Short note: zbMATH Open

Klaus Hulek

A long term project has finally become reality: zbMATH has become an open access database as of 1st January 2021, allowing free usage worldwide. This is the result of a process which has lasted several years and our thanks go to all the individuals and institutions who have made this possible.

The mathematical community is invited to participate in the future development of the database. We will now be open to share data and links with other non-commercial databases. This opens up the way to new cooperations and, we hope, novel and potentially unexpected new developments in the future. We are currently working on an API which will allow much of our data for research to be downloaded for research and non-commercial purposes. Please share you ideas on the future of the database with us via editor@zbmath.org. And of course, we are always looking for new reviewers: you can register via our website: zbmath.org/become-a-reviewer/

Last but not least: we thank Springer Verlag for many years of good cooperation. zbMATH would not exist without Springer Verlag. At the same time the landscape of publishing is undergoing fundamental changes and we believe that the new model is the right direction to follow in the future.

—

Klaus Hulek is professor of mathematics at Leibniz University Hannover and Editor-in-Chief of zbMATH Open. His field of research is algebraic geometry.

hulek@math.uni-hannover.de

DOI 10.4171/MAG-11

zbMATH Open: Towards standardized machine interfaces to expose bibliographic metadata

Moritz Schubotz and Olaf Teschke

In this article, we give motivation for the need for standardized machine interfaces to zbMATH open data, outline the target audience, describe our preliminary strategy to develop API interfaces, and report on the details of the first interface we implemented.

1 Target audience

As announced in the previous note, zbMATH is becoming open access from the 1st of January 2021.[1] For most working mathematicians, this means that they can access zbMATH from anywhere in the world without a subscription or authentication. Additionally, we envision benefits to the community through our efforts to connect zbMATH data with information systems of research data, collaborative platforms, funding agencies, and intradisciplinary efforts, as outlined in [2]. We expect that our efforts to disseminate the results of mathematical research will provide this research with increased visibility. However, to target domain-independent information systems, we need to comply with standardized information exchange protocols and interfaces.

In what follows, we describe potential partners that might interact with zbMATH. We will offer the data via so-called Application Programming Interfaces (APIs). Moreover, in this report, we focus on how others can make use of zbMATH open data, rather than how zbMATH can use other data sources. As depicted in Figure 1, the potential consumers can be clustered into at least five groups, which we will describe below.

[1] The open web platform is now available under the name zbMATH Open, while we will address the zbMATH content and services under the traditional umbrella name zbMATH for convenience.

Bibliographic consumer	MathOverflow Wikimedia arXiv Zotero	a_1 Selection of individual items a_2 High throughput a_3 End-user-friendly formats a_4 Various representations a_5 Fuzzy search
Aggregators	OpenAIRE/ERC NFDI/DFG	b_1 Standard compliance b_2 Incremental updates b_3 Projection on properties
Archives	Software Heritage Internet archive	c_1 Fetch everything c_2 Reduce traffic c_3 Traceability of versions
Search engines	Firefox search plugin	d_1 Selection of individual items d_2 High throughput d_3 End-user-friendly formats d_4 Various representations d_5 Fuzzy search d_6 Formula search
Individuals	Blog on specific topic Personal reference list	e_1 Easy to setup e_2 Long term stability

Figure 1. Envisioned consumer (left) and desiderata (right) [5]

Bibliographic consumers are information systems that display references to scientific publications. They often deal with user-generated content that references individual research articles. For websites like Wikipedia or MathOverflow, users interactively search for references to support their statements. The remote information system, e.g., MathOverflow, sends the user's search-string to a designated zbMATH API endpoint, which then returns a ranked list of possible references. The remote information system takes care of the formatting. While MathOverflow, for instance, might use zbMATH exclusively, others, such as Wikipedia, might want to fuse results from zbMATH with results from other providers of bibliographic metadata. Standardized protocols drastically reduce the implementation effort for intra-domain information systems. Even before the transformation to zbMATH open, we provided a simple API for MathOverflow [4], which was limited to the top three search results. This legal restriction has now vanished. In contrast, to interactive bibliographic customers described before, arXiv and other publishers might use zbMATH's bibliographic metadata to disambiguate references, which is an essential prerequisite for many information retrieval tasks such as recommendations, semantic searches, or plagiarism detection [3].

Aggregators such as the OpenAIRE research explorer, SemanticScholar, DataCite, or Altmetric extract information from different sources, transform them to standardized representations and load them into their specific data models. Additionally, in some countries, researchers are also required to report their publications to government platforms. At the end of the process, funding agencies or other decision-makers can use these data sources for so-called data-driven decision making. Here standardized interfaces and formats evolved to simplify the aggregators' job, as crawling through web-pages optimized for human consumption is error-prone and involves complicated heuristics that are fragile and vulnerable to layout changes.

