Klaus Hulek
Octavio Paniagua Taboada
Olaf Teschke *(editors)*

90 Years of zbMATH

EMS
PRESS

Klaus Hulek
Octavio Paniagua Taboada
Olaf Teschke *(editors)*

90 Years of zbMATH

EMS PRESS

Editors

Klaus Hulek
Institut für Algebraische Geometrie, Fakultät für Mathematik und Physik
Leibniz Universität Hannover
Welfengarten 1, 30167 Hannover, Germany

Email: hulek@math.uni-hannover.de

Octavio Paniagua Taboada
Subject-specific services, Department of Mathematics
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin
Franklinstraße 11, 10587 Berlin, Germany

Email: octavio.paniaguataboada@fiz-karlsruhe.de

Olaf Teschke
Subject-specific services, Department of Mathematics
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin
Franklinstraße 11, 10587 Berlin, Germany

Email: olaf.teschke@fiz-karlsruhe.de

2020 Mathematics Subject Classification: Primary 01A74; Secondary 01-06, 01A60, 01A61, 01A65

Keywords: open access, zbMATH Open, swMATH, gender gap, reviewer community, API solutions, digital math libraries, digital shadow, formula search, author identification, interconnectivity, community platforms, bibliometrical tool

ISBN 978-3-98547-073-0, eISBN 978-3-98547-574-2, DOI 10.4171/ZBL90

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Published by EMS Press, an imprint of the

European Mathematical Society – EMS – Publishing House GmbH
Institut für Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin, Germany

https://ems.press

Typeset by EMS Press
Printed in Germany
♾ Printed on acid free paper

Preface

zbMATH Open – a personal introduction

When I started working on my PhD in 1977, it was the done thing to regularly spend time in the library in order to browse through the new journals and books that had arrived. Typically, I set Friday afternoon aside to go there and systematically browse through the newly delivered publications. This, together with the preprints which my supervisor (and later I myself) received, were, at least at the beginning of my career, my primary source of information.

Spending my afternoons in the library of the Institute in Erlangen, I was struck by two large bookcases. One contained an impressive array of yellow volumes, the other a similar collection of red books. These were the *Zentralblatt für Mathematik und ihre Grenzgebiete*, as its full title was then, in short the *Zentralblatt*, and its counterpart, the *Mathematical Reviews*. Whenever a new volume of either arrived, I checked the relevant sections in search of yet undiscovered articles which might be relevant to my research.

Studying the title page of Zentralblatt, I was struck particularly by one detail: Zentralblatt was edited jointly by the Heidelberg Academy of Sciences and the (East German) Akademie der Wissenschaften. This cooperation was indeed ended by East Germany in 1977, but the fact that it had existed until then came as a big surprise. And also now, in retrospect, this is very remarkable. After all, the wall had been built in 1961 and by the mid 1970s the two German states had developed very different political structures; there were virtually no areas in which they collaborated. Amazingly, the cooperation of mathematicians remained intact long into the Cold War.

As the years went by, the importance of Zentralblatt and Mathematical Reviews for my own research decreased. This was for a variety of reasons. My mathematical network had grown considerably more extensive and, consequently, I was sent many more preprints by colleagues. Also, it had become much easier to travel in order to attend conferences or collaborate with colleagues. This in turn opened up many new ways for obtaining re- and preprints and up-to-date information.

The arrival of email and then the Internet in the early 1990s made a tremendous difference. My own field of research (algebraic geometry) was among the first branches in mathematics to use the arXiv systematically for the dissemination of new mathematical literature. Naturally, this also affected my use of Zentralblatt; it did not become obsolete, but its role shifted. Zentralblatt became less important for finding out what was new, but remained very useful for tracing and evaluating mathematical literature (after all only a small part of the mathematical literature was available electronically at that time). This applied in particular to research areas which were further away from my own field of expertise.

The fact that the databases were available online meant also that it became very easy to access the information at any time and from virtually any place. Thus, the regular visits to the library to check the recent editions of journals, or to check Zentralblatt and Mathematical Reviews, all but came to an end. I must also confess that at that time I more or less lost sight of Zentralblatt for a while. Both services, Zentralblatt and Mathematical Reviews, went online at about the same time – in fact Zentralblatt was slightly earlier – but due to the close connection of the AMS with the mathematical community, in particular in the US, MathSciNet, the online database of Mathematical Reviews, was more successful in the early years of digitisation in establishing its web presence. But: zbMATH has clearly been catching up steadily and the fact that we are now open access will further accelerate this process.

I started to hear more about zbMATH, as it had been renamed by then, when Gert-Martin Greuel, whom I know well as a mathematical colleague, became Editor-in-Chief in 2012. He often talked to me about zbMATH and MathReviews and he argued very forcefully that the mathematical community can only gain from having two reviewing services available. The competition of the two services clearly helps to improve the performance of each, and thus benefits the mathematical community. At this point, I would like to emphasize that the relationship between the two databases, although they are natural competitors, has, in my experience, always been a cordial one. The joint development of the mathematics subject classification MSC 2020 is just one proof of this.

It came as a big surprise to me when FIZ Karlsruhe and Springer approached me early in 2015 to ask whether I would be prepared to take over as Editor-in-Chief of zbMATH from 2016. At that time, I was spending half a year at the IAS in Princeton, having just finished a 9-year period as Vice President for Research at Leibniz Universität Hannover. My plan after that had been to give all administrative tasks a wide berth and to concentrate purely on research. For this reason, I was at first quite reluctant to accept the offer. On the other hand, during my time as Vice President in Hannover, I had come into close contact with questions concerning the future of publishing, open access and other topics such as research data. Based on this experience I felt that the position of Editor-in-Chief would be both interesting and challenging. And I strongly believed that zbMATH should be supported and developed further. So, I finally decided to accept the offer to become Editor-in-Chief of zbMATH.

After taking up my new position, I noticed just how sophisticated a database zbMATH had become. I was also very soon confronted with crucial questions about the future direction zbMATH should take, and not least what an appropriate business model could be. We quickly came to the decision, which had already been considered by Gert-Martin Greuel, that the best way forward was to go open access. Needless to say, extra resources were required for this, and we started the long process of applying for a suitable grant in the context of the evaluation of FIZ Karlsruhe, an exercise which takes place every 7 years. With the help of many, the application was finally successful and we were indeed able to go open access on 1st January 2021. I still believe that this is a great step forward and that this will consequently enable zbMATH Open, as it has now been renamed, to realise its full potential. There are many exciting new challenges, including the whole realm of mathematical research data, and non-textual information which zbMATH Open will have to address, and this will only be possible with the close involvement of the mathematical community. The last 90 years of its history, when Zentralblatt metamorphosed into zbMATH Open, show that we are able to adapt to new conditions and environments without losing sight of our main goal, namely that of providing high quality information on a very wide range of mathematical publications in all the different formats this may take. I strongly believe that zbMATH Open will become an even more important tool for the working mathematician in the future.

Klaus Hulek

Greetings from the president and former president of the European Mathematical Society

On behalf of the European Mathematical Society (EMS) we congratulate zbMATH on its 90 anniversary. The EMS is very proud to be one of the three partners that publish zbMATH. Although we were not in this partnership 90 years ago (the EMS is just about 33 years old), we consider this as one of our most important tasks. The distribution of mathematical knowledge, the recording of what is known, to make clear who did what and when, and to make the scientific developments (the content of publications and software) openly available to the whole community, is essential for the well-being and future progress of the mathematical sciences. This is the reason why we are in the publishing team, and even increased our role when zbMATH turned open access, which we think is a great move that will have a major impact. We follow this direction and our publishing house EMS Press is now publishing under the subscribe-to-open concept.

In a human life, someone of 90 years is very old, but looking at zbMATH we see a young and very energetic teenager with a lot of revolutionary ideas and visions. And the best of all, we see the strengths to make these ideas reality. One of these ideas is to join the initiative to make our scientific data, such as publications, software, and bibliometric data to be FAIR (findable, accessible, interoperable, and reusable). For this reason we are very happy to support also the zbMATH initiative to join the German National Data Infrastructure.

We wish that zbMATH maintains its great momentum and we are happy to stay a strong partner in its future endeavours.

