zbMATH Open: API Solutions and Research Challenges

Matteo Petrera[1], Dennis Trautwein[2], Isabel Beckenbach[1], Dariush Ehsani[1], Fabian Müller[1], Olaf Teschke[1], Bela Gipp[2] and Moritz Schubotz[1,2]

[1]zbMATH / FIZ Karlsruhe, Berlin, Germany, first.last@fiz-karlsruhe.de
[2]Bergische Universität Wuppertal, Wuppertal, Germany, last@gipplab.org

Abstract

We present zbMATH Open, the most comprehensive collection of reviews and bibliographic metadata of scholarly literature in mathematics. Besides our website zbMATH.org which is openly accessible since the beginning of this year, we provide API endpoints to offer our data. APIs improve interoperability with others, i.e., digital libraries, and allow using our data for research purposes. In this article, we (1) illustrate the current and future overview of the services offered by zbMATH; (2) present the initial version of the zbMATH links API; (3) analyze potentials and limitations of the links API based on the example of the NIST Digital Library of Mathematical Functions; (4) and finally, present the zbMATH Open dataset as a research resource and discuss connected open research problems.

1. Introduction

Since the beginning of 2021, zbMATH is open for public access. Currently, zbMATH Open[1] contains over 4 million bibliographic entries with reviews contributed by more than 7,000 active reviewers and abstracts drawn from more than 3,000 journals and book series, and more than 190,000 books. For most working mathematicians, this means that they can access zbMATH from anywhere in the world without subscription nor authentication. Additionally, we envision benefits to the community by our efforts to connect zbMATH data with information systems of research data, collaborative platforms, funding agencies, and intra-disciplinary efforts, as outlined in [8, 18]. We expect that our commitment in disseminating mathematics research results will increase the visibility of mathematics for any scientific audience. We invite the mathematical community to participate actively in the further development of the platform.

Very recently, at zbMATH, efforts have been spent to develop Application Programming Interface (API) solutions to facilitate and optimize open-access to mathematical research data.

In Figure 1, we sketch a conceptual overview of zbMATH's services. The boxes "Reviewer Interface", "Internal Interfaces", and "zbMATH.org Website" show the well-established components of zbMATH and are outside the scope of this paper. The box "OAI-PMH API" was released in April 2021 [18]. This protocol is widely used for metadata-harvesting. Via the OAI-PMH API[2], researchers can harvest the entire dataset or only specific subsets of our collection. We offer the data in two flavors, the standardized Dublin Core[3] metadata format and a second format, that is closer to zbMATH's internal data model. The content generated by zbMATH Open, such as reviews, classifications, software, or author disambiguation data are distributed under CC-BY-SA 4.0. This defines the license for the whole dataset, which also contains non-copyrighted bibliographic metadata and reference data derived from I4OSC (CC0). Note that the API does only provide a subset of the data in the zbMATH Open Web interface since in several cases third-party information, such as abstracts, cannot be made available under a suitable license through the API. In those cases we replaced the data with a placeholder string. We envision that for researchers dealing with different data providers, the Dublin Core format is more suitable. We expect that for people used to our website, our own format is more appealing to use. From

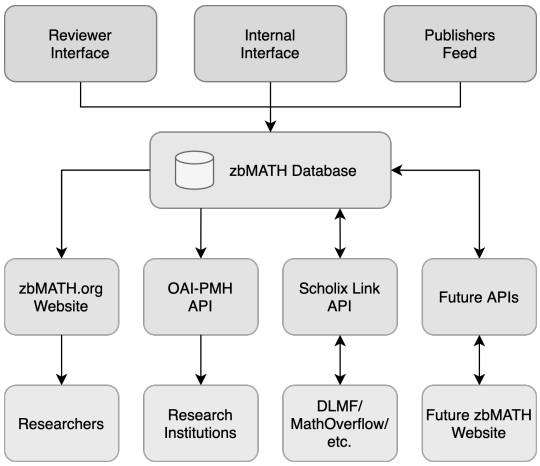

Figure 1: Overview of the zbMATH database and its associated data flows. This paper focuses on the "Scholix Links API". "Future APIs" are under construction.

Digital Infrastructures for Scholarly Content Objects (DISCO2021) at JCDL2021

[1]https://zbmath.org/
[2]https://oai.zbmath.org
[3]https://dublincore.org/

the API one can fetch the entire dataset or a well-defined subset using a metadata harvester[4]. One harvest output will be permanently stored as a research dataset of the Special Interest Group on Maths Linguistics data repository. This data repository also contains annual snapshots of arXiv[5] articles in different formats optimized for mathematical information retrieval research challenges. As the zbMATH open data links to many arXiv preprints, we plan to synchronize the release cycles to create consistent snapshots of zbMATH data and associated fulltext sources.

In this paper, we describe a new service offered by zbMATH, namely an API, called "zbMATH Links API", represented by the box stating "Scholix Link API" in Figure 1. At present, this new API is focused on the interconnections between zbMATH and the Digital Library of Mathematical Functions (DLMF)[6], even though more partners are expected to be hosted soon (e.g., MathOverflow, arXiv, Online Encyclopedia of Integer Sequences). Search engines or researchers from mathematics or the field of bibliometric research might use our zbMATH Links API to present and use the search results. Furthermore, the source code of our API has been released in the form of a Python package[7], so that any interested user can use it for similar purposes in any context where the interconnection between bibliographic data and links has to be studied and documented. In this way, we hope to serve the needs of a wide range of potential users.

