objects of interest (MOI) [4], we analyzed mathematical notations in the two largest scientific datasets of mathematical documents: arXiv and zbMATH Open. Our analysis of 2.5 billion mathematical expressions revealed that mathematical notations follow a frequency distribution pattern similar to that of words in natural languages.

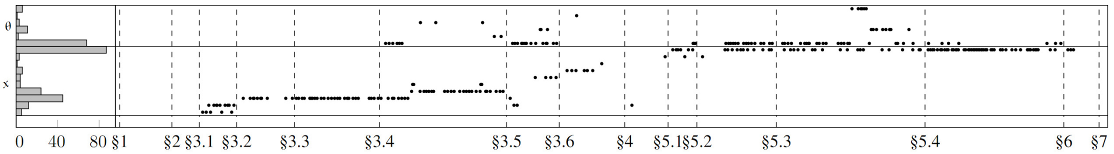


Figure 1: Different semantic annotations of θ and x within the same article.

The discovery of this pattern allowed us to use similar techniques as in natural languages to measure the significance of mathematical expressions and remove potential mathematical *stop words*. One of the most well-known schemes to measure significance based on frequency distributions is the term frequency-inverse document frequency measure (TF-IDF). By applying identical calculations to mathematical expressions, we created a novel context-sensitive math search system, enabled the first assistance system for mathematical inputs, and provided useful information for plagiarism detection and scientific recommender systems.

Grounding of Formulae The MathIR community often neglects that the semantics of mathematical notations can change rather frequently, even within the same document. While authors can redefine variables, constants, and functions at their discretion (e.g., π), little is known about the extent of this phenomenon. Mathematical language follows certain notation styles and rules that have developed over thousands of years. We expected that mixing semantics for identical expressions would be rare, at least within qualitative research articles.

To test this assumption and ground formulae in mathematical objects, we manually annotated mathematical objects with mathematical concepts derived from the context of the objects [11]. Figure 1 illustrates how the mathematical concepts for θ and x change multiple times within one article, and sometimes even change back and forth within a single subsection. Although one might expect such erratic semantic changes for heavily overloaded identifiers like x, we also observed this behavior for rather specific, less frequent identifiers like θ.

For instance, the connection between Euler numbers and Euler polynomials demonstrates how E_n can change its semantic meaning within the scope of a single formula:

$$E_n = 2^n E_n\left(\frac{1}{2}\right) \tag{1}$$

Our study [11] showed that the scope of mathematical semantic information can be very narrow and change back and forth within subsections, paragraphs, or other context windows.

AnnoMathTex The lack of high-quality annotated datasets for mathematical literature is a major challenge for MathIR. To address this issue, we developed AnnoMathTex [22, 9], a system that provides artificial intelligence (AI) guidance to improve the workflow of curating annotated datasets for mathematical literature. AnnoMathTex can recommend annotations for Wikipedia articles, drawing from arXiv, Wikipedia, Wikidata, the text surrounding the formula to be annotated, and previous user-made annotations. The system further expands these sources through fuzzy searches and linked Wikidata properties. To address the unknown scope for annotation, AnnoMathTex distinguishes global from local annotations. Our evaluation of AnnoMathTex demonstrated that the system accelerates the process of manually annotating a dataset by a factor of 1.4 for entire formulae and 2.4 for single identifiers. Moreover, we employed the system to automatically annotate Wikipedia articles and update properties in Wikidata. Remarkably, 80% of these changes were accepted in Wikipedia and 67% in Wikidata. Overall, AnnoMathTex shows great potential to significantly advance the field of MathIR by accelerating the process of annotating and updating Wikidata properties.

MathMLben The MathMLben project [10] addressed the lack of standard datasets for well-formatted and semantically enhanced content MathML. This benchmark contained 305 formulae taken from English Wikipedia articles, which we semantically enriched manually to obtain error-free and accurate MathML data. As a follow-up, we evaluated state-of-the-art LaTeX to MathML conversion tools to verify their accuracy and practicality in generating MathML. This evaluation was a crucial step forward for all MathIR-related tasks because MathML is hardly ever written manually, due to its convoluted XML structure. Typically, conversion tools are employed to generate MathML from LaTeX or image sources. However, we found that none of the tools achieved sufficient accuracy, and only three were able to generate content MathML, while all others exclusively provided presentation MathML with varying accuracy.

Our analysis identified the lack of context awareness as the main weakness of all conversion tools. Mathematical expressions are highly context-dependent and require substantial fundamental knowledge, often referred to as common knowledge. For example, the simple expression $\pi(x+y)$ may represent a multiplication between the mathematical constant π and $x+y$. However, the formula could also occur in a number theory context and discuss the number of primes, in which case π more likely refers to the prime counting function $\pi(n)$. Without context analysis, no tool could distinguish one from the other, although the difference is crucial.

The MathMLben dataset was designed to be extendable and has been used for follow-up projects. Most noteworthy are the extensions for Wikimedia [21, 23, 22, 26].

1.2.2 LaCASt

The LaTeX to Computer Algebra Systems translator (LaCASt) [12, 7, 2, 1] was a major project of this DFG grant. LaCASt is the first context-sensitive translator that analyzes the structure of mathematical LaTeX inputs and considers the textual context of a formula to disambiguate mathematical expressions. This translation process requires solutions to all objectives of the DFG grant, i.e., a syntactic analysis of mathematical expressions, a semantic enrichment pipeline, and novel quality metrics and evaluation techniques.