Jan Philip Solovej, president of the EMS
Volker Mehrmann, former president of the EMS

Congratulations from the Heidelberg Academy of Sciences and Humanities

1931 saw the publication of the first volume of the "Zentralblatt für Mathematik und ihrer Grenzgebiete", a new mathematical review journal founded by professors Richard Courant and Otto Neugebauer (Göttingen), Harald Bohr (Copenhagen), and the publisher Ferdinand Springer. Their aim was to provide a more up-to-date, comprehensive account of progress in mathematics and related disciplines, and to improve international scientific communication. In its history, Zentralblatt für Mathematik has undergone a remarkable series of transformations.

The first Editorial Office was established on the premises of the Springer publishing house in Berlin. The start under the first Editor-in-Chief of Zentralblatt, Otto Neugebauer, was promising. However, the Nazis' rise to power in Germany in 1933 and its devastating global consequences also had a major effect on Zentralblatt. Courant fled Germany for the USA in 1933 to escape the Nazi regime. Neugebauer followed in 1939, after a period with Bohr in Copenhagen.

As Zentralblatt came under increasing political pressure and its independence was under threat, the American Mathematical Society founded Mathematical Reviews, instigated by Neugebauer and supported by Courant. Its first volume appeared in 1940. Mathematical Reviews now serves the mathematical community in its electronic version MathSciNet, a database which is in many ways comparable to zbMATH.

In Germany the management of Zentralblatt was taken over in 1939 by the Prussian Academy of Sciences in conjunction with the German Mathematical Society. It was at this stage that Academies started to play an essential role for Zentralblatt.

The Prussian Academy was reopened in 1946 as the German Academy of Sciences. Together with Springer, this academy relaunched Zentralblatt in 1947. What followed was a remarkable German-German collaboration. This was severely affected by the building of the Berlin wall in 1961. The Editorial Office of Zentralblatt and the German Academy of Sciences were located in the Eastern part of Berlin, and thus fell under the rule of the GDR government. An additional Editorial Office was consequently established in West Berlin, again on the premises of Springer, and manuscripts were taken physically from one part of Berlin to another — a trip across the Iron Curtain. It was at this stage that the Heidelberg Academy of Sciences and Humanities took on responsibilities for Zentralblatt. The Heidelberg Academy (based in the West) and the German Academy of Sciences in the German Democratic Republic (in the East) began to edit Zentralblatt jointly, while Springer was responsible for printing and distribution. The GDR government finally ended this cooperation in 1977.

The Heidelberg Academy of Sciences and Humanities was happy to continue its involvement in Zentralblatt after 1977. The increasing role of electronic tools began to fundamentally change the landscape of publishing. This led to the foundation of FIZ Karlsruhe, now operating as “Leibniz Institute for Information Infrastructure”, which became responsible for the Editorial Office of Zentralblatt. In the 1990’s a major effort was undertaken to put Zentralblatt, whose electronic version was renamed zbMATH, on a more European level, with the European Mathematical Society (EMS) as a key player. In particular, the European Mathematical Society and the Heidelberg Academy of Sciences and Humanities share the task of scientifically supervising both the technical processing of the data and the development of the tools for information processing. Thus, starting in 1999, Zentralblatt has had three editorial institutions: the EMS, FIZ Karlsruhe and the Heidelberg Academy of Sciences and Humanities, with Springer responsible for marketing and distribution (until 2020).

The Heidelberg Academy of Sciences and Humanities enthusiastically supported the move to make zbMATH open access. In January 2021, this became reality thanks to a decision by the German Joint Science Conference (Gemeinsame Wissenschaftskonferenz – GWK) based on an evaluation by the Leibniz Association: zbMATH became zbMATH Open. While EMS, FIZ Karlsruhe and the Heidelberg Academy of Sciences and Humanities renewed their editorial contract, the involvement of Springer came to an end. The EMS and the Heidelberg Academy of Sciences and Humanities remain important for quality control and the involvement of the mathematical community.

There are many reasons why the Heidelberg Academy of Sciences and Humanities and zbMATH Open are natural partners. By now this goes far beyond providing highly reliable information on mathematical publications. Many new aspects have arisen, not least the use of mathematical software, now addressed by swMATH, a database which has meanwhile become an integral part of zbMATH Open. Another topic, which concerns the Academy and zbMATH Open alike, is the responsible handling of (open) research data. The Heidelberg Academy of Sciences and Humanities is happy to see that zbMATH Open plays an active role in the NFDI consortium MaRDI (Mathematical Research Data Initiative). Research data and artificial intelligence are intimately linked, and tools such as zbMATH Open will play an essential role in future developments.

zbMATH open has the potential for becoming an essential tool for computer-aided search, supply and use of scientific information in mathematics and its applications. We are proud and happy to be an editorial institution of zbMATH Open.

We congratulate zbMATH Open on its 90th anniversary, and look forward to fruitful cooperation in the years to come.

Bernd Schneidmüller
President of the Heidelberg Academy of Sciences and Humanities

Greetings from the Head of Mathematics Department on behalf of FIZ Karlsruhe – Institute for Information Infrastructure

FIZ Karlsruhe - Institute for Information Infrastructure is happy to celebrate the 90th anniversary of zbMATH. Since 1979, FIZ Karlsruhe has been developing and maintaining the editorial and technical infrastructure of Zentralblatt MATH. During this time, the service has undergone a substantial development – the latest, and arguably most significant, being the recent transformation to zbMATH Open. Hosting the world's most comprehensive and longest running abstracting and reviewing service makes us very proud, and the achievement to make it open with the support of German federal and state funds fits perfectly into our mission of providing accessible and sustainable infrastructures for the scientific community.

zbMATH Open has both pursued traditional scientific values and procedures – especially, the editorial process which ensures the quality of the content by its large reviewer network and the editorial staff – and continuously implemented state-of-the-art techniques for information retrieval. Already the first years of Zentralblatt at FIZ Karlsruhe were marked by the establishment of a digital back-end system and the introduction of TEX. Since the 1980s, the service has been electronically available, since 1990 as CD-ROM, and since 1996 on the internet. In 2004, the Jahrbuch has been added, and the 2010s saw the introduction of the software layer swMATH and a fine-grained author database connected with a large number of external sources.

Today, propelled by the open interfaces and data which are available through its transformation, zbMATH Open develops into a hub of mathematical knowledge which provides scientists with interlinked information from a large variety of sources. The service is currently extending into various new directions, like the integration of mathematical research data and community platforms, the addition of affiliation information, or semantic entity linking. Many of these developments are possible through national and European projects, like the German National Data Infrastructure and the European Open Science Cloud.

The peculiar connection of modern developments and valuable sustainable scientific information specific for zbMATH Open is perhaps best demonstrated by the recent activities to convert the first volumes into LATEX. What was beyond an affordable effort for a long time, appears now feasible by new AI technologies.

FIZ Karlsruhe is happy to provide, with its strong partners, also in the next phase of zbMATH Open's life services for the mathematical community which take advantage of new technologies while sustaining the traditional quality and reliability.

Olaf Teschke
Head of Mathematics Department, FIZ Karlsruhe

Contents

Chapter 1

swMATH: Publication-based indexing of software

Hagen Chrapary, Wolfgang Dalitz, Wolfram Sperber, and Moritz Schubotz

The quote “For a scientific institute that does not need experimental facilities, the library is the most valuable asset” (Begehr, *Mathematik in Berlin*[1]) is an exemplary assessment of the importance of mathematical publications in the mid-20th century. Mathematical knowledge was mainly available in the form of publications. With the development of computers and their use for solving mathematical problems, experiments have now also found their way into mathematical research. Computerisation and digitisation have added new forms of mathematical knowledge to the traditional ones. This has led to broad impetus in the application of mathematical knowledge. In particular, mathematical modeling, the development of algorithms, and their implementation, generally summarised under the term Scientific Computing, have become indispensable tools in the industry, the service sector, and administration.

Therefore, the extension of adequate infrastructure for the access and use of this knowledge is required. This knowledge manifests itself in the form of mathematical research data, such as the versions of software and the underlying mathematical models. The management of the different classes of mathematical research data is more complex than that for publications. This is due to several reasons, particularly the dynamic nature of software development, the dependencies on hardware, operating systems, programming languages, data formatting, modeling languages, etc. Nevertheless, mathematical publications continue to form the core of mathematical knowledge and are a source and tool for Scientific Computing. This fact can also be used to develop specific services for research data, shown in the following using the swMATH Open service as an example. With the database swMATH, FIZ Karlsruhe and Zuse Institute Berlin (ZIB) have developed the world’s largest catalog for mathematical software, which currently lists almost 40,000 software objects, classifies them, and links them to software archives, such as Software Heritage, for open source software. The idea for the development is based on the evaluation of software references from the mathematical literature. The freely accessible database zbMATH is the world’s largest bibliographic database of mathematical publications, with currently about 4,300,000 entries, and forms the basis of swMATH.