The main contributions of this paper are:

1. We provide an overview of the new API implementation using the example of how DLMF makes use of it. An analysis of the currently available dataset will be outlined.

2. We present other natural candidates for the API, thus proving the potential coverage of the current mathematical literature.

3. We highlight implications and new research potentials by showing how existing research can be transferred to make use of zbMATHs open APIs.

In the following section 2, we motivate the choice of DLMF as the first partner for our new API and how it is currently used in their environment. Afterward, in section 3, we present the implementation details, analyze the DLMF link data and give some details about other potential partners. In section 4, we discuss the technical capabilities of the new API and compare the capabilities of the open APIs of zbMATH with its pendant of PubMed. The last section is devoted to some concluding remarks and open problems.

[4] https://www.openarchives.org/pmh/tools/
[5] https://arxiv.org/
[6] https://dlmf.nist.gov/
[7] https://purl.org/zb/13

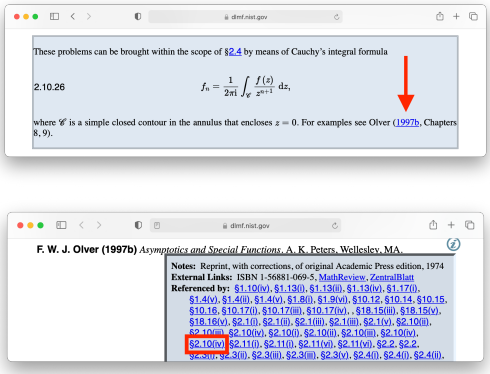

Figure 2: A reference in DLMF, available at https://dlmf.nist.gov/bib/O (below), and a link to it, https://dlmf.nist.gov/2.10#iv.p2 (above)

2. DLMF as a zbMATH partner

Among all possible partners that may interact with zbMATH, we selected the aforementioned Digital Library of Mathematical Functions (DLMF) as a first partner. In addition to being an important reference tool for mathematicians, DLMF offers a relatively small bibliographic catalog and is therefore very well suited for testing our API.

DLMF is a well-established web resource that enlarges and translates the classical "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", edited by M. Abramowitz and I. A. Stegun in 1964 into a modern and functional digital library. As the original book's title inspiring this web service suggests, DLMF is a digital handbook about theoretical and computational aspects of special functions. Its primary purpose is to provide a modern reference tool for researchers in mathematics, physical sciences, and engineering. It contains hundreds of definitions and theorems, presented with a standardized notation, together with tables, figures, and references to peer-reviewed papers and books. It was published online on the May 7th 2010 and is continuously maintained, reviewed, and updated ever since. Indeed, the field of special functions still receives great attention from the mathematics community, and new contributions enrich the contents of the library year by year. DLMF presents its contents in 36 chapters, and the bibliography currently consists of 2,748 references[8] of which 2,053 directly link to zbMATH (i.e., about 75%). This is a valuable service offered independently by DLMF and zbMATH since each user has the possibility of accessing all selected publications' bibliographic data. Let us note that of the remaining 25% of publications not linked to zbMATH, most of them do not belong to the zbMATH database.

Before providing more details about our Links API, let

[8] https://dlmf.nist.gov/bib/

us mention a few details about the links' structure we are interested in. Each reference in the DLMF bibliography may be cited many times in the DLMF pages. Each of these instances carries its own link to zbMATH. For example, the book "Asymptotics and special functions" by F. W. J. Olver (Reprint, 1997; Zbl 0982.41018)[9] is referenced 332 times. Each citation defines a link to zbMATH uniquely. An example of one of these links is: https://dlmf.nist.gov/2.10#iv.p2 (see Figure 2). In this case, Olver's book is referenced in Part 2 of Section §2.10(iv) Taylor and Laurent Coefficients: Darboux's Method. In Figure 2, we also see that the Section §2.10(iv) is cited 3 times. Each instance corresponds to a link that points to a different destination site in the DLMF library. The highlighted §2.10(iv) points to what we see in the first screenshot of Figure 2.

3. zbMATH Links API

This section presents the main features of the new "zbMATH Links API" by explaining its structure and various technical capabilities. Then, we give an analysis of the link statistics associated with our DLMF collaboration.

3.1. Structure of the API

The API itself has been implemented in Python and is described using the OpenAPI Specification[10], a language-agnostic interface description standard for APIs. At present, it hosts only one partner, DLMF, but it will soon host other partners. The underlying dataset has been generated by scraping the DLMF bibliography. As a result, we got 2,053 references (indexed at zbMATH) and 6,526 distinct links. In this framework, the links are objects belonging to the *source* (of a given partner; DLMF in the present case), and zbMATH objects are objects belonging to the *target*.

The API offers eight endpoints, more specifically six GET routes, one POST route, and one PUT route. The Swagger UI of the zbMATH Links API is available online[11]. Here is a concise listing of the provided functionalities:

- `GET /partner` retrieves data of a given zbMATH partner.
- `PUT /partner` edits data of a given zbMATH partner.
- `GET /link` retrieves links for a given zbMATH object. The parameters are: Authors, MSC codes[12], X-Field[13].