To achieve this, LaCASt first builds a dependency graph for mathematical notations within a document to link relevant document sections and retrieve semantic information about specific formulae. For example, let us assume the formula $\pi(x+y)$ has previously been introduced in the document as the prime counting function. The created link in the graph allows LaCASt to retrieve the relevant information and disambiguate π. The next step is to annotate nodes in the graph with textual descriptions surrounding the formula. These textual annotations are used to retrieve standard notations from the Digital Library of Mathematical Functions (DLMF). Many well-known formulae, such as the prime counting function $\pi(n)$, have standardized notations. The standard notation is essential to link the mathematical concept, here 'prime counting function', to the relevant subexpression, such as $\pi(\cdot)$. The standard notations from the DLMF helped to identify the logical syntax of an expression and annotate the syntactic elements with semantic information retrieved from the context.

LaCASt then tries to replace the general LaTeX expressions, such as `\pi(n)`, with semantically enhanced LaTeX taken from the DLMF, here `\nprimes@{n}` by considering the gathered information. After the disambiguation, the expression can be mapped to the syntax of a Computer Algebra System (CAS). For example, a translation to Mathematica would map `\nprimes@{n}` to `PrimePi[n]`. In our study, the only option to determine if a translation was correct, was to consult an expert for both the mathematical formulae and the CAS. To evaluate LaCASt, we created a dataset of 95 formulae, which we randomly selected from Wikipedia articles and manually translated to two CAS syntaxes (Maple and Mathematica) with the help of an expert. LaCASt correctly translated 27% of the formulae. Notably, the expert could only translate 81% of the formulae, which underscores the task's complexity and indicates a possible upper bound. A more comprehensive common knowledge dataset could have improved LaCASt's performance by 20%. Nonetheless, LaCASt outperformed the state-of-the-art baseline—Mathematica's LaTeX import function—which only translated 9% of the formulae correctly.

Another option to check the correctness of translations, which we evaluated in our study, is to compute the translated formula in the target Computer Algebra System and analyze the results. The equations should originate from a reliable source that established the equation's correctness—we used the DLMF. If the translated equation is invalid, there are three possible explanations: (1) the source equation was incorrect, (2) the translation was incorrect, or (3) the CAS exhibits a bug. Given that the source of the equation and the target CAS (in our study Mathematica or Maple) are highly reliable, reasons (1) and (3) are unlikely.

LaCASt could translate 72% of the DLMF to the two CAS and evaluated 48% of the 4,713 translated equations successfully. The other 52% of the equations could not be evaluated due to missing semantic information, such as branch cut positions, domains, and other constraints. Thus, a failed verification still required a manual investigation. However, we can conclude that the translations performed by LaCASt are reliable. Interestingly, the approach could detect errors in the DLMF and the two major CAS, Maple and Mathematica.

Recently, we demonstrated the potential of LaCASt by verifying Wikipedia edits. Although LaCASt is not yet productive, a demo page showcasing LaCASt's capabilities is available at `tpami.wmflabs.org`.

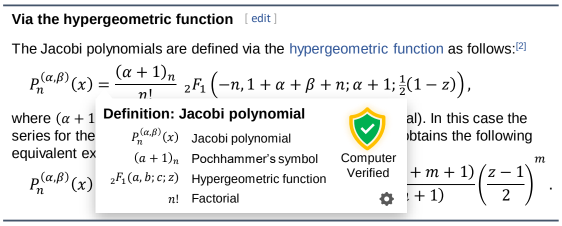
1.2.3 DLMF & DRMF

The LaCASt project also contributed to improving the Digital Library of Mathematical Functions and the Digital Repository of Mathematical Formulae (DRMF). The evaluation approach of computing translated formulae in the target CAS (see above) allowed the detection of erroneous mathematical content within the prestigious DLMF. The various identified errors included missing semantic information, incorrect links, sign errors, and others.

The DLMF is manually written using semantic LaTeX macros, which enables LaCASt to achieve optimal performance as the typically uncertain disambiguation of formulae can be skipped in most cases. This benefit of the DLMF allows for reliable translations of formulae to widely used CAS such as Mathematica and Maple, which can be very beneficial for other researchers.

We are currently discussing the integration of LaCASt into the DLMF or the expansion of the DRMF with the National Institute of Standards and Technology (NIST). A demo page showcasing all translations and evaluation results for the DLMF is available at `lacast.wmflabs.org`.

1.2.4 Wikipedia and Wikidata Extensions & Projects



Via the hypergeometric function [edit]

The Jacobi polynomials are defined via the hypergeometric function as follows:[2]

$$P_n^{(\alpha,\beta)}(x) = \frac{(\alpha+1)_n}{n!}\,{}_2F_1\left(-n, 1+\alpha+\beta+n; \alpha+1; \tfrac{1}{2}(1-z)\right),$$

where $(\alpha + 1$ … al). In this case the series for the … btains the following equivalent ex …

$$P_n^{(\alpha,\beta)}(x) \quad \ldots + m + 1) \ldots + 1) \left(\frac{z-1}{2}\right)^m .$$

Definition: Jacobi polynomial

$P_n^{(\alpha,\beta)}(x)$	Jacobi polynomial
$(a+1)_n$	Pochhammer's symbol
${}_2F_1(a,b;c;z)$	Hypergeometric function
$n!$	Factorial

Computer Verified

Figure 2: Mathematical semantic annotation in Wikipedia.