Today, software development is often accompanied with a publication describing the software’s essential aspects (underlying algorithms, functionalities, hardware, and

[1] Heinrich Begehr, *Mathematik in Berlin – Geschichte und Dokumentation*, Erster Halbband, Shaker Verlag 1998, ISBN 3-8265-4225-8, p. 246

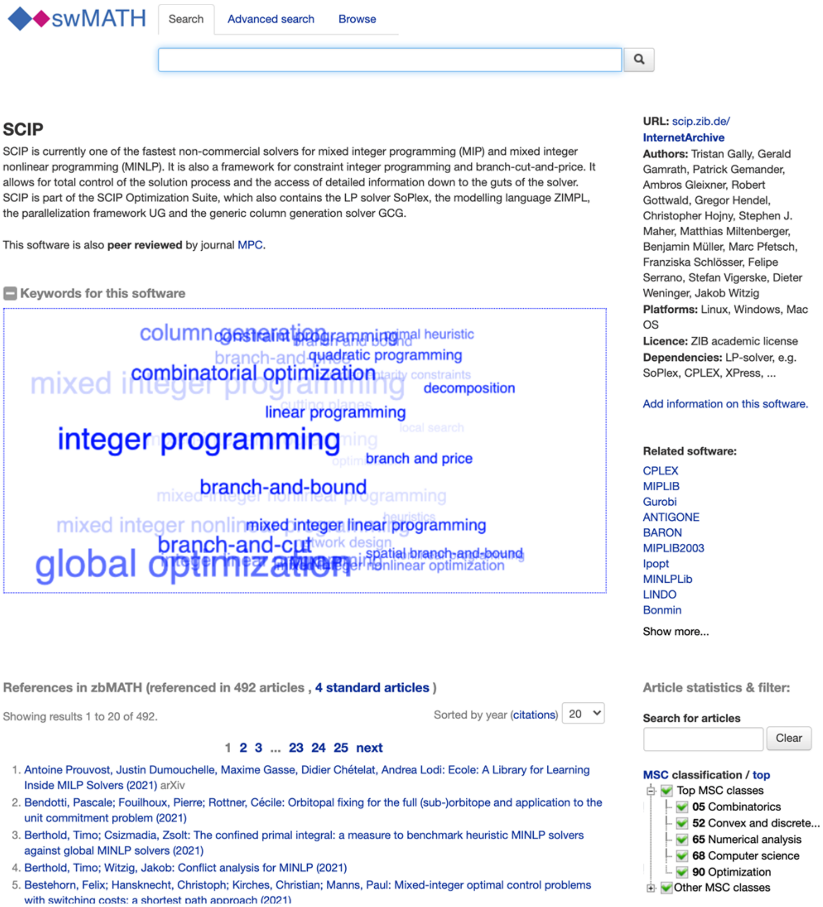
software requirements). These types of publications, referred to as standard publications in swMATH, are particularly highlighted in swMATH. Standard publications are often found in mathematical journals that specialize in mathematical software, such as the journal “ACM Transactions on Mathematical Software” (TOMS). A second relevant class of publications (“user publications”) containing information about a software cite this software in connection with the results they have achieved with its use. Many user publications on software can be taken as an indication of the high dissemination and quality of the software.

In the case of standard publications, the name of the software usually appears in the title. In the case of user publications, references to the software can be found primarily in the full text, in the citation lists, and in the keywords. The bibliographic data fields of the database zbMATH, particularly title, keywords, Mathematics Subject Classification (MSC), citations of the database zbMATH, and other sources, such as arXiv and journals and web sources specialized in mathematical software, are evaluated for information about software. The title of the software is used to search for publications in which this software has been used.

The swMATH service provides information about all software versions (under a common name). These objects are uniquely citable via the identifier introduced in swMATH. Every software goes through a life cycle, expressed in the different versions of a software. Unfortunately, there is still no standard for referencing software. In particular, information about the version of the software is often missing, which is, however, indispensable for verifying the results. The biographical information of the zbMATH database is also used to provide information about the software, such as the mathematical areas that were starting points for the development of the software, or application fields in which the software has been used. In turn, the swMATH project led to the recording of references of mathematical software in a further data field in the database zbMATH. On the one hand, this new data field facilitates software identification. On the other hand, it underlines mathematical software’s growing importance, especially for users from other scientific disciplines, industry, the service sector, and administration. The increasing importance of software citations is also expressed in the high number of software products in zbMATH; at the end of September 2021, 475,011 software references were found in 244,084 entries in zbMATH Open. Together with zbMATH Open and other publicly accessible sources such as arXiv, almost 250 thousand scientific publications are thus evaluated in swMATH.

Software user groups include software developers, users, providers, and service operators who require different information. Software data models, therefore, comprise a variety of information on the content, tasks, solution approaches, algorithms and limitations of the software, the software code, the requirements for and dependencies on hardware and software, the development status, the business model, and licenses, the developers and providers, test data and application examples, etc. These can differ significantly for individual versions and are sometimes incompatible.

A typical presentation of software entries in swMATH is demonstrated in Figure 1 using the swMATH entry of SCIP, a well-known optimization software:



SCIP

SCIP is currently one of the fastest non-commercial solvers for mixed integer programming (MIP) and mixed integer nonlinear programming (MINLP). It is also a framework for constraint integer programming and branch-cut-and-price. It allows for total control of the solution process and the access of detailed information down to the guts of the solver. SCIP is part of the SCIP Optimization Suite, which also contains the LP solver SoPlex, the modelling language ZIMPL, the parallelization framework UG and the generic column generation solver GCG.

This software is also **peer reviewed** by journal MPC.

Keywords for this software

URL: scip.zib.de/
InternetArchive
Authors: Tristan Gally, Gerald Gamrath, Patrick Gemander, Ambros Gleixner, Robert Gottwald, Gregor Hendel, Christopher Hojny, Stephen J. Maher, Matthias Miltenberger, Benjamin Müller, Marc Pfetsch, Franziska Schlösser, Felipe Serrano, Stefan Vigerske, Dieter Weninger, Jakob Witzig
Platforms: Linux, Windows, Mac OS
Licence: ZIB academic license
Dependencies: LP-solver, e.g. SoPlex, CPLEX, XPress, ...

Add information on this software.

Related software:

CPLEX
MIPLIB
Gurobi
ANTIGONE
BARON
MIPLIB2003
Ipopt
MINLPLib
LINDO
Bonmin

Show more...

References in zbMATH (referenced in 492 articles , 4 standard articles)

Showing results 1 to 20 of 492.

Sorted by year (citations) 20

1 2 3 ... 23 24 25 next

1. Antoine Prouvost, Justin Dumouchelle, Maxime Gasse, Didier Chételat, Andrea Lodi: Ecole: A Library for Learning Inside MILP Solvers (2021) arXiv
2. Bendotti, Pascale; Fouilhoux, Pierre; Rottner, Cécile: Orbitopal fixing for the full (sub-)orbitope and application to the unit commitment problem (2021)
3. Berthold, Timo; Csizmadia, Zsolt: The confined primal integral: a measure to benchmark heuristic MINLP solvers against global MINLP solvers (2021)
4. Berthold, Timo; Witzig, Jakob: Conflict analysis for MINLP (2021)
5. Bestehorn, Felix; Hansknecht, Christoph; Kirches, Christian; Manns, Paul: Mixed-integer optimal control problems with switching costs: a shortest path approach (2021)

Article statistics & filter:

Search for articles

Clear

MSC classification / top

- ☑ Top MSC classes
  - ☑ **05** Combinatorics
  - ☑ **52** Convex and discrete...
  - ☑ **65** Numerical analysis
  - ☑ **68** Computer science
  - ☑ **90** Optimization
- ☑ Other MSC classes

Figure 1. SCIP (https://swmath.org/software/1091). The URL is persistent and can be used for referencing the software.

First, swMATH provides the information extracted from the zbMATH database. The swMATH entries are based on two pillars:

- the analysis and evaluation of the information from the publications;
- the reference to external sources.