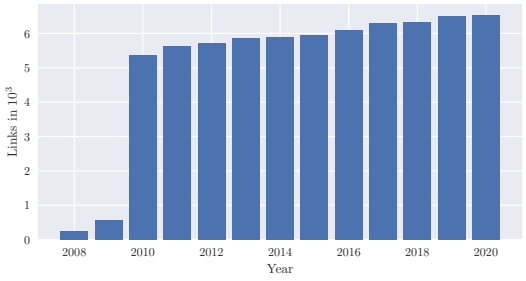

Figure 3: Number of links to the zbMATH API. One can see a huge increase in 2010 – the year DLMF officially started.

- `GET /link/item` checks relations (if any) between a given link identifier (e.g., 2.10#iv.p2) and a given zbMATH object (e.g., Zbl 0982.41018). The parameters are: Zbl code, Source identifier, Partner name, X-Field.
- `POST /link` allows any user of the API to create a new link (for a given partner) related to a zbMATH object. The parameters are: Zbl code, Source identifier, Partner name, Link relation.
- `GET /source` gives a list of all links of a given zbMATH partner.
- `GET /statistics/msc` shows the occurrence of primary MSC codes (2-digit level) in the source.
- `GET /statistics/year` shows the occurrence of years of publication of references in the source.

Our JSON response body is modeled on the Scholix metadata schema[14]. The models used to pack the data are explicitly reported in the API web interface. It is worth recalling that Scholix is a well-established framework to exchange information between data and literature links. The schema's architecture is designed to allow for bulk exchange of link information, which contains all necessary data to keep track of bibliographic parameters identifying scholarly links.

3.2. Analysis of DLMF Data

Based on our available DLMF dataset, it is possible to draw some conclusions:

- In the JSON response body of our `GET /link` methods, one can see that each link is equipped with a publication date. This date refers to the date the link itself has been added in the DLMF bibliography. We scraped the historical bibliography between 2008 and 2020 (December is

[9] https://zbmath.org/?q=an%3A0982.41018
[10] https://swagger.io/specification/
[11] https://purl.org/zb/14
[12] Mathematics Subject Classification Scheme 2020, https://msc2020.org/
[13] The X-Field is an optional parameter that can be used when one is running a query that can pull back a lot of metadata, but only a few fields in the output are of interest. Example: in the GET/link one is interested only in retrieving the id identifier of sources where the name of the author is Abramowitz. Then, Author: Abramowitz, X-Field: {Source{Identifier{ID}}}.
[14] http://www.scholix.org/schema/3-0

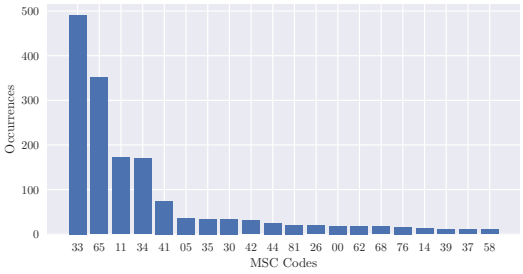

Figure 4: Distribution of primary 2-digit MSC codes in the DLMF dataset

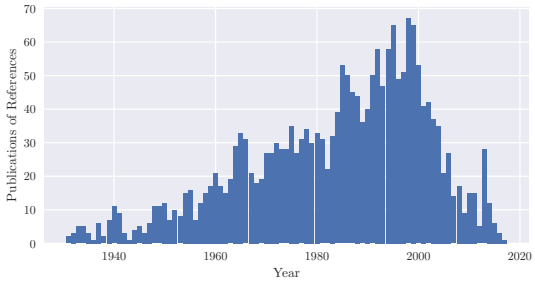

Figure 5: Distribution of years of publication of references in the DLMF dataset

the scraping's reference month) and found the growth numbers depicted in Figure 3. Clearly, the growth of population of references changed drastically in 2010, the year when DLMF started officially.

- The two statistics routes show results concerning the distribution of primary MSC codes (2-digit level) and years of publication of the references in the current dataset. As one may expect, the most frequently cited primary MSC codes are:

MSC Code	References	Area
33	491	Special functions
65	351	Numerical analysis
11	172	Number theory

See Figure 4 for more details. On the other hand, the most frequent years of publication of cited references in the current dataset are:

References	67	65	65
Year	1998	1999	1995

See Figure 5 for more details. Looking at both Figures 3 and 5 we could infer that the DLMF bibliography suffers from a delay in updating its references. More precisely, the fact that the maximum peak is centered at the end of the 90s makes us think of some kind of difficulty in identifying relevant references referring to the last twenty years.

- The references in the current DLMF dataset which have the most citations are:
  - *F. W. J. Olver*, Asymptotics and special functions. Wellesley, MA: A K Peters (1997; Zbl 0982.41018): 332 citations,
  - *M. Abramowitz* (ed.) and *I. A. Stegun* (ed.), Handbook of mathematical functions with formulas, graphs and mathematical tables. Washington: U.S. Department of Commerce. (1964; Zbl 0171.38503): 118 citations,
  - *A. Erdélyi et al.*, Higher transcendental functions. Vol. I. New York: McGraw-Hill Book Co. (1953; Zbl 0051.30303): 110 citations.