Archives such as the Internet Archive and Software Heritage capture the digital history in the forms of websites or software code. Since their mission is digital preservation, an API that enables replaying the entire history of the website would be ideal. Moreover, they strive to reduce traffic consumption and avoid redundancy. Their infrastructure is optimized to preserve HTML websites in the form presented to a user at a particular point in time.

Search Engines might use our API to present search results in a different format. For instance, Mozilla Firefox has a built-in feature to include custom search engines that implement the OpenSearch standard. One interesting feature to consider is if and how mathematical formulæ are represented in OpenSearch.

Individuals or small groups of people with particular needs are of particular importance to us. We aim to provide as much support as possible to research groups, either from mathematics or from the field of bibliometric research. Highly motivated individuals who aim to use our data creatively are also on our schedule. Here, we are open to requests, and need to investigate potential uses case-by-case. A typical, not too exceptional use case we envision would be to set up a personal publication list or to enrich a personal website with the latest news of specific Mathematics Subject Classification (MSC) classes. While many of these functionalities are already possible with zbMATH's news-feed functionality, we expect the API functionality to be more flexible.

2 First steps towards APIs

Given the diverse expectations and needs described above, we do not see a one-size-fits-all solution that would fulfill the diverse requirements. Therefore, we decided to pursue an iterative approach to building API solutions. As a first step, we aim to start with a first API version that is well-established, easy to implement, and has a substantial positive impact on working mathematicians. According to our analysis, aggregators, archives, and bibliometric researchers commonly use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH seems to be well-established, sufficiently documented, and relatively easy to implement. This protocol is also well-suited to downloading, i.e. to harvesting the entire open collection of zbMATH document data. These data come along with a CC-BY-SA license, which facilitates both reusability and allowing derived work to remain in the open ecosystem, although this comes at the cost that some third-party content (such as abstracts) is not included due to legal constraints. Additionally, one may harvest well-defined subsets and consume updates since the last download without requiring to redownload a dump. We expect that this format will also be well-suited to individuals working with zbMATH data. Especially consumers that work with other datasets besides zbMATH will appreciate the standardized functionality of the protocol. However, the format is not well-suited for bibliographic consumers, given the overhead caused by the standard and the lack of flexibility. Because of this, we decided to create at least two APIs, with the OAI-PMH API as a starting point, and other more flexible APIs to be determined.

In Figure 2, we display a possible scenario for zbMATH's future API development efforts. The blue boxes (Reviewer Interface, Internal Interfaces, zbMATH database, and zbMATH Website) show the well-established components of zbMATH. The dark gray box (OAI-PMH API) shows the newly released API described in this paper. With this setup, all write operations to the database will be performed from the reviewer interface and other internal interfaces. In contrast, the Website and the OAI-PMH interfaces are read-only interfaces that present the zbMATH database's contents without modifying it.

As a next step, we are working on an API to create links from zbMATH to external sources and vice versa. One commonly accepted format to describe these connections is the Scholix format. Therefore we labeled this project "Scholix Link API" in Figure 2. Independently of this task, we are also working on a general-purpose API (Gray box Custom API in the figure) that replicates the current website's functionality but produces the results in a better machine-processible form, such as JSON instead of HTML. In theory, one might use this API for a far-future version of the zbMATH website, given that efficient caching layers are implemented. While we are pursuing the linking and custom API efforts in parallel, our goal is to limit as far as possible the number of distinct API endpoints. Another vital link will be a bidirectional link to research data in the context of the German Mathematical Research Data Initiative (MaRDI), a consortium formed for applications within the National Research Data Infrastructure (see [1]). In the MaRDI project (light gray box at the top), we plan to repurpose the generic WikiBase knowledge graph software that supports many well-established structured graph data exchange protocols, such as RDF and SPARQL among others.

3 Implementation

Given the above motivations, we have implemented a first version of the OAI-PMH interface. The current demo is available from purl.org/zb/10. As required by the protocol, our OAI-PMH api offers six endpoints, namely (1 Identify, 2 ListMetadataFormats, 3 ListSets, 4 ListIdentifiers, 5 ListRecords, 6 GetRecord):

- Endpoint 1 helps aggregators and archives to discover the new API fully unsupervised, identify which version of the OAI-PMH standards we are using, and other technicalities.
- Endpoint 2 lists the formats that we use to expose zbMATH data. We implemented two flavors. The first is the required standard Dublin Core Metadata Record format, which contains standardized fields like abstract, publisher, creator, or title. For the second, we implemented a format that is closer to zbMATH's internal data model. Many domain-specific classifications can be expressed in terms of Dublin core vocabulary. However, the MSC is not predefined in the Dublin core standard, even though it seems to us that it could be modeled. However, expressing all the details of zbMATH's data in Dublin core terms would require an immense effort of coordination with librarians to ensure that our encodings are modeled according to common best practices for modelling specifics in Dublin core. In other words, we are addressing the issue from two ends. With the DC standard, we encode the low hanging fruits in a standard way. With our additional zbMATH custom format, we ensure that we expose all the data we are legally allowed to by the API.
- Endpoint 3 lists the subsets of the zbMATH dataset that we think could be relevant. In the first version, we provide the following sets: document type, year, document author, classification, keyword, document language, author variation, author reference, biographic reference, software, review type, review language, reviewer, serial publisher. For example, the set `document_author:Noether, Emmy` is the subset of all zbMATH entries authored by Emmy Noether. As an extension to OAI-PMH's built-in set logic, we implemented magic characters `|&~` that indicate the standard set operations 'or', 'and', and 'not' respectively, allowing users to combine sets at will. Obviously, in endpoint 3, we only enumerate the 1 125 144 base sets.
- Endpoints 4 and 5 list the currently 4 206 870 list zbMATH identifiers and records, respectively. This endpoint is predestined to