For analysis and evaluation of the publications, the zbMATH Open data provide an excellent basis for analysis and evaluation. From this, the following data are generated via the software for the swMATH entry:

- The description of the software: For this purpose, a summary of the current standard publication of the software is used, if available.
- The keywords from the standard, and user publications: A keyword cloud is generated from all standard and user publications. It contains all keywords and weights them according to the number of frequencies (the font size is chosen according to the frequency).
- The list of standard, and user publications that cite a software: This is of considerable value; a large number of publications indicates high relevance. The authors of the standard publications are usually the software developers.
- An MSC classification of the software: For this purpose, the number of MSC classes assigned to the publications is determined. We assume that the most frequently mentioned MSC classes of the publications also characterize the essential mathematical contents and application areas.
- References to similar software and dependencies between software products: The common mention of software names and a match in MSC classifications in standard and user publications are indicative of a relationship between different software products. However, the nature of this relationship cannot be inferred from mere common naming without a deeper analysis. Frequently, however, reference is made to similar alternative software products.
- Current development status of the software: From the publication dates of the publications, the life cycle of the software is inferred, expressed by the S-curve cycle typical in business administration for describing the life cycle of a product. After publication (date of the first standard publication), the dissemination of the software begins, leading to increasing publication numbers. After the numbers stabilize, the number of publications gradually decreases and ends after development ceases due to new development or the use of alternative products, which is expressed in the decrease of publication numbers. A major reason for the end of software development is the constant development of the framework conditions for a software (hardware and software).

swMATH essentially provides the bibliographical data of a software. If the Web contains more information, e.g., the code or a documentation of a software, then the swMATH entries link to the following information:

- The web pages of the software: The web page of a software usually offers a detailed overview of the current version of a software, contains information about the content and goals of the software development, about the code or the license terms, about the hardware and software requirements, the installation, about the developers and providers, etc.
- Software Heritage: For developers, the software code is of particular interest, for example, further joint development. The Software Heritage service developed by

INRIA (France) is the world's largest archive for software codes, offering information on all versions.

- The code of SCIP is not open source: Instead of the link to the code, the swMATH entry gives the license terms in the swMATH entry of SCIP.
- Internet Archive: The various versions of a software's web pages are archived periodically by the Internet Archive and thus provide an overview of a software's development history. However, only the top levels of the web pages are usually publicly accessible.

The advantage of the swMATH approach is that the information can be mainly generated automatically from the zbMATH data and Internet sources. Of course, further efforts and activities are needed to support the infrastructure for this type of research data. In particular, introducing a standard for software citations, which has been worked on for years, opens new perspectives for improving the swMATH service. The swMATH database is enjoying increasing use, which is also reflected in the visibility of swMATH via the major search engines.

The concept of swMATH can also be used for other classes of research data. Currently, a prototype of a database for mathematical algorithms is being developed. A mathematical software implements a mathematical algorithm, although the distinction between algorithm and software is not handled uniformly, which does not simplify the automatic identification of algorithms.

The publication-based approach of swMATH is the first successful step towards the comprehensive information of all classes of mathematical research data and an essential pillar of the mathematical information of FIZ Karlsruhe.

In the future, the swMATH team will collaborate closely with the German initiative for Mathematical Research Data (MaRDI). By doing so, we bring software and other mathematical research data such as formulae, numerical data, models, statistical data, and interactive notebooks, among many others, closer together. Additionally, we participate in the task force for Infrastructure Quality Research Software of the Europen Open Science Cloud. Therefore, we build a community of infrastructure providers for research software not only within mathematics. Combining forces with institutions with similar intents ensures that swMATH is well connected to similar initiatives from other disciplines.

Chapter 2

The zbMATH reviewer community

Dirk Werner

Our reviewer community is as diverse as the community of mathematicians itself, ranging from PhD students to Fields medallists. They come from nearly every country on this planet and can read dozens of languages. Nowadays, there are very few articles or monographs not written in English and hardly anything is published in languages beyond English, French, German and Russian (not forgetting Chinese), but we do have a review (in English) of an article written in Irish.

There are more than 7,000 reviewers, most of whom do their work diligently and reliably. The COVID-19 pandemic has, somewhat surprisingly, helped to attract more new reviewers and to convince many of those already on our roster to accept more material for review.

The percentage of papers that are reviewed vs. those that are just indexed with their summary, differs greatly among the mathematical fields. Typically, core subjects of pure mathematics fare much better in this respect than the physics or economics oriented applied fields. For instance, in algebraic geometry, about 55% of recent publications in zbMATH will find a reviewer, compared to less than 5% in solid and fluid mechanics, control theory or statistics. This is reflected by the structure of the reviewer community: of about 7,200 active zbMATH reviewers, most are in number theory and algebraic geometry (both 11%, while these areas make up only about 1% and 3% of the overall publications, respectively), followed by 10% of reviewers in each of the three areas of PDEs, functional analysis and operator theory (with publication shares of 6%, 1%, and 2%).

Even within a mathematical field, the percentage of reviewed papers differs. This is not a result of our assessment of the quality of the articles in question, but a clear indication of a lack of reviewers. For example, in C^*-algebra theory (MSC 46L) the review rate is only 30%, whereas MSC 46G (vector measures, infinite dimensional holomorphy and the like) sports a much higher 70%, which by no means should be mistaken to mean that one field is twice as important as the other.

It is sometimes argued that actual reviews are no longer needed these days since every paper comes with an abstract, readily available on the Internet, presenting a very short version of the introduction. However, apart from providing a more balanced synopsis of the paper, a review written by an experienced reviewer will, in addition, give some background and pointers to related literature and will highlight the main ideas, or maybe voice concerns about the validity of the arguments. (Recently, the latter happened in the reviews of the alleged solutions of the Navier–Stokes problem or of the abc conjecture.)

One can distinguish several types of reviewers when it comes to criticism. There are, for one thing, those for whom a review is incomplete unless there is an inkling of criticism, however petty, like pointing out trivial spelling errors. There are others who prefer not to take up an assignment when criticism is unavoidable. A third type is doubtful on how to spell out shortcomings; they often ask us in advance whether it is becoming at all to be critical and, if so, how to phrase this warily in order not to offend the authors.

When speaking about the vast majority of reviewers who do their work properly, a comment is pertinent about the others. They might just plagiarise the existing abstract; usually it is then enough to remind them in a short email to play honestly. There are other cases that indicate a much deeper problem in contemporary mathematical publishing, and we'd like to share a recent experience with you. One reviewer submitted what he thought was an acceptable contribution about a paper published in a top-notch journal written by a native English speaker. The submission, however, rather had the format of a referee's report giving hints at what the author should do upon acceptance of the paper. Being vague in its formulations the report did not allude to a single concrete result in the paper, but offered the advice to (a) check the paper for grammatical mistakes, (b) add some numerical examples (the paper was on monodromy groups), and (c) add a reference to a paper of his that had appeared in a journal that is, for good reason, not covered by zbMATH. The lesson to be learned here is that such one-size-fits-all "reviews" seem to be accepted by a brand of publications occasionally termed predatory.

Finishing on a positive note, we stress that our reviewers fulfill an important task, and some of them have shared their expertise for more than 60 years! Among the longest serving reviewers one should mention Johann Jakob Burkhard and János Aczél, both having contributed for 65 years, from 1939 until 2004 and 1946 until 2011, respectively. As remarked above, the mere abstracts of articles are easily found on the Internet; but the work of a gifted reviewer provides an added value that benefits the readership at large. Therefore we hope that many new reviewers will sign up[1] in the future; the mathematical community will surely appreciate their commitment.

[1] https://zbmath.org/become-a-reviewer

Chapter 3

The gender publication gap in mathematics: A bibliometric analysis of zbMATH data

Helena Mihaljević, Lucía Santamaría

The achievement of *gender equality and empowerment of all women and girls* is one of the 17 goals listed by the United Nations' 2030 Agenda for Sustainable Development towards a more peaceful, inclusive, equal, prosperous and sustainable world. According to the most recent Global Gender Gap Report 2021 of the World Economic Forum, *the COVID-19 crisis has increased pre-existing gender inequalities*, meaning that "another generation of women will have to wait for gender parity".[1] Recent investigations have collected evidence that the pandemic has affected female academics in STEMM fields (science, technology, engineering, mathematics, and medicine) particularly hard along multiple dimensions, such as productivity, boundary setting and control, and the ability to engage actively in collaborations and network building.[2] In order to fully understand the gender gap in academia and its development, for instance to assess and counteract the effects of crises such as pandemics, fine-grained data are needed. These typically need to go beyond the often-employed high-level statistics such as those measured by the Global Gender Gap Index applied in the WEF Report.