  In Figure 6 one can see the references, identified by Zbl code, with more than 50 citations.

3.3. Usage

The motivation behind the recent implementation of APIs at zbMATH is twofold. On the one hand, we want to offer to the scientific community an efficient and open access to our data. On the other hand, we wish to expose the dynamic interaction between our bibliographic data and those coming from other resources. It is essential to note that both of these targets are made possible by zbMATH becoming an open web service. This provides a boost for disseminating scientific knowledge, and our work may help to understand how it spreads and auto-correlates in a functional way.

The zbMATH links API with its first partner DLMF represents a tool that can be used in various ways and contains many properties that are advantageous for the research process. Here, we want to present concrete usage instances where a user of either DLMF or zbMATH can generally benefit from the service:

- A DLMF user can access all bibliographic resources indexed at zbMATH relating to a specific topic of interest. This may help to get a consistent overview of the scientific development of the topic itself.
- A researcher interested in a publication indexed at zbMATH can use our API to verify if and possibly where that publication is cited in DLMF. A search of this type can also be very diversified thanks to the filters that our routes offer. For example, one might be interested in identifying which DLMF links are related to a particular

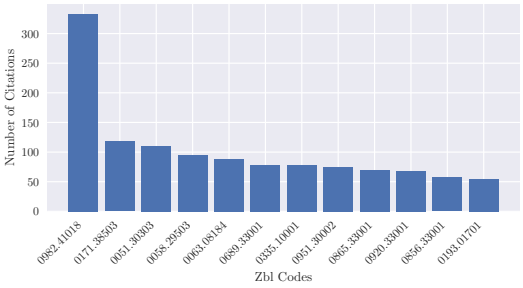

Figure 6: References (identified by Zbl code) in the DLMF dataset cited more than 50 times

MSC code or a particular author. This means that a targeted use of our API can allow a detailed bibliographic search that otherwise would not be possible.

- A researcher more interested in the history of mathematics can use our API to trace the bibliography related to a certain topic covered in DLMF and observe the historical development of the topic itself in terms of the literature related to it. Such research can be very rich and diverse. It is sufficient to think that in the field of special functions there are classical topics, such as the "gamma function" or "elliptic integrals", which have a long history behind them.

When other partners will be included in our API, the covered spectrum will expand considerably, thus providing the user with an efficient and flexible bibliographic searching tool.

In section 4, we will try to compare the service offered by API solutions at zbMATH with those offered by similar platforms. Therefore, the goal will be to understand in what aspects we can and must improve in the near future.

3.4. Limitations and Future Partners

While in general the Scholix API format, was a very good fit for our project we experienced some inconveniences. For one, the link description in the DLMF sometimes contains mathematical expressions. However, the API specification allows only string fields. It would be good if the standard could be expanded to allow for HTML or another way of expressing mathematical expressions within descriptions. Moreover, one of the problems we faced was modelling the MSC codes in the API. We chose the field "subtype" of the "type" attribute in Scholix. However, this does not appear to be the original intent of that field. Additionally, all MSC codes are joined to one string, which implies that those would be better modeled as an array, which is not allowed by the specification.

We are working on adding further partners to the zbMATH Links API. Three natural candidates are MathOverflow[15], arXiv[16], and the Online Encyclopedia of Integer Sequences[17].

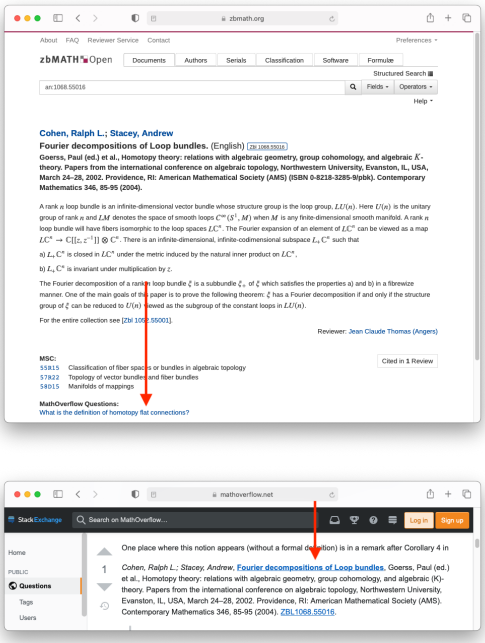

Figure 7: A reference in MathOverFlow (below), and a link to it (above)

MathOverflow is a question-and-answer platform for mathematics that is part of the StackExchange Network[18]. In a previous collaboration, zbMATH and MathOverflow added the possibility to cite zbMATH entries in a MathOverflow post directly, see [11]. The zbMATH citations on the MathOverflow website link to the corresponding zbMATH record. On the zbMATH side, we use the StackExchange API to generate links to MathOverflow questions citing a zbMATH record. This bidirectional linking is shown exemplarily in the two screenshots in Figure 7. These data will soon be added to the zbMATH Links API.

arXiv is one of the most used open-access repositories of electronic preprints in mathematics. Roughly 250k zbMATH records contain links to their specific arXiv preprints that were added manually or thanks to information provided by the publishers. However, many arXiv preprints are still missing. To have access to an arXiv preprint of a zbMATH record is not only important for mathematicians, who might not have access to the journal version, but also to researchers who want to use the available arXiv data, which includes full-texts of many preprints, and combine this data with the metadata from

[15] https://mathoverflow.net/
[16] https://arxiv.org/
[17] https://oeis.org/
[18] https://stackexchange.com/

zbMATH. Therefore, a suitable algorithm is needed to find a corresponding preprint for a zbMATH record if one exists. This problem can be seen as an entity matching problem, and there exists software for it, for example, JedAI, see [12]. For our purpose, the existing software was not suitable. Therefore, we implemented our own matching algorithm. Let us provide a few details about such a matching process, although an accurate and critical description is beyond the scope of this article.