We focused on creating demonstrators primarily for two systems within the Mediawiki platform—the encyclopedia Wikipedia and the knowledgebase Wikidata. For Wikipedia, we realized several extensions that enable providing semantically enhanced mathematical content as part of the encyclopedia [29]. One of the extensions enables users to annotate mathematical formulae in Wikipedia articles with items from Wikidata to provide additional semantic information. For example, the famous formula $E = mc^2$ in the English Wikipedia article on mass-energy equivalence is now linked with the Wikidata item `Q35875`. This linking enables providing additional information on the formula to the end-user directly within the Wikipedia article. When someone hovers over an annotated mathematical formula, a popup appears with information such as the name, a description, and a list of elements and their meaning. In the future, we aim to extend these popups to display verified equations, e.g., checked by LaCASt, as shown in Figure 2.

We also explored potential applications that can be derived from well-maintained knowledgebase systems such as Wikidata [22]. We elaborated on the possibility of using Wikidata items to automatically generate OpenMath content dictionaries, which are required to annotate more complex mathematical concepts in content MathML data [27]. The ability to link mathematical expressions and concepts with natural language expressions (which is possible in the case of Wikidata, even for multi-lingual expressions) is beneficial for a large variety of MathIR systems.

Hereafter we explain two projects that emerged from this idea—PhysWikiQuiz and MathQA.

1.2.5 PhysWikiQuiz & MathQA

PhysWikiQuiz [26] and its predecessor MathQA [3, 30] are systems that use Wikidata to automatically generate mathematical questions. Wikidata is a knowledge graph with items connected via a large variety of properties. It can be exploited to generate math-related questions, e.g., for question-answering systems and educational purposes. The PhysWikiQuiz system first takes an example formula from a Wikidata item, such as $v = s/t$ (speed). It then retrieves the names and units of each symbol involved in the formula via Wikidata properties. Using a computer algebra system (in this case, SymPy), the system rearranges the formula and generates test calculations that can be presented to the user and verified by the computer algebra system. For instance, given the input speed', the system may generate the question: `What is the distance` s`, given speed` $v = 10ms^{-1}$ `and duration` $t = 6s$?, with an expected correct answer of $60m$. The system also checks the value and unit of the answer.

1.3 Project Adjustments

Significant adjustments to the research agenda were not necessary but the project team had to relocate multiple times, moving from the University of Konstanz to the University of Wuppertal, and later to the University of Göttingen. These relocations caused staff shortages, which required applying for an extension of the originally planned project timeline. While the extension was necessary, it also provided opportunities that would have been difficult to achieve otherwise, such as collaborating with new partners from new institutions.

1.4 Commercial Viability

Although some of our projects, such as PhysWikiQuiz and LaCASt, have the potential for commercial exploitation, we do not plan to use the project results commercially. We are committed to making the findings and systems resulting from our research accessible to everyone free of charge as we firmly believe that open access is crucial for advancing research and innovation.

1.5 Follow-up Projects

The improved accessibility and retrievability of mathematical knowledge achieved in this project have enabled research on downstream applications.

One such application is *Math-based plagiarism detection (MathPD)*, i.e., the search for mathematical content in academic documents that is similar to the content in prior publications without that being justified and acknowledged according to academic standards. MathPD poses numerous application-specific questions and challenges that we investigate in the DFG project *Analyzing Mathematics to Detect Disguised Academic Plagiarism* (project no. 437179652)

In addition to enabling checks for plagiarized mathematical content, we seek to improve the plagiarism detection process by eliminating data privacy concerns and the dependence on near-monopolistic service providers. To achieve this goal, we are researching privacy-preserving methods to identify similar content in academic publications, which includes text, images, citations, and mathematical content. Moreover, we investigate the use of distributed ledger technology to create a distributed, trustless system for running privacy-preserving plagiarism checks. We will submit a funding proposal for this research to the DFG in July 2023.

The technology developed in this project serves as the foundation for the portal of the National Research Data Initiative (NFDI) project for mathematics (MaRDI) (project no. 460135501).

Specifically, the semantically enhanced version of the DLMF formulae was the initial seed of mathematical research data imported to the portal `portal.mardi4nfdi.de`. In the MaRDI project, we continue our efforts to enhance the semantics of mathematical research data and improve its accessibility and retrievability.

1.6 Contributors

This section provides an overview of the primary contributors and project partners.

1.6.1 Project Employees

This project supported four research associates: Corinna Breitinger, Dr. André Greiner-Petter, Dr. Norman Meuschke, and Dr. Moritz Schubotz.

1.6.2 Partner Institutions

Besides the involved host institutions, the universities of Konstanz, Wuppertal and Göttingen in Germany, we collaborated with four international institutions:

1. FIZ Karlsruhe - Leibniz Institute for Information Infrastructure (zbMATH Open), Berlin, Germany
2. National Institute of Informatics (NII), Tokyo, Japan
3. National Institute for Standards and Technology (NIST), Gaithersburg, USA
4. Wikimedia Foundation

1.6.3 Qualification of Junior Researchers

The following doctoral dissertations have been completed during the project:

1. André Greiner-Petter, Making Presentation Math Computable - A Context-Sensitive Approach for Translating LaTeX to Computer Algebra Systems. Dissertation, University of Wuppertal, 2022. [13]
2. Norman Meuschke, Analyzing Non-Textual Content Elements to Detect Academic Plagiarism. Dissertation, University of Konstanz, 2021. [17]

The following doctoral dissertations are still in progress:

1. Phillipp Scharpf, Mathematical Entity Linking, University of Konstanz
2. Corinna Breitinger, Academic Recommender Systems, University of Konstanz

In addition, 7 master's and 17 bachelor's theses were completed in relation to the project.