1 The Gender Gap in Science Project

In 2017 eleven scientific organizations, led by the International Mathematical Union (IMU) and the International Union of Pure and Applied Chemistry (IUPAC), joined efforts to conduct an interdisciplinary, cross-national project to gather and analyse comprehensive data on the situation of women in mathematics, computing and natural sciences. The project "A Global Approach to the Gender Gap in Mathematical, Computing, and Natural Sciences: How to Measure It, How to Reduce It?"[3] was funded for the period 2017–2020 by the International Science Council (ISC). Annual

[1] https://www3.weforum.org/docs/WEF_GGGR_2021.pdf

[2] E. Higginbotham and M. Lund Dahlberg (eds.), *The impact of COVID-19 on the careers of women in academic sciences, engineering, and medicine*. A Consensus Study Report of the National Academies of Sciences, Engineering, and Medicine. The National Academies Press, Washington, DC, 2021 https://doi.org/10.17226/26061

[3] https://gender-gap-in-science.org

coordination meetings were held by partners to discuss goals, approaches and methodology. A well-attended final conference was organized in November 2019 at the ICTP in Trieste, after which the project's final report was made public.[4]

The Gender Gap project was articulated around three central themes. Besides a Global Survey of Scientists and the creation of a Database of Good Practices, the third working package consisted of the *examination of the situation of academic authors and their publication practices in different academic fields across world countries and regions with respect to the scientists' gender*. This type of analysis makes it possible to identify common, and discipline-specific issues that might require interventions in view of the measured gender gap.

The reason for a focus on publishing practices lay in the importance of publications for academic careers. Scientific publications are not only the major outlet for scholarly communication, they are regarded as a proxy for a researcher's scientific credo and play a key role in achieving and maintaining a successful career in academia. Decisions on tenure and other academic promotions are mostly based on evaluations of the candidate's research portfolio that pay special attention to research publications like journal articles, in addition to grants, conference presentations, and how visible or well-recognized a scholar is. Thus, *the understanding of publication practices, obtained through measurable data on research output, is of great interest* to academic institutions, science policymakers, and researchers alike.

Multiple studies based on bibliometric data have concentrated on the variable of gender. The literature also comprises discipline-specific findings from the area of mathematics, albeit in small numbers. Much of the existing scientometric research builds on cross-discipline corpora such as Scopus and, accordingly, focuses less on individual fields. Research directed to a specific discipline or subfield, in turn, typically examines a limited selection of journals or conferences or a narrow time period. In the aforementioned Gender Gap project, we built on existing results and focused on data sources managed by community organisations and curated by experts, encompassing the respective disciplines as comprehensively as possible in terms of content and temporal coverage. *The analyses of publication behavior in mathematics were performed on zbMATH data*, made available to us at regular intervals in order to provide the most up-to-date status of the additional information gathered by the zbMATH office, such as improved author profiles or extracted geo-entities.

Below we present various key findings from the Gender Gap in Science project related to mathematics. Further results related to gender, as well as to mathematical

[4] https://gender-gap-in-science.org/2019/11/09/celebration-of-the-conference-on-global-approach-to-the-gender-gap-in-mathematical-computing-and-natural-sciences-how-to-measure-it-how-to-reduce-it

publishing in general, plus additional context information e.g. on the data processing algorithms that were employed, can be found in the final project report.[5]

2 The Gender Gap in mathematical publications: Cohorts and gender analyses

We analysed the full collection of publications by scientists with a main research focus in mathematics ("core mathematicians") from 1970 until July 2019. This *data set comprises more than 3 million documents corresponding to more than 5.2 million authorships* (pairs of author and document), yielding an average of 1.7 authors per article. We inferred the gender of these authorships from the authors' names via various statistical name-gender databases and services, resulting in approximately 3.6 million being assigned to men, 0.5 million to women, and 1.2 million that could not be matched to any gender. Omitting authors for which our gender assignment procedure led to no reliable result, authorships of women accounted for about 12% of the total. These[6] in turn belong to ca. 65,000 authors labeled as women and ca. 260,000 authors labeled as men, which yields around 21% women among all recorded authors in zbMATH in the mentioned time span. Figure 1 shows the number of authors according to the year of their first publication ("cohort"), and the percentage of women among them. The proportion of women has increased steadily, growing from less than 10% in the 1970s to over 27% after 50 years. Moreover, nowadays, *more than 14,000 new mathematicians start publishing per year, corresponding to 4,000 women that enter the field of mathematics annually*.

While more and more women become part of academic mathematical research, the question arises how many of them continue to pursue scientific careers in the field several years later. After all, numerous studies show that the percentage of women decreases drastically the higher one looks up the career ladder. Therefore we analysed how many authors "drop out" after a given number of years: we checked, per author and time span, whether each author still appears in zbMATH a number of years after their first publication. Figure 2 visualizes the proportions grouped by cohort and gender for all authors who had been initially active for five years. The assumption of an initial period of five years of activity serves as a proxy for the post-doctoral stage,

[5] M.-F. Roy, C. Guillopé, M. Cesa, R. Ivie, S. White, H. Mihaljevic, L. Santamaría, R. Kelly, M. Goos, S. Ponce Dawson, I. Gledhill, and M.-H. Chiu, A global approach to the gender gap in mathematical, computing, and natural sciences: How to measure it, how to reduce it? International Mathematical Union (2020) https://doi.org/10.5281/zenodo.3882609

[6] Not all authorships can be assigned to a unique author, in particular if the author's name is frequent.

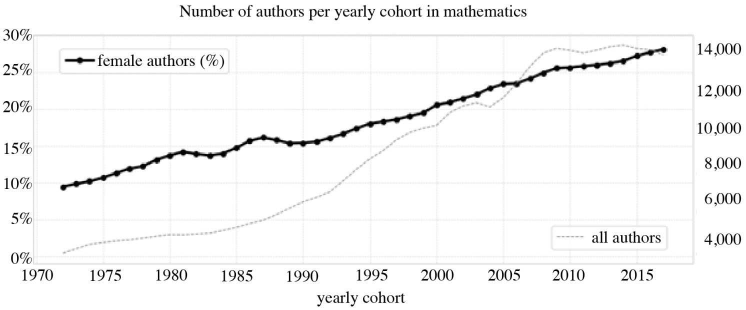


Figure 1. (Dotted grey line; right axis) Defining a zbMATH author's cohort as the year of their first publication, number of authors found in the database per cohort from 1970 until 2017. (Solid black line; left axis) Percentage of all authors that could be algorithmically assigned as female.

thus the figure implies the following: the number of authors that stay in academia further 6 to 10 years has reduced enormously when comparing the 1970s cohorts with those from the 2000s. If we associate the subsequent 10-year period with the time when a permanent academic position is secured, then around 60% of the male "post-docs" from the most recent cohorts manage to achieve such a career milestone. For women, the percentages have been, and continue being, lower than for men. However, the differences between women and men have reduced over time. Likely, this is mainly due to the fact that the number of PhD students and post-docs has grown much faster than the available permanent positions in mathematical research.

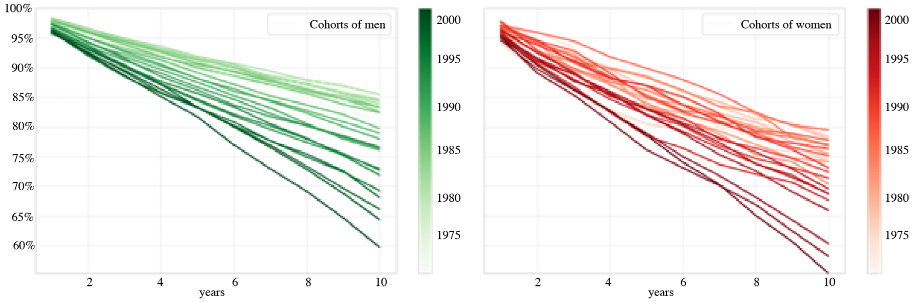


Figure 2. Percentage of male (left) and female (right) mathematicians that continue publishing for another 1 to 10 years after having been active for 5 years. The colors indicate cohorts, with dark colors indicating the most recent ones. The figure exposes a "publishing drop-out rate" in mathematics throughout the past four decades.