For each search record we generate a small set (default: 3) of possible matching records (called candidates), and compare them with the search record. The candidate records are generated via an Elasticsearch[19] query, where we search for the title and authors of a search record. To decide whether a search and a candidate record match, a three-dimensional feature vector is computed. We use the similarity of the titles, authors, and abstracts as features. The similarity of two titles is their Levenshtein distance divided by the maximum length of the titles. To compare the similarity of two abstracts, we use the cosine distance of their tf-idf vectors (based on words). For the similarity of the authors of two articles we use a more involved approach, which is based on the Levenshtein distance of the author names, but also can handle changes in the order of the author names and incorporates information on different author spellings. Using these feature vectors, we train a decision tree classifier on our training data and test it on some test data using sklearn[20]. If multiple candidates match according to the trained classifier, we take the one whose feature vector has the smallest Euclidean norm.

The training and test data is generated as follows. For every arXiv preprint with a DOI in its metadata we search for a zbMATH entry with the same DOI. If we find one, we add this pair to our ground truth file. We also add some arXiv preprints with a DOI for which no zbMATH entry with the same DOI exists. Finally, we split the ground truth into a training set and a test set. We currently obtain a precision of 99.51 % and a recall of 96.89 % on the test set.

The **Online Encyclopedia of Integer Sequences** is a renowned online database of sequences of numbers launched in November 2010. It currently contains 342.422 sequences, each of them with its own list of metadata: first terms of the sequence, formulas for generating the sequence, references to books, articles, and scholarly links where the sequences have appeared, and more. At present, we are working on retrieving all references listed under "References" and "Links" for each sequence. Such references will be matched with our internal zbMATH Citation Matcher[21] and then stored in our Links API.

[19] https://www.elastic.co/elasticsearch/
[20] https://scikit-learn.org/stable/index.html
[21] https://zbmath.org/citationmatching/

Table 1
Side-by-side comparison of zbMATH Open and PubMed. These are the numbers from 2020

	zbMATH Open	PubMed
Open Access since	2021	1996
Annual Bib. Entries	$>.13$ M	>1.5 M
Bib. Entries Total	>4.0 M	>31.5 M
Journal Titles	>3.0 K	>5.0 K
Search Queries 2020	closed access	>3300 M

4. Research Opportunities

This section presents research opportunities arising from the newly released open data and API solutions at zbMATH in a broader perspective. Moreover, we compare our service with PubMed to put it in a broader context. PubMed, with its underlying MEDLINE dataset and PubMed Central free full-text archive, is another well-known search engine within the biomedical scientific research and digital libraries community [1, 4, 5, 6, 7, 9, 13, 19]. It is available to the public since 1996, indexes over 32 million bibliographic references of biomedical literature, and is supported by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH)[22]. On the other hand, zbMATH Open has over four million bibliographic entries and was made public on 1st January 2021. Table 1 shows a side-by-side comparison of PubMed and zbMATH Open.

We work out strengths and weaknesses by presenting selected research publications that leverage PubMeds APIs and analyze their applicability to the current state of zbMATH. This serves the purpose of uncovering immediate research opportunities in applying existing methods to the new open dataset of zbMATH and highlighting development prospects in areas where existing methods can not yet readily be applied due to missing interfaces or generally missing capabilities. The following paragraphs are to be understood as an inspiration for projects that can be based on the new open-access zbMATH data. After each paragraph, we propose one or multiple research questions that could follow from the described use case.

4.1. Immediate Research Opportunities

In this subsection, we focus on research publications that have leveraged PubMeds open APIs and on general research opportunities.

4.1.1. Tagging of Scientific Publications

Assigning keywords or tags to scientific publications is a crucial tool to increase discoverability. However, assign-

[22] https://pubmed.ncbi.nlm.nih.gov/about/

ing such tags to scientific literature is an expensive and cumbersome process as human reviewers often assign them manually. This, in turn, leads to inconsistencies as different reviewers may assign different tags to the same publications. In [19] Veytsman proposes an automated approach to measure tag consistency across research publications based on a metric that captures how predictive a tag is for a citation. The author conducted experiments based on the MeSH[23] tags that human reviewers manually attach to documents of the PubMed database corpus. He concluded that their simple metric, whether a tag is predictive of citations, indeed can be used to measure tagging consistency. Each indexed publication of zbMATH contains one or many MSC codes[24] and a set of keywords. The former is a hierarchical, alphanumerical identifier indicating the area of mathematics a certain research paper touches and the latter are free-text keywords that the authors suggest. Both classifiers, i.e., MSC codes and keywords, are eventually adjusted by the editors of zbMATH.