1.7 Contributions of Project Publications to Grant Objectives

Table 1 shows the project's work packages as defined in the proposal and Table 2 how the publications resulting from this project contributed to the work packages. The proposal (in German) is available here: `https://doi.org/10.5281/zenodo.7908591`.

Table 1: Short description of work packages.

WP	Work Package Description
1	**Syntactic Analysis of Mathematical Expressions**
1.1	Differentiating Math and Non-Math Content
1.2	Classification of Mathematical Expressions
1.3	Tokenization of Complex Mathematical Expressions
1.4	Pattern Matching and Interaction
2	**Semantic Enrichment with Math Concepts**
2.1	Annotating Math Tokens with Natural Language Tokens
2.2	Annotating Math Tokens with Mathematical Concepts
2.3	Consistency and Quality Checks
3	**Quality Metrics, Demonstrators, and Evaluation**
3.1	Development of Quality Metrics for Mathematical Markups
3.2	Demonstrators
3.3	Evaluations

Table 2: Overview of project publications and their contributions to the work packages.

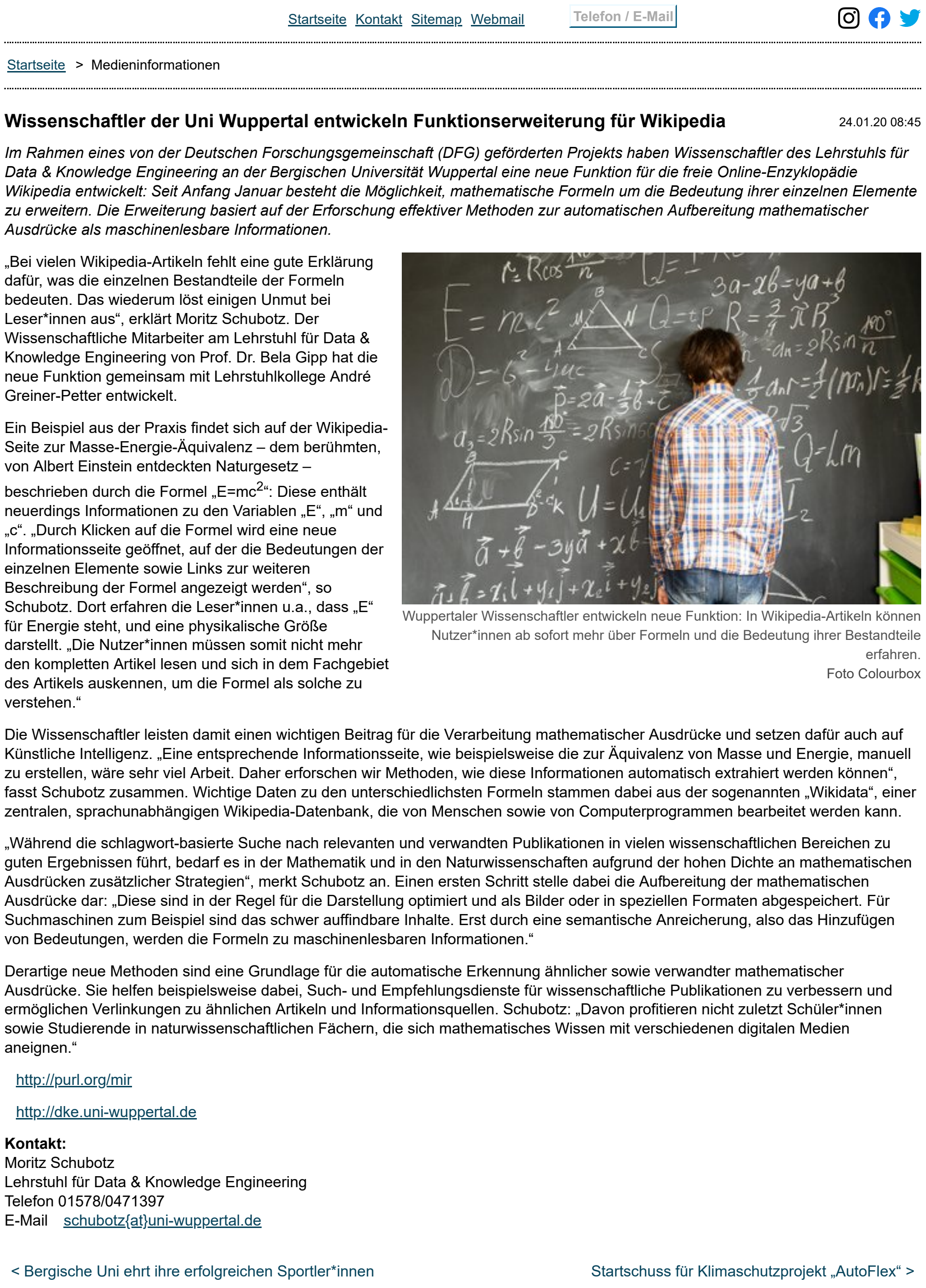
Year	Publication	WP 1				WP 2			WP 3		
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2022	A. Greiner-Petter et al. “Do the Math: Making Mathematics in Wikipedia Computable”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (Aug. 2022), pp. 4384–4395. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2022.3195261; [1]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	A. Greiner-Petter et al. “Comparative Verification of the Digital Library of Mathematical Functions and Computer Algebra Systems”. In: *Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, (TACAS).* Springer, Apr. 2022, pp. 87–105. DOI: 10.1007/978-3-030-99524-9_5; [2]		✗	✗		✗	✗		✗		✗
	P. Scharpf, M. Schubotz, and B. Gipp. “Mining Mathematical Documents for Question Answering via Unsupervised Formula Labeling”. In: *Proc. ACM/IEEE JCDL.* ACM, June 2022, pp. 1–11. DOI: 10.1145/3529372.3530925; [3]					✗	✗	✗			
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2021	P. Scharpf, M. Schubotz, and B. Gipp. “Fast Linking of Mathematical Wikidata Entities in Wikipedia Articles Using Annotation Recommendation”. In: *Proc. WWW.* ACM, Apr. 2021, pp. 602–609. DOI: 10.1145/3442442.3452348; [22]					✗	✗	✗			
	P. Scharpf, M. Schubotz, and B. Gipp. “Mathematics in Wikidata”. In: *Proc. WWW.* vol. 2982. CEUR-WS.org, 2021; [23]					✗	✗				
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2020	A. Greiner-Petter et al. “Discovering Mathematical Objects of Interest — A Study of Mathematical Notations”. In: *Proc. WWW.* ACM, Apr. 2020, pp. 1445–1456. DOI: 10.1145/3366423.3380218; [4]	✗		✗		✗					
	T. Asakura et al. “Towards Grounding of Formulae”. In: *Proceedings of the First Workshop on Scholarly Document Processing (SDP@EMNLP).* ACL, 2020, pp. 138–147. DOI: 10.18653/v1/2020.sdp-1.16; [11]	✗		✗		✗		✗	✗		✗
	M. Schubotz et al. “Mathematical Formulae in Wikimedia Projects 2020”. In: *Proc. ACM/IEEE JCDL.* ACM, Aug. 2020, pp. 447–448. DOI: 10.1145/3383583.3398557; [29]		✗			✗	✗	✗		✗	
	A. Greiner-Petter et al. “Math-Word Embedding in Math Search and Semantic Extraction”. In: *Scientometrics* 125.3 (Dec. 2020), pp. 3017–3046. ISSN: 0138-9130. DOI: 10.1007/s11192-020-03502-9; [5]	✗				✗					✗