3 The Gender Gap in renowned mathematical journals

As already mentioned, scholarly journals are a crucial vehicle for the forging of academic careers in STEMM, as decisions on tenure, funding, and promotions strongly depend on the researchers' publication record. Moreover, it is not just the number of articles a scientist writes that matters, but also the venue where they appear. Publishing in highly renowned journals in one's discipline is a powerful determinant of tenure in many STEMM fields including mathematics, and an important predictor of professional success. Thus, any bibliometric study on publication practices ought to take into account their impact in the making of academic careers.

In previous research,[7] also based on zbMATH data, we had already demonstrated that authorships by women are vastly underrepresented in journals with a high reputation in terms of two common ranking methods, the manually compiled Australian ERA indicator and the journal impact factor (JIF). In this project, we intended to offer the scientific community the opportunity of examining gender distributions in journals of particular relevance to them or their subfield. We made this possible via a dedicated web interface that allows readers to filter specific publication venues of their interest.

Additionally, we have taken *a close look at selected journals published by mathematical societies as well as those particularly renowned in individual topical subfields*. Figure 3 illustrates that the percentages of authorships from women in said selected journals are predominantly constrained below 20%. Around half of the society journals show a rising tendency over the past decades. The *Bulletin de la Société Mathématique de France* shows a rather noisy behavior and no clear chronological trend, with close to no publications by women at all in various years. The average share is around 10%, similar to the *Journal of the European Mathematical Society*. The lowest percentages are found in the *Journal of the American Mathematical Society*, where the proportion of women is around 5% or less, and shows no noticeable increase over time. The bottom three topical journals on the right-hand column, which mainly feature works in areas of applied mathematics, display a rising development over time with shares above 10% in recent years. Except for the *Journal of Differential Geometry*, all journals reveal a slight positive trend. The renowned journals *Inventiones Mathematicae* and *Annals of Mathematics*, which for the most part publish work in pure mathematics, stand out with percentages of women authorships predominantly in the single-digit range.[8]

[7] H. Mihaljević–Brandt, L. Santamaría, and M. Tullney, The effect of gender in the publication patterns in mathematics. PLOS ONE 11 (10): e0165367 (2016) https://doi.org/10.1371/journal.pone.0165367

[8] For more details, see H. Mihaljević and L. Santamaría, Authorship in top-ranked mathematical and physical journals: Role of gender on self-perceptions and bibliographic evidence. Quantitative Science Studies 1 (4): 1468–1492 (2020) https://doi.org/10.1162/qss_a_00090

There may be several potential causes for the measured underrepresentation, but these cannot be determined from the bibliographic data. As an alternative data source we have leveraged the 2018 Global Survey of Mathematical, Natural, and Computing Scientists that was conducted as another working package of the project to obtain answers from almost 10,000 mathematicians, physicists, and astronomers about their submission practices to top-ranked journals in their disciplines. More precisely, we asked the following question: "*During the last five years, how many articles have you submitted to journals that are top-ranked in your field?*" Respondents were expected to provide a number between 0 and 30; larger values were clustered together. According to the obtained responses, women and men self-report to have submitted similar numbers of articles in the past 5 years, with no major statistically significant differences in subgroup analyses broken down by disciplines or world regions. What matters much more than gender in the computed model is strong research activity, a professional network, and overall academic success.

The reported perceived submission practices do not support the hypothesis that the underrepresentation of women in prestigious journals is mainly rooted in them submitting less manuscripts for consideration than men. Considering the importance of publishing in renowned journals on the one hand and the conflicting bibliographic analysis on the other, this begs the question on the role of peer review. We observe that the refereeing system in mathematics lacks homogeneity and relies substantially on the authors' credit and the level of trust between editors and reviewer(s). In this regard, we stress that *there are hardly any systematic studies on the peer review process in mathematics*,[9] a need that very much ought to be addressed.

4 Learnings and perspectives on the Gender Gap in mathematical publications

Inspired by the UN's agenda to reach gender equality and empowerment of all women and girls within the next decade, we set out to investigate the existence and characteristics of a particular gender gap: the *underrepresentation of female authors in academic publishing in mathematics with respect to their male counterparts*. The comprehensive data collection from zbMATH as well as our usage of algorithmic methods at scale make this bibliometric analysis feasible.

There are various aspects to consider when speaking of a gender gap. We have provided insights on the gap defined by the proportional presence of women as authors

[9] C. Geist, B. Löwe and B. Van Kerkhove, Peer review and knowledge by testimony in mathematics. In *PhiMSAMP: Philosophy of mathematics: Sociological aspects and mathematical practice*, pp. 155–178. London, College Publications, 2010.

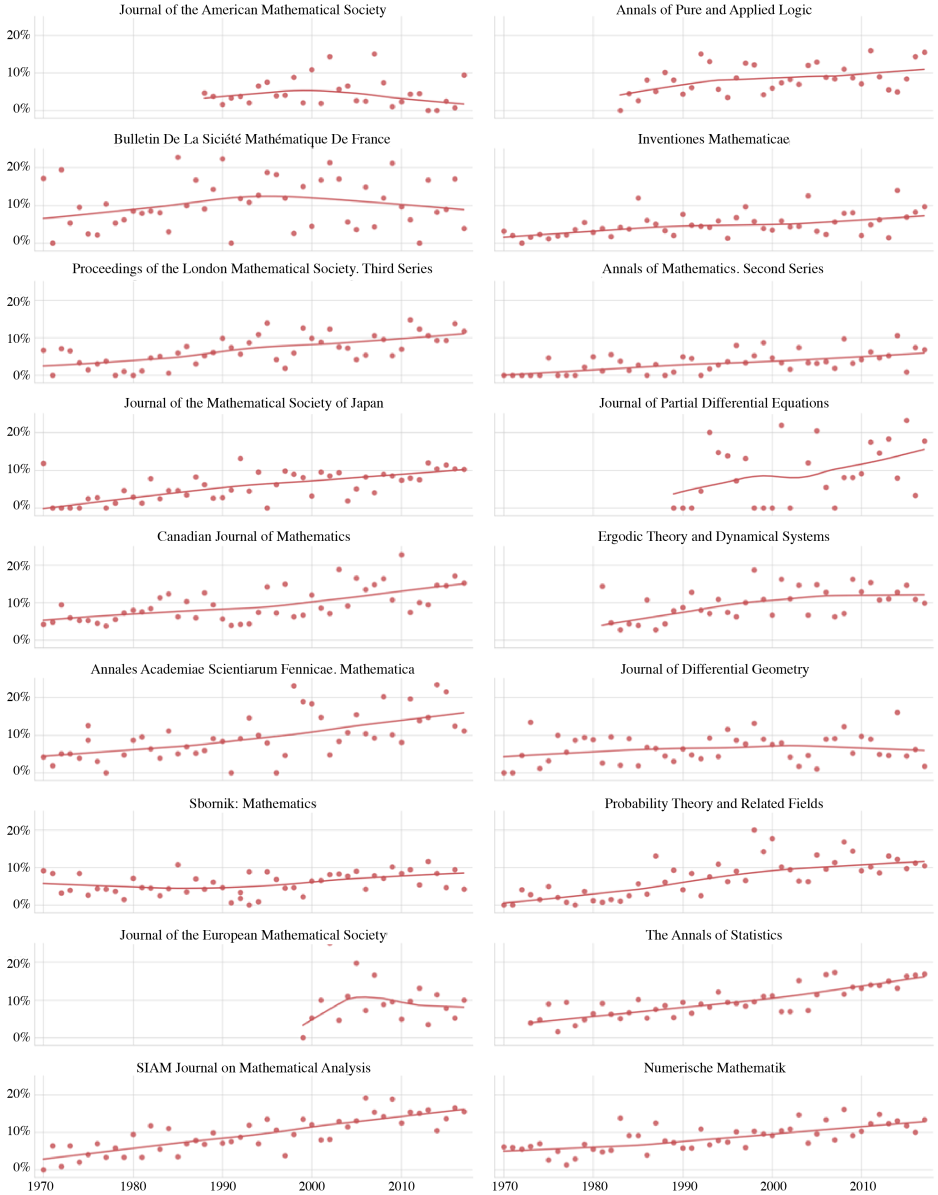


Figure 3. Percentage of authorships from women in renowned mathematics journals per year between 1970 and 2017.

of core mathematics publications; we have also investigated whether there is a gender gap in the dropout rates that affect the length of mathematicians' academic careers; finally, we have focused on the gender gap in renowned, high-impact mathematical journals.