We can imagine that the same experiments that Veytsman in [19] carried out can now be done based on the corpus of zbMATH Open. There would even be the possibility to further integrate with MathOverflow and recommend citations based on the tags given in their platform when a post is created.

Potential research questions:

1. **How to measure tagging consistency across mathematical research publications?** Here, one can investigate how the methods developed in [19] can be applied to mathematics data. The required data can be derived via our API.
2. **What can be learned from crowd-sourced tagging in MathOverflow compared to curated tagging in zbMATH?** Especially interesting is here, if the tags from one service can help to search in the other service. The differences in the tagging behavior might also give insights on the learning curve as only known concepts will be tagged by individuals.

4.1.2. PDF Text Extraction Benchmark

As the Portable Document Format (PDF) is the ubiquitous and standard format for scientific publications, its layout-based nature makes it hard to extract semantic meaning from the content. There exist a variety of tools that apply certain heuristics to identify which parts of a document represent, e.g., the title or a paragraph of text. Bast et al. [1] established a benchmark for text extraction performance of 14 tools by taking over 12,000 PDF documents from arXiv and obtaining their semantic information from associated `tex` files and then comparing the outputs of those tools to the semantic information present in the `tex` files. zbMATH Open also provides semantic information in the form of the XML format. While the investigated PDF files also contained some mathematics literature, the idiosyncrasies of mathematical typesetting may be worth a reevaluation with the sole focus on mathematics literature. Here especially the link between zbMATH entries and `tex` sources on arxiv which are provided by the API are helpful.

Furthermore, zbMATH Open provides high-resolution scans of early publications that were not yet typeset in a digital form alongside their corresponding `tex` source files for over 15,000 research article reviews. This corpus constitutes a huge potential for improving optical character recognition (OCR) techniques in the domain of mathematics as outlined in [2].

Potential research questions:

3. **How do the state-of-the-art PDF text extraction tools perform for mathematical literature?**
4. **What are the main challenges in optical character recognition of mathematical formulas?**

4.1.3. Training Dataset

The opening up of zbMATH means that new training data can be used for artificial intelligence applications. The following listing provides inspiration for new possibilities that the dataset could be used for:

Formula Search The search mask of zbMATH Open already offers a formula search. However, the new open API allows building ones own or improving the formula search functionality by leveraging meta information provided alongside with the indexed articles.

Potential research questions:

5. **What influence do different search options in digital libraries have on the scientific discovery process?** It is save to assume, that the discovery options for scientific literature will have an effect on the outcomes on ones own research. Here, one could try to qualitatively or even quantitatively assess this influence.
6. **What are the state-of-the-art approaches to formula search, and what are the main challenges to overcome?**

Recommender Systems The provided data allow building a comprehensive recommendation system. This system could incorporate not only the meta information of the OAI-PMH APIs like MSC tags or keywords but also leverage the information on other platforms that a certain research article is linked in. E.g., mentions of related research papers in conversations on MathOverflow may be a good indicator for other relevant literature. As we

[23] https://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/index.html

[24] Mathematics Subject Classification 2020,https://msc2020.org/

continue to attract more and more partners for our Link API the context increases from which a potential recommender system can draw meaningful conclusions.

Potential research questions:

7. **Which features are most significant for related literature recommendations in mathematics?**
8. **What are the distinguishing challenges in feature extraction from mathematical literature?** The challenge of this research question is to identify how state-of-the-art recommender systems of other disciplines need to be tuned to excel at mathematical literature recommendations.

Formula Disambiguation I Similar formulas can have vastly different meanings in different contexts [14, 15, 16, 17]. This is especially true for single symbols used in these formulas as researchers in different fields will certainly have assigned a different meaning to symbols. A system that tries to understand in which context a formula appears and draw meaning from that could especially leverage the MSC classification that is assigned to all articles on zbMATH Open. Most results from the OAI-PMH API contain an abstract where one can often find typeset formulas that can be used as training data along with full-text data that can be obtained through arXiv.

Potential research questions:

10. **How can similarly typeset formulas describing different concepts be disambiguated?** The main challenge of this research question is to devise criteria that make a formula ambiguous.
11. **What are the distinguishing factors in formula typesetting to avoid ambiguity?** In this research question it would be the goal to devise guidelines to avoid typesetting ambiguous formulas in the first place.

Formula Disambiguation II Following the above disambiguation, it is also possible for a single concept to be expressed in different ways. Imagine the circumference U of a circle being expressed in one paper as $U = 2\pi r$ and in another $U = \pi d$ with radius r and diameter d. Indeed, both formulas describe the same concept but are typeset differently. This kind of disambiguation will be of immediate relevance for academic plagiarism detection. State-of-the-art plagiarism detection systems already consider paraphrased text but lack capabilities to effectively detect "paraphrased" formulae [10].

Potential research questions:

12. **How can differently typeset formulas describing the same concept be disambiguated?** The main challenge of this research question is to devise ways to identify such formula combinations.
13. **What factors make a formula more readable than a differently typeset formula describing the same concept?** Here, one can investigate factors for readability and if there are objectively better ways to typeset a certain formula.