Year	Publication	WP 1				WP 2			WP 3		
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2020	A. Greiner-Petter et al. “Making Presentation Math Computable: Proposing a Context Sensitive Approach for Translating LaTeX to Computer Algebra Systems”. In: *Proc. ICMS*. vol. 12097. Springer, 2020, pp. 335–341. DOI: 10.1007/978-3-030-52200-1_33; [15]	✗				✗	✗	✗	✗	✗	✗
	P. Scharpf et al. “ARQMath Lab: An Incubator for Semantic Formula Search in zbMATH Open?” In: *Working Notes of (CLEF) 2020 - Conference and Labs of the Evaluation Forum*. Vol. 2696. CEUR-WS.org, 2020; [25]					✗		✗			
	P. Scharpf et al. “Classification and Clustering of arXiv Documents, Sections, and Abstracts, Comparing Encodings of Natural and Mathematical Language”. In: *Proc. ACM/IEEE JCDL*. ACM, Aug. 2020, pp. 137–146. DOI: 10.1145/3383583.3398529; [6]					✗					
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2019	N. Meuschke et al. “Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations”. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, June 2019, pp. 120–129. DOI: 10.1109/jcdl.2019.00026; [8]	✗								✗	✗
	P. Scharpf et al. “AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Sept. 2019, pp. 532–533. DOI: 10.1145/3298689.3347042; [9]					✗	✗	✗			
	A. Greiner-Petter et al. “Why Machines Cannot Learn Mathematics, Yet”. In: *Proc. BIRNDL at ACM SIGIR*. vol. 2414. CEUR-WS.org, 2019; [14]	✗				✗					✗
	A. Greiner-Petter et al. “Semantic Preserving Bijective Mappings for Expressions Involving Special Functions between Computer Algebra Systems and Document Preparation Systems”. In: *Aslib Journal of Information Management* 71.3 (May 2019), pp. 415–439. ISSN: 2050-3806. DOI: 10.1108/AJIM-08-2018-0185; [7]			✗		✗	✗		✗		✗
	P. Scharpf et al. “Towards Formula Concept Discovery and Recognition”. In: *Proc. ACM SIGIR*. vol. 2414. CEUR-WS.org, 2019, pp. 108–115; [21]						✗		✗		
	M. Schubotz et al. “Forms of Plagiarism in Digital Mathematical Libraries”. In: *Intelligent Computer Mathematics*. Vol. 11617. Springer International Publishing, 2019, pp. 258–274. DOI: 10.1007/978-3-030-23250-4_18; [31]				✗				✗		✗