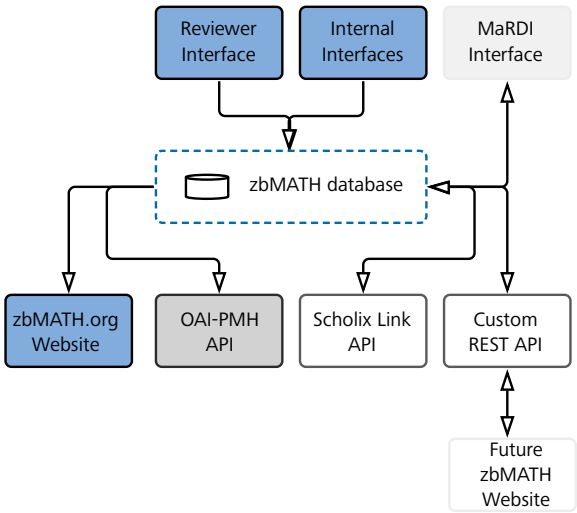

Figure 2. A conceptual overview of the zbMATH database and the data flow from an to the database

obtain a dump of all public zbMATH open data. Here OAI-PMH's built-in cursor and resumption mechanism ensure an efficient and seamless retrieval of the data. For the convenience of end-users, one can use one of the many available OAI-PMH metadata harvesters from www.openarchives.org/pmh/tools/ to retrieve all the data.

- Endpoint 6 gets individual zbMATH entries.

To conclude with a real example, assume that we want to retrieve the OAI-PMH metadata for the entry with zblnumber 1200.35057. We would first need to retrieve the corresponding internal identifier (DE number), which can be done by clicking on the BibTeX button below the article. In this example, a BibTeX entry with key zbMATH05797851 will be downloaded open. The last digits following the word zbMATH, i.e., 5797851, are the DE number. One can then use this number to retrieve the metadata from the API by appending the query

```
verb=GetRecord&identifier=oai:zbmath.org:
5797851&metadataPrefix=oai_zb_preview
```

to the root of the API endpoint in the browser. Here "verb" identifies the endpoint (6=GetRecord), and the DE number is prefixed with identifier prefix and postfixed with the desired metadata format. The browser will then display a large XML file that contains the review text and other public information available on the zbMATH website. See purl.org/zb/11 for comparison to the website view. One can use this method to obtain any document from the zbMATH open database without downloading large sets of articles.

4 Conclusion

We have introduced the target audience of our API, discussed our strategy of rolling out APIs to cover a wide range of potential users, and described details of our API infrastructure's first pillar. While our plans for future endpoints are subject to change and the current OAI-PMH endpoint is subject to continual improvement, we have taken the first step towards standardized machine interfaces to make the data of zbMATH available to a broader audience.

References

[1] K. Hulek, F. Müller, M. Schubotz and O. Teschke, Mathematical research data – an analysis through zbMATH references. *Eur. Math.Soc. Newsl.* **113**, 54–57 (2019)

[2] K. Hulek and O. Teschke, Die Transformation von zbMATH zu einer offenen Plattform für die Mathematik. *Mitt. Dtsch. Math.-Ver.* **28**, 108–111 (2020)

[3] N. Meuschke, V. Stange, M. Schubotz, M. Kramer and B. Gipp, Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, IL, USA, 120–129 (2019)

[4] F. Müller, M. Schubotz and O. Teschke, References to research literature in QA forums – a case study of zbMATH links from MathOverflow. *Eur. Math. Soc. Newsl.* **114**, 50–52 (2019)

[5] M. Schubotz, D. Trautwein and O. Teschke, zbMATH is open: A practical guide to open an informationservice. *Proceedings of The Open Science Conference 2021 (OSC '21), February 17–19, 2020*, online

Moritz Schubotz is a senior researcher for mathematical information retrieval and open science. He maintains the support for mathematical formulæ in Wikipedia and is off-site collaborator at NIST.

moritz.schubotz@fiz-karlsruhe.de

Olaf Teschke is Managing editor of zbMATH and Vice-chair of the EMS Committee on publications and electronic dissemination.

olaf.teschke@fiz-karlsruhe.de