Consistent with the global trend in higher education, we observe *increasing proportions of women entering the field of mathematics with each passing year*. The understanding of the extent to which those newcomers will progressively attain senior academic positions is crucial to address the "leaky pipeline" phenomenon. Thanks to our cohort analysis based on zbMATH publication data, we are able to provide insights on this issue. We show that dropout rates of mathematicians after their postdoctoral stage, which used to be higher for women, are converging on similar figures for both genders. These data certainly offer optimistic prospects regarding the eventual closure of this particular aspect of the gender gap.

On the other hand, our analysis of women's presence in renowned journals is a good measure of the gender gap in relation to achieving a prestigious academic career. In this regard, a non-negligible number of the *prestigious mathematical journals under consideration show a meager representation of women among their authors*. All other factors being equal, the expectation is that the proportion of women among all authors should roughly resemble the percentage of established female mathematicians in the profession, a number that has been steadily growing and that is estimated to be currently around 25%. Remarkably, several of the analysed journals publish very few articles authored by women and exhibit no signs of turnaround over the last couple of decades. An explanation for this fact might lie in the characteristics of the peer review process in mathematics, which favors close interactions and trust relationships between editors and reviewers and opens the door to conscious and unconscious biases. Regarding subfields, applied areas display a better situation for women than pure ones, which in itself introduces a series of discussion points regarding the intrinsic differences among subfields of mathematics.

The above remarks provide a compelling starting point for future research questions. Is the increasing number of young female mathematicians enough to stop the pipeline from leaking? Which factor in the retention of women in academia is played by the professional atmosphere in pure versus applied mathematics? What is the importance of informal academic networks to make a mathematician's career thrive? Is the lack of double-blindness in peer review hindering women and other underrepresented groups in mathematics? *It would be excellent to discuss our data-backed findings with experts from the respective subfields in the mathematical community*, with the goal of formulating plausible hypotheses that could explain the observations found by our work in the Gender Gap project.

Chapter 4

Quality control at zbMATH

Dirk Werner

zbMATH is a reviewing and abstracting service, which, according to its own definition, sets out to cover all mathematical publications presenting a “genuinely new point of view.”

Whereas before 2010 almost all periodicals could be assumed to satisfy this criterion, the situation changed with the advent of Open Access platforms, a number of which were dubbed “potential, possible or probable predatory publishers” by J. Beall in his now defunct list. Indeed, in some of the journals falling into this category, papers “proving” the Riemann Hypothesis or Fermat’s Last Theorem in a couple of lines can be found; sadly, they were indexed in Zentralblatt because it tried to be as complete as possible at the time. However, on closer inspection these papers revealed a deeper problem of those journals: improper or missing peer review. Since peer reviewing is an indispensable prerequisite for getting indexed in zbMATH, this was a clear indication to discontinue indexing such periodicals.

But another problem became evident. Every week we receive enquiries from editors of newly founded journals asking us to index their papers. Most of them do not publish nonsense like 3-line proofs (or refutations) of the Riemann Hypothesis, but still most of the papers are at the level of exercises, where the authors reproduce a known proof under a formally less restrictive hypothesis. We do not consider such ε-perturbations of known facts as really new, and after tightening our indexing policy some years ago, leading to the requirement of a “genuinely new point of view,” we decided not to index journals in this quality segment. In the last two years there were more than 100 enquiries concerning indexation, but only 25 were granted, mostly for the reason explained above. (The other class of nonindexable journals are those that carry no, or hardly any mathematics.)

When this policy is implemented properly, readers of zbMATH can reasonably expect that only papers from serious journals are indexed. (Here, serious is meant in a wide sense; there are loads of reasonable journals that certainly do not match Acta Mathematica.) Alas, erring is human, and hence a small percentage of published papers contain errors or gaps, sometimes small and sometimes big. The mistakes are often found by the authors themselves, but sometimes by our reviewers, which might or might not lead to a correction or, when the worst comes to the worst, to a retraction. Criticism by reviewers is generally welcome if it is based on facts rather than prejudice and formulated in polite terms. That authors might still not accept the critique is another matter; a case in point is the alleged solution of one of the Millennium Problems that was proved incorrect in the zbMATH review of the corresponding paper.

Reviewers who find mistakes are sometimes reluctant to point them out in public and seek refuge in asking to publish the abstract instead or not to index the paper altogether. However, we think this does a disservice to the community at large, and we try to convince such reviewers to state the problematic parts matter-of-factly, to the advantage of all readers. Incidentally, publishing the abstract of a paper instead of a review is not an indication of lacking quality, but one of lacking reviewers.

Duplication of papers is another matter of concern. We distinguish between two types of duplications. The first one, considered legitimate, is when an author presents his own paper in a seminar-type volume before the "official" journal version is published. As opposed to this there are those (self-) duplications where authors publish the same paper twice in different journals, naturally without citing the other version. Worse than this are duplications when author A republishes a paper of author B. Though such a behaviour is widely known as plagiarism, we stick to the facts and say that the papers are identical; it is then practically always clear who has copied whom. Again, our reviewers help detect such cases that went unnoticed by antiplagiarism software.

Finally, we also monitor the quality of the reviews themselves. Each review is edited by at least one editor to make sure that the number of typos and language slips remains below a critical barrier. But more importantly, we aim at publishing reviews that convey information which cannot be trivially gleaned from the abstract of the paper. Every week some (however few) reviewers try to make us believe that a submitted text identical to the abstract, just with "we show" replaced by "the authors show" (and sometimes even without this amendment), is an acceptable review that should justifiably carry the signature of the person who submitted it; it is not, and we gently indicate to those reviewers that our review request forms explicitly ask for extensive quotes to be labelled as such.

In conclusion, quality control is a multi-faceted endeavour, from the choice of journals suitable for indexing to the editing of reviews of individual papers.

Chapter 5

The digital shadow of mathematics and its ramifications

Howard S. Cohl, Moritz Schubotz

1 The natural habitat of mathematics

Mathematics is ubiquitous. It is the incredibly creative and widely spoken language of logic. Mathematical knowledge resides in the minds of mathematicians and those who want or need to use mathematics. They have learned mathematics through the process of thought; by listening to people speak and through conversations; through the reading of books; and more recently by browsing on the internet; by listening and watching videos; and through written and computational practice. Through practice, understanding is accomplished and has led to concept extension and generalization. Ultimately, through writing, publication and presentation, dissemination is obtained and one hopes to achieve the global blossoming, conceptualization and description of mathematical notions. This continuing process generates a wealth of data which should be shared globally, but is in practice subject to restrictions to access. Through refinement and use, important results precipitate and become more widely available. The use of computers has greatly facilitated this process.

Mathematical functions and the operations they satisfy are widespread. The so-called special functions are mathematical functions which are so useful and have appeared so often (in applications) that they have been given special names. As well as special functions, there are also special constants, numbers and special sequences of numbers (see the On-Line Encyclopedia of Integer Sequences[1]). There is also a large collection of mathematical objects or operations which have commonly appeared and these have been given names as well. The names of these special objects summarise and provide an organisational structure to mathematics.

What are special function names? These names are often ascribed to the discoverer or to a person who greatly exploited their use, or simply to a description of their action, or sometimes, out of the void. Special functions arise in a variety of contexts. Historically, some of the most common special functions arose in areas of classical analysis and natural mathematics such as in the study of the figure of the

[1] https://oeis.org (The mention of specific products, trademarks, or brand names is for purposes of identification only. Such mention is not to be interpreted in any way as an endorsement or certification of such products or brands by the National Institute of Standards and Technology, nor does it imply that the products so identified are necessarily the best available for the purpose. All trademarks mentioned herein belong to their respective owners.)

earth by Pierre-Simon Laplace, through the separation of linear partial differential equations. Special functions include classical orthogonal polynomials, Bessel functions, associated Legendre functions, the gamma function, and elliptic integrals. Even more esoteric functions arise such as parabolic cylinder functions, Mathieu functions, Lamé functions, ellipsoidal harmonics, elliptic functions and so on. This is the tip of the iceberg. There are many more and today one can find an excellent summary of the most important ones in the NIST Digital Library of Mathematical Functions [1] in which there are 36 chapters, each focusing on their most important properties. More generally, Scharpf et al. [3] define the term of *Formula Concept* “as a collection of equivalent formulae with different representations.” The question and description of equivalence must be pinpointed so that a wider audience may understand the conversation. Special function and number data is a subset of the much larger collection of mathematical knowledge.