Year	Publication	WP 1				WP 2			WP 3		
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2018	H. S. Cohl, A. Greiner-Petter, and M. Schubotz. “Automated Symbolic and Numerical Testing of DLMF Formulae Using Computer Algebra Systems”. In: *Proc. CICM*. vol. 11006. Springer International Publishing, 2018, pp. 39–52. doi: 10.1007/978-3-319-96812-4_4; [12]		✗	✗		✗	✗		✗		✗
	M. Schubotz et al. “Improving the Representation and Conversion of Mathematical Formulae by Considering Their Textual Context”. In: *Proc. ACM/IEEE JCDL*. ACM, May 2018, pp. 233–242. doi: 10.1145/3197026.3197058; [10]		✗	✗	✗	✗	✗		✗		✗
	N. Meuschke et al. “HyPlag: A Hybrid Approach to Academic Plagiarism Detection”. In: *Proc. ACM SIGIR*. ACM, June 2018, pp. 1321–1324. doi: 10.1145/3209978.3210177; [19]									✗	✗
	A. Greiner-Petter et al. “MathTools: An Open API for Convenient MathML Handling”. In: *Proc. CICM*. vol. 11006. Springer International Publishing, 2018, pp. 104–110. doi: 10.1007/978-3-319-96812-4_9; [16]			✗					✗		✗
	F. Petersen, M. Schubotz, and B. Gipp. “Towards Formula Translation Using Recursive Neural Networks”. In: *WiP at CICM*. vol. 2307. CEUR-WS.org, 2018; [20]				✗				✗		✗
	P. Scharpf, M. Schubotz, and B. Gipp. “Representing Mathematical Formulae in Content MathML Using Wikidata”. In: *Proc. BIRNDL at ACM SIGIR*. vol. 2132. CEUR-WS.org, 2018; [24]					✗	✗			✗	
	M. Schubotz. “Generating OpenMath Content Dictionaries from Wikidata”. In: *WiP at CICM*. vol. 2307. CEUR-WS.org, 2018; [27]						✗				
	M. Schubotz. “Mathematische Formeln in Wikipedia”. In: *Beiträge zum Mathematikunterricht 2018*. Gesellschaft für Didaktik der Mathematik, 2018, pp. 1635–1638. doi: 10.17877/DE290R-19676; [28]	✗	✗	✗	✗	✗	✗	✗		✗	
	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; [30]					✗	✗	✗			
		1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3
2017	N. Meuschke et al. “Analyzing Mathematical Content to Detect Academic Plagiarism”. In: *Proceedings ACM Conference on Information and Knowledge Management (CIKM)*. ACM, Nov. 2017, pp. 2211–2214. doi: 10.1145/3132847.3133144; [18]	✗								✗	✗

1.8 Press Releases

The press relations office at the University of Wuppertal reported on the project's contributions to improving the access to mathematical content in Wikipedia. The original press release is no longer available due to a complete overhaul and relaunch of the University's website. We attach a copy of the article (in German) below.



Startseite Kontakt Sitemap Webmail Telefon / E-Mail

Startseite > Medieninformationen

Wissenschaftler der Uni Wuppertal entwickeln Funktionserweiterung für Wikipedia

24.01.20 08:45

Im Rahmen eines von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts haben Wissenschaftler des Lehrstuhls für Data & Knowledge Engineering an der Bergischen Universität Wuppertal eine neue Funktion für die freie Online-Enzyklopädie Wikipedia entwickelt: Seit Anfang Januar besteht die Möglichkeit, mathematische Formeln um die Bedeutung ihrer einzelnen Elemente zu erweitern. Die Erweiterung basiert auf der Erforschung effektiver Methoden zur automatischen Aufbereitung mathematischer Ausdrücke als maschinenlesbare Informationen.

„Bei vielen Wikipedia-Artikeln fehlt eine gute Erklärung dafür, was die einzelnen Bestandteile der Formeln bedeuten. Das wiederum löst einigen Unmut bei Leser*innen aus", erklärt Moritz Schubotz. Der Wissenschaftliche Mitarbeiter am Lehrstuhl für Data & Knowledge Engineering von Prof. Dr. Bela Gipp hat die neue Funktion gemeinsam mit Lehrstuhlkollege André Greiner-Petter entwickelt.

Ein Beispiel aus der Praxis findet sich auf der Wikipedia-Seite zur Masse-Energie-Äquivalenz – dem berühmten, von Albert Einstein entdeckten Naturgesetz – beschrieben durch die Formel „E=mc^2": Diese enthält neuerdings Informationen zu den Variablen „E", „m" und „c". „Durch Klicken auf die Formel wird eine neue Informationsseite geöffnet, auf der die Bedeutungen der einzelnen Elemente sowie Links zur weiteren Beschreibung der Formel angezeigt werden", so Schubotz. Dort erfahren die Leser*innen u.a., dass „E" für Energie steht, und eine physikalische Größe darstellt. „Die Nutzer*innen müssen somit nicht mehr den kompletten Artikel lesen und sich in dem Fachgebiet des Artikels auskennen, um die Formel als solche zu verstehen."

Wuppertaler Wissenschaftler entwickeln neue Funktion: In Wikipedia-Artikeln können Nutzer*innen ab sofort mehr über Formeln und die Bedeutung ihrer Bestandteile erfahren.
Foto Colourbox

Die Wissenschaftler leisten damit einen wichtigen Beitrag für die Verarbeitung mathematischer Ausdrücke und setzen dafür auch auf Künstliche Intelligenz. „Eine entsprechende Informationsseite, wie beispielsweise die zur Äquivalenz von Masse und Energie, manuell zu erstellen, wäre sehr viel Arbeit. Daher erforschen wir Methoden, wie diese Informationen automatisch extrahiert werden können", fasst Schubotz zusammen. Wichtige Daten zu den unterschiedlichsten Formeln stammen dabei aus der sogenannten „Wikidata", einer zentralen, sprachunabhängigen Wikipedia-Datenbank, die von Menschen sowie von Computerprogrammen bearbeitet werden kann.