Math Spell-Checking Popular tools like Grammarly[25] scan your text for common grammatical mistakes and provide the user hints about potential improvements. A similar offering could be developed for typesetting formulas by, for example, giving simple warnings of missing closed parentheses (if applicable) or other common mistakes. Such a spell-checking system could make use of the data of zbMATH Open and linked peripheral services. The linking to arXiv could be used to retrieve the full-text `tex` information, and the connection to MathOverflow could be used to detect common mistakes by taking into account the edit history of formulas in posts.

Potential research questions:

14. **What are common errors in mathematical formula typesetting, and how to identify them?** The main challenge of this research question is to derive a method to identify erroneous formulas; and as a second step to investigate what common errors are.
15. **What impact had formulas containing errors in the mathematics research community?** Here, one can research the consequences that errors in formulas and the research that built on them had. This could be extended to the influence of errors in formulas on widespread websites like Wikipedia to contemporary incidence.

Classification and clustering While zbMATH Open provides MSC tags and keywords for the research articles, we can imagine that there are different classification and clustering approaches that are not represented through the meta information of zbMATH. The open-access to the APIs allows building use case specific search and clustering systems.

Potential research questions:

16. **Do different logical classification and clustering schemes emerge from the zbMATH Open metadata besides the MSC classification scheme?**

Review generation At present, many research papers and books indexed at zbMATH are supplemented with a review written by external experts in the field. Currently more than 7,000 active experts participate in compiling reviews for research papers and books. They critically analyze the contribution of the publication under consideration, often summarize the content and judge it in

[25] https://grammarly.com/

reference to a bigger context. With the advancements of text generating deep learning models such as language models, it is not far to seek to train models on these handwritten reviews in conjunction with their full-text articles and metadata of zbMATH Open.

Potential research questions:

17. **What are the significant properties that a mathematical review should include?** In this research question one should distill the essential properties of what makes a "good" mathematical review.
18. **How do mathematical reviews generated by AI language models compare with manually written reviews according to the aforementioned significant properties?** Here, it is interesting to understand if artificial intelligence is capable of meeting the aforementioned properties.
19. **What impact can AI language models have on the mathematical review process?** In this research question, one should work out the implications of potentially machine written reviews.

4.2. Development Prospects

In this subsection, we focus on research publications that have leveraged PubMeds open APIs to which there is no pendant yet in zbMATH Open. The uses-cases in this section serve as inspiration for development opportunities.

4.2.1. Retraction Tracking

There are manifold reasons why a scientific publication could get retracted. It can range from erroneous study design to deliberate misconduct like plagiarism or generating artificial data to support a hypothesis. With the increasing amount of scientific literature at an accelerating rate, the number of retracted papers naturally increases as well. Therefore, it is crucial to notify researchers early in the research process about possible retracted publications. In [4] Dinh et al. present a Zotero[26] plugin called *ReTracker* that helps to identify retracted papers from PubMed. *ReTracker* uses the full paper titles as they are present in the Zotero library to query PubMed on its retraction status. This status is persisted in a local cache and displayed to the user. With the opening of zbMATH this plugin could now not only cover articles of biomedical literature but to also inform researches about retracted publications in the field of mathematics. Currently, zbMATH Open does not provide information about the retraction status, but we can imagine that collecting this information from various trustworthy sources and making it accessible through the API would be a valuable addition to the current service. The authors in [4] underline the need for such a tool by stating that the citation rate of retracted publications can even increase after they got their retraction status [4], so, literature is still cited even years after retraction.

Potential research questions:

20. **How does the retraction of mathematical papers influence their citations?** This question follows the observation of [4] that the citation count of literature still increases after it got retracted, so the intuitive answer that citations stop after retraction does not hold true. Here, it would be interesting to identify the reasons why literature is still cited.
21. **What are the most common reasons for the retraction of mathematical research papers, and how can publication of such papers be minimized?** Here, one can think in the direction of computer assisted quality assurance on the publisher side and how this could help the publishing process.

4.2.2. Collaboration Identification

While digital libraries nowadays offer comprehensive and advanced search interfaces to retrieve and explore related scientific literature, they often lack the understanding of how authors have collaborated and to which extent their collaboration was fruitful. The same statement is true for zbMATH Open. In [3] Cagliero et al. explored ways to identify collaboration patterns of authors and to measure to what extent the collaboration was fruitful. They harvested digital libraries and online databases for research publications and applied a pattern-based approach to identify collaborations among researchers. By making the APIs of zbMATH open-access, we believe that Cagliero et al. [3] can serve as inspiration to motivate further insights generation techniques like author collaboration identification.

Potential research questions:

22. **How can the open data of zbMATH be used to construct collaboration graphs among mathematics researchers?** The main contribution in this research question would be a comprehensive collaboration graph based on the zbMATH open dataset.
23. **What conclusions can be drawn from an author collaboration graph concerning collaboration effectiveness?** Here, one can investigate how the methods developed in [3] can be applied to the data of our APIs.

5. Conclusions and Future Work

In this article, we have presented the recent innovations made to zbMATH. We implemented API solutions following the OAI-PMH and Scholix standards. Those solutions allow the scientific community to use our open database

[26] https://www.zotero.org/

in an efficient and reproducible way. We demonstrated the capabilities of API solutions on the basis of existing links between DLMF and zbMATH. By combining classification information from zbMATH with reference information from DLMF, we could derive new insights on references in the DLMF. In the future, we will incorporate MathOverflow, arXiv, and the Online Encyclopedia of Integer Sequences to the new zbMATH Links API. Moreover, we gave inspiration for research opportunities arising from the APIs. In this context, we proposed 23 open research questions that can be immediately approached by leveraging the open access model and new programming interfaces.