„Während die schlagwort-basierte Suche nach relevanten und verwandten Publikationen in vielen wissenschaftlichen Bereichen zu guten Ergebnissen führt, bedarf es in der Mathematik und in den Naturwissenschaften aufgrund der hohen Dichte an mathematischen Ausdrücken zusätzlicher Strategien", merkt Schubotz an. Einen ersten Schritt stelle dabei die Aufbereitung der mathematischen Ausdrücke dar: „Diese sind in der Regel für die Darstellung optimiert und als Bilder oder in speziellen Formaten abgespeichert. Für Suchmaschinen zum Beispiel sind das schwer auffindbare Inhalte. Erst durch eine semantische Anreicherung, also das Hinzufügen von Bedeutungen, werden die Formeln zu maschinenlesbaren Informationen."

Derartige neue Methoden sind eine Grundlage für die automatische Erkennung ähnlicher sowie verwandter mathematischer Ausdrücke. Sie helfen beispielsweise dabei, Such- und Empfehlungsdienste für wissenschaftliche Publikationen zu verbessern und ermöglichen Verlinkungen zu ähnlichen Artikeln und Informationsquellen. Schubotz: „Davon profitieren nicht zuletzt Schüler*innen sowie Studierende in naturwissenschaftlichen Fächern, die sich mathematisches Wissen mit verschiedenen digitalen Medien aneignen."

http://purl.org/mir

http://dke.uni-wuppertal.de

Kontakt:
Moritz Schubotz
Lehrstuhl für Data & Knowledge Engineering
Telefon 01578/0471397
E-Mail schubotz{at}uni-wuppertal.de

< Bergische Uni ehrt ihre erfolgreichen Sportler*innen

Startschuss für Klimaschutzprojekt „AutoFlex" >

References

[1] A. Greiner-Petter, M. Schubotz, C. Breitinger, P. Scharpf, A. Aizawa, and B. Gipp. "Do the Math: Making Mathematics in Wikipedia Computable". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (Aug. 2022), pp. 4384–4395. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2022.3195261.

[2] A. Greiner-Petter, H. S. Cohl, A. Youssef, et al. "Comparative Verification of the Digital Library of Mathematical Functions and Computer Algebra Systems". In: *Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, (TACAS)*. Springer, Apr. 2022, pp. 87–105. DOI: 10.1007/978-3-030-99524-9_5.

[3] P. Scharpf, M. Schubotz, and B. Gipp. "Mining Mathematical Documents for Question Answering via Unsupervised Formula Labeling". In: *Proc. ACM/IEEE JCDL*. ACM, June 2022, pp. 1–11. DOI: 10.1145/3529372.3530925.

[4] A. Greiner-Petter, M. Schubotz, F. Müller, et al. "Discovering Mathematical Objects of Interest — A Study of Mathematical Notations". In: *Proc. WWW*. ACM, Apr. 2020, pp. 1445–1456. DOI: 10.1145/3366423.3380218.

[5] A. Greiner-Petter, A. Youssef, T. Ruas, et al. "Math-Word Embedding in Math Search and Semantic Extraction". In: *Scientometrics* 125.3 (Dec. 2020), pp. 3017–3046. ISSN: 0138-9130. DOI: 10.1007/s11192-020-03502-9.

[6] P. Scharpf, M. Schubotz, A. Youssef, F. Hamborg, N. Meuschke, and B. Gipp. "Classification and Clustering of arXiv Documents, Sections, and Abstracts, Comparing Encodings of Natural and Mathematical Language". In: *Proc. ACM/IEEE JCDL*. ACM, Aug. 2020, pp. 137–146. DOI: 10.1145/3383583.3398529.

[7] A. Greiner-Petter, M. Schubotz, H. S. Cohl, and B. Gipp. "Semantic Preserving Bijective Mappings for Expressions Involving Special Functions between Computer Algebra Systems and Document Preparation Systems". In: *Aslib Journal of Information Management* 71.3 (May 2019), pp. 415–439. ISSN: 2050-3806. DOI: 10.1108/AJIM-08-2018-0185.

[8] N. Meuschke, V. Stange, M. Schubotz, M. Kramer, and B. Gipp. "Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations". In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, June 2019, pp. 120–129. DOI: 10.1109/jcdl.2019.00026.

[9] P. Scharpf, I. Mackerracher, M. Schubotz, J. Beel, C. Breitinger, and B. Gipp. "AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Sept. 2019, pp. 532–533. DOI: 10.1145/3298689.3347042.

[10] M. Schubotz, A. Greiner-Petter, P. Scharpf, N. Meuschke, H. S. Cohl, and B. Gipp. "Improving the Representation and Conversion of Mathematical Formulae by Considering Their Textual Context". In: *Proc. ACM/IEEE JCDL*. ACM, May 2018, pp. 233–242. DOI: 10.1145/3197026.3197058.

[11] T. Asakura, A. Greiner-Petter, A. Aizawa, and Y. Miyao. "Towards Grounding of Formulae". In: *Proceedings of the First Workshop on Scholarly Document Processing (SDP@EMNLP)*. ACL, 2020, pp. 138–147. DOI: 10.18653/v1/2020.sdp-1.16.

[12] H. S. Cohl, A. Greiner-Petter, and M. Schubotz. “Automated Symbolic and Numerical Testing of DLMF Formulae Using Computer Algebra Systems”. In: *Proc. CICM.* Vol. 11006. Springer International Publishing, 2018, pp. 39–52. DOI: 10.1007/978-3-319-96812-4_4.

[13] A. Greiner-Petter. *Making Presentation Math Computable: A Context-Sensitive Approach for Translating LaTeX to Computer Algebra Systems.* Springer Fachmedien Wiesbaden, 2023. ISBN: 978-3-658-40473-4. DOI: 10.1007/978-3-658-40473-4.

[14] A. Greiner-Petter, T. Ruas, M. Schubotz, A. Aizawa, W. I. Grosky, and B. Gipp. “Why Machines Cannot Learn Mathematics, Yet”. In: *Proc. BIRNDL at ACM SIGIR.* Vol. 2414. CEUR-WS.org, 2019.

[15] A. Greiner-Petter, M. Schubotz, A. Aizawa, and B. Gipp. “Making Presentation Math Computable: Proposing a Context Sensitive Approach for Translating LaTeX to Computer Algebra Systems”. In: *Proc. ICMS.* Vol. 12097. Springer, 2020, pp. 335–341. DOI: 10.1007/978-3-030-52200-1_33.

[16] A. Greiner-Petter, M. Schubotz, H. S. Cohl, and B. Gipp. “MathTools: An Open API for Convenient MathML Handling”. In: *Proc. CICM.* Vol. 11006. Springer International Publishing, 2018, pp. 104–110. DOI: 10.1007/978-3-319-96812-4_9.

[17] N. Meuschke. “Analyzing Non-Textual Content Elements to Detect Academic Plagiarism”. PhD thesis. University of Konstanz, Dept. of Computer and Information Science, 2021. DOI: 10.5281/zenodo.4913345.

[18] N. Meuschke, M. Schubotz, F. Hamborg, T. Skopal, and B. Gipp. “Analyzing Mathematical Content to Detect Academic Plagiarism”. In: *Proceedings ACM Conference on Information and Knowledge Management (CIKM).* ACM, Nov. 2017, pp. 2211–2214. DOI: 10.1145/3132847.3133144.

[19] N. Meuschke, V. Stange, M. Schubotz, and B. Gipp. “HyPlag: A Hybrid Approach to Academic Plagiarism Detection”. In: *Proc. ACM SIGIR.* ACM, June 2018, pp. 1321–1324. DOI: 10.1145/3209978.3210177.

[20] F. Petersen, M. Schubotz, and B. Gipp. “Towards Formula Translation Using Recursive Neural Networks”. In: *WiP at CICM.* Vol. 2307. CEUR-WS.org, 2018.

[21] P. Scharpf, M. Schubotz, H. S. Cohl, and B. Gipp. “Towards Formula Concept Discovery and Recognition”. In: *Proc. ACM SIGIR.* Vol. 2414. CEUR-WS.org, 2019, pp. 108–115.

[22] P. Scharpf, M. Schubotz, and B. Gipp. “Fast Linking of Mathematical Wikidata Entities in Wikipedia Articles Using Annotation Recommendation”. In: *Proc. WWW.* ACM, Apr. 2021, pp. 602–609. DOI: 10.1145/3442442.3452348.

[23] P. Scharpf, M. Schubotz, and B. Gipp. “Mathematics in Wikidata”. In: *Proc. WWW.* Vol. 2982. CEUR-WS.org, 2021.

[24] P. Scharpf, M. Schubotz, and B. Gipp. “Representing Mathematical Formulae in Content MathML Using Wikidata”. In: *Proc. BIRNDL at ACM SIGIR.* Vol. 2132. CEUR-WS.org, 2018.

[25] P. Scharpf, M. Schubotz, A. Greiner-Petter, M. Ostendorff, O. Teschke, and B. Gipp. “ARQMath Lab: An Incubator for Semantic Formula Search in zbMATH Open?” In: *Working Notes of (CLEF) 2020 - Conference and Labs of the Evaluation Forum.* Vol. 2696. CEUR-WS.org, 2020.

[26] P. Scharpf, M. Schubotz, A. Spitz, A. Greiner-Petter, and B. Gipp. “Collaborative and AI-aided Exam Question Generation Using Wikidata in Education”. In: *Proc. WWW.* Vol. 3262. CEUR-WS.org, 2022.

[27] M. Schubotz. “Generating OpenMath Content Dictionaries from Wikidata”. In: *WiP at CICM.* Vol. 2307. CEUR-WS.org, 2018.

[28] M. Schubotz. “Mathematische Formeln in Wikipedia”. In: *Beiträge zum Mathematikunterricht 2018.* Gesellschaft für Didaktik der Mathematik, 2018, pp. 1635–1638. DOI: 10.17877/DE290R-19676.

[29] M. Schubotz, A. Greiner-Petter, N. Meuschke, O. Teschke, and B. Gipp. “Mathematical Formulae in Wikimedia Projects 2020”. In: *Proc. ACM/IEEE JCDL.* ACM, Aug. 2020, pp. 447–448. DOI: 10.1145/3383583.3398557.

[30] M. Schubotz, P. Scharpf, K. Dudhat, Y. Nagar, F. Hamborg, and B. Gipp. “Introducing MathQA - a Math-Aware Question Answering System”. In: *Proc. ACM/IEEE JCDL.* June 2018. DOI: 10.1108/idd-06-2018-0022.

[31] M. Schubotz, O. Teschke, V. Stange, N. Meuschke, and B. Gipp. “Forms of Plagiarism in Digital Mathematical Libraries”. In: *Intelligent Computer Mathematics.* Vol. 11617. Springer International Publishing, 2019, pp. 258–274. DOI: 10.1007/978-3-030-23250-4_18.