We will optimize our API interfaces to the needs of the scientific community and zbMATHs data partners in the future. Depending on the needs of the communities, we will evolve and adapt our data formats. Moreover, we are working for open access publications and permissive licenses for the reuse of scholarly metadata. We aim to convince publishers to distribute abstracts and references under permissive licenses. We will also continue to integrate mathematics-related research software and research data besides traditional publications.

References

[1] H. Bast and C. Korzen. "A Benchmark and Evaluation for Text Extraction from PDF". In: *Proc. ACM/IEEE JCDL*. Toronto, ON, Canada: IEEE, June 2017, pp. 1–10. DOI: 10/ghchxm.

[2] M. Beck et al. "Transforming Scanned zbMATH Volumes to LaTeX: Planning the Next Level Digitisation". In: *EMS Newsletter* 2020-9.117 (Sept. 2020), pp. 49–52. DOI: 10.4171/news/117/11.

[3] L. Cagliero et al. "Identifying Collaborations among Researchers: a pattern-based approach". In: *Proc. BIRNDL at ACM SIGIR*. Ed. by P. Mayr, M. K. Chandrasekaran, and K. Jaidka. Vol. 1888. CEUR-WS.org, 2017, pp. 56–68.

[4] L. Dinh, Y.-Y. Cheng, and N. N. Parulian. "ReTracker: an Open-Source Plugin for Automated and Standardized Tracking of Retracted Scholarly Publications". In: *Proc. ACM/IEEE JCDL*. Ed. by M. Bonn et al. IEEE, 2019, pp. 406–407. DOI: 10.1109/JCDL.2019.00092.

[5] S. Eggers et al. "Visualizing aggregated biological pathway relations". In: *Proc. ACM/IEEE JCDL*. 2005, pp. 67–68. DOI: 10.1145/1065385.1065400.

[6] T. Erekhinskaya et al. "Knowledge Extraction for Literature Review". en. In: *Proc. ACM/IEEE JCDL*. Newark New Jersey USA: ACM, June 2016, pp. 221–222. DOI: 10.1145/2910896.2925441.

[7] J. M. González Pinto, J. Wawrzinek, and W. Balke. "What Drives Research Efforts? Find Scientific Claims that Count!" In: *Proc. ACM/IEEE JCDL*. 2019, pp. 217–226. DOI: 10.1109/JCDL.2019.00038.

[8] K. Hulek and O. Teschke. "The Transition of zbMATH Towards an Open Information Platform for Mathematics". In: *EMS Newsletter* 2020-6.116 (June 2020), pp. 44–47. DOI: 10.4171/news/116/12.

[9] K. Jhawar et al. "Author Name Disambiguation in PubMed using Ensemble-Based Classification Algorithms". In: Aug. 2020, pp. 469–470. DOI: 10.1145/3383583.3398568.

[10] N. Meuschke et al. "Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations". In: *Proc. ACM/IEEE JCDL*. Urbana-Champaign, Illinois, USA, June 2019. DOI: 10.1109/JCDL.2019.00026.

[11] F. Müller, M. Schubotz, and O. Teschke. "References to Research Literature in QA Forums – A Case Study of zbMATH Links from MathOverflow". In: *EMS Newsletter* 2019-12.114 (Nov. 2019), pp. 50–52. DOI: 10.4171/news/114/15.

[12] G. Papadakis et al. "The return of jedAI: end-to-end entity resolution for structured and semi-structured data". In: *Proc. VLDB* 11.12 (Aug. 2018), pp. 1950–1953. DOI: 10.14778/3229863.3236232.

[13] H. Saggion and F. Ronzano. "Scholarly Data Mining: Making Sense of Scientific Literature". In: *Proc. ACM/IEEE JCDL*. Toronto, ON, Canada: IEEE, June 2017, pp. 1–2. DOI: 10.1109/jcdl.2017.7991622.

[14] 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. DOI: 10.1145/3442442.3452348.

[15] P. Scharpf, M. Schubotz, and B. Gipp. "Representing Mathematical Formulae in Content MathML using Wikidata". In: *BIRNDL@SIGIR*. Vol. 2132. CEUR-WS.org, 2018, pp. 46–59.

[16] P. Scharpf et al. "*AnnoMath TeX* - a formula identifier annotation recommender system for STEM documents". In: *RecSys*. ACM, 2019, pp. 532–533.

[17] P. Scharpf et al. "Towards Formula Concept Discovery and Recognition". In: *BIRNDL@SIGIR*. Vol. 2414. CEUR-WS.org, 2019, pp. 108–115.

[18] M. Schubotz and O. Teschke. "zbMATH Open: Towards standardized machine interfaces to expose bibliographic metadata". In: *EMS Newsletter* 2021-4 (2021). DOI: DOI10.4171/MAG-12.

[19] B. Veytsman. "How to Measure the Consistency of the Tagging of Scientific Papers?" en. In: *Proc. ACM/IEEE JCDL*. Champaign, IL, USA: IEEE, June 2019, pp. 372–373. DOI: 10.1109/jcdl.2019.00076.