Nowadays one can peruse the constantly evolving collection of mathematical knowledge by visiting and examining the online arXiv preprint server. This dataset gives a good sample of the breath and depth of mathematical knowledge which is constantly evolving. In an even more refined collection, the journals of mathematics and mathematical physics provide an even more carefully curated collection of information. There are as well collections of monographs published by the mathematical science publishers. Together, we have focused on the mathematical knowledge associated with the real and complex analysis of special functions and numbers. However, there exist alternative and more abstract mathematical knowledge such as that which is connected with group theory, abstract algebra, number theory, differential geometry, topology, graph theory, category theory, set theory, type theory, logic, and so on. There is often a deep underlying connection between these fields which all have footprints in the entirety of mathematical knowledge.

Of supreme wealth has been those mathematicians who explore the mathematical terrain through their research – those who have discovered and revisited areas, and have provided extensions, generalizations – new results. Usually these individuals provide the benefit of sharing their discoveries through publication of journal articles and perhaps in monographs. In the future, AI may more significantly play the role of these mathematicians, but there exist significant obstacles to this transition [2, 5].

2 The ongoing and future invasion of mathematics into the digital space

In order for there to be full computational access to the data associated with mathematical knowledge, one must transform this data into a form in which it is understandable by a machine. In order to accomplish this, one should enhance the machine so that it is clever enough to understand and use the data. This is the problem of semanticisation or semantic augmentation [4] and it lies at the heart of the problem. We must develop a confidence that correct mathematical meanings may be inferred by the machine. With the mathematics of special functions, this journey is well underway, and our special route is through the preparation of mathematical documents, the most common way to spread and communicate mathematical description.

The most common method for mathematical data to be entered into the literature is connected to the problem of typesetting mathematical information. In today's literature, the most common method for typesetting mathematical information is with the use of TEX or LATEX. These are programming languages which center around typesetting mathematical expressions. Even though LATEX produces readable presentation, the content may be shrouded. In order to remedy this, more precise methods for describing mathematical syntax is necessary. For now, we enjoy communication to and through Computer Algebra Systems (e.g., Mathematica, Maple, MATLAB, Reduce, Magma, SageMath, SymPy, etc.). In the case of online mathematical content, the use of XML or MATHML is powerfully opportune. Our team captures this semantic data initially though the use of LATEX and important metadata connected with the content is provided to the user.

Even more thorough prescription for describing mathematical content on a machine is provided through languages used to develop formalized mathematics – such as those used in Automated or Interactive Theorem Proving (e.g., Lean (proof assistant), Isabelle (automated theorem prover), Coq (interactive theorem prover)). As one moves further in this direction, the ability for humans to read the mathematics starts to fade away, but the ability for computers to process such information is greatly enhanced. This is the question of human comprehension vs. the question of machine comprehension and the ability to rid oneself of ambiguity while enhancing precision – a principal goal for the field of mathematical knowledge management.

3 Our conclusion and the eventual payoff

Many features of mathematics are clear to its readers. However, in reality there are many assumptions which the reader understands without explanation. This becomes apparent, when looking at theorem proving systems and digital mathematical compendia which describe the semantics of mathematics. Semanticisation is the horizon where humans and description meet and will play a fundamental role for the forthcoming evolution of mathematical knowledge. Once there is a critically large machine-readable collection of mathematical content, through the bootstrapping process, artificial intelligence should be able to ascertain missing semantic through the same process that humans use. When mathematical knowledge is fully accessible to machines, only our imagination will provide a boundary to possible routes of mathematical exploration in the digital realm. We have only just begun.

References

[1] NIST Digital Library of Mathematical Functions, http://dlmf.nist.gov, Release 1.2.0 of 2024-03-27 visited on 14 March 2024

[2] A. Greiner-Petter, T. Ruas, M. Schubotz, A. Aizawa, W. I. Grosky, and B. Gipp, Why machines cannot learn mathematics, yet. In *Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2019) co-located with the 42nd international ACM SIGIR conference on research and development in information retrieval (SIGIR 2019), Paris, France, July 25, 2019*, edited by M. K. Chandrasekaran and P. Mayr, pp. 130–137, CEUR Workshop Proceedings 2414, CEUR-WS.org, 2019

[3] P. Scharpf, M. Schubotz, H. S. Cohl, and B. Gipp, Towards formula concept discovery and recognition. In *Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2019) co-located with the 42nd international ACM SIGIR conference on research and development in information retrieval (SIGIR 2019), Paris, France, July 25, 2019*, edited by M. K. Chandrasekaran and P. Mayr, pp. 108–115, CEUR Workshop Proceedings 2414, CEUR-WS.org, 2019

[4] M. Schubotz, *Augmenting mathematical formulae for more effective querying & efficient presentation*. Ph.D. thesis, Technical University of Berlin, Germany, 2017, https://d-nb.info/1135201722 visited on 14 March 2024

[5] M. Schubotz, N. Meuschke, M. Leich, and B. Gipp, Exploring the one-brain barrier: A manual contribution to the NTCIR-12 MathIR task. In *Proceedings of the 12th NTCIR conference on evaluation of information access technologies, National Center of Sciences, Tokyo, Japan, June 7-10, 2016*, edited by N. Kando, T. Sakai, and M. Sanderson, pp. 309–317, National Institute of Informatics (NII), 2016

Chapter 6

Digital math libraries and the commitment to open access at zbMATH

Dariush Ehsani

In addition to providing open access to the reviews at zbMATH Open, zbMATH attempts to connect articles within its database with online digital versions. This is to be seen, for instance, in the DOI and arXiv links on an article's information page at zbmath.org. A goal of zbMATH Open is not just to provide a link to a digital version of an article, whose full-text may be off-limits to non-subscribers of a particular journal, but also to provide an open access link to the article whenever possible. The ideal case would be the DOI pointing to the digital version at an open access journal, but an often well-suited alternative is to link to a preprint on an open access preprint server.

Furthermore, the potential to provide digital mathematical content is significant. Based on [2], we can estimate that more than 60% of around 130 million pages of math research since 1868 is digitally available. The use of DOIs has become a standard method to identify digital content, and the emergence of repositories such as arXiv, EuDML, Gallica, JSTOR, Math-Net.ru, or Project Euclid has enhanced the access to math articles and preprints.

In the case of linking to open access journals, several considerations must be made. For instance, the stability of the journal must be taken into account (change of publisher, change in open access policy of publisher, discontinuation of journal could all affect access to digital content), and even if there are no changes in a journal, the journal's open access policy could present some problems. In some cases there is a "moving wall" policy, so that access to an article is only available after some period of time.

Peer-review is essential to ensuring a standard and quality of mathematical research, and mathematicians consider the quality peer-review the most important asset of a journal [3]. In that sense, the open access policy of a journal may not come into consideration when an author is publishing research. Self-archiving on a repository or on another website in this situation becomes an important tool in open access; this type of open access, where an author publishes a document, and self-archives is known as "green" open access. Of course, the question of what to do with historical publications published before the time of preprint servers remains. In this regard, moving walls (eventually making a publication open access) on the part of journals or publishers is an important step, and such goals are being increasingly adopted and encouraged by EuDML, ProjectEuclid, and MathNet.Ru. But in light of the peer-review concerns of mathematicians, encouraging publishers to adopt open access policies, where the journal provides open access to an article seems to be essential in moving towards open access. Even here, though, some distinctions between types of open access need to be kept in mind. In some cases, journals agree to an open access policy for digital content in exchange for an article processing charge (APC); such access is characterized as "gold." While APC journals account for most of the growth of open access journal publications, they rarely fall into the category of core mathematics journals, which are defined to relate solely to mathematical content and belong to the top two indices in [4]. An open access journal which provides access without issuing additional charges is termed "platinum" or "diamond," where additional costs are usually picked up by a third party, such as a large institution or library. While the advantages of such a policy is clear (no obstacles on the part of the author to providing access to research), the financial obstacles are obvious; funding has to come from somewhere, and that funding has to be stable. Still, the impact of green open access is most significant; it basically accounts for all progress made in open access share in core mathematics journals during the last two decades [1]. This evolution can be seen in Figure 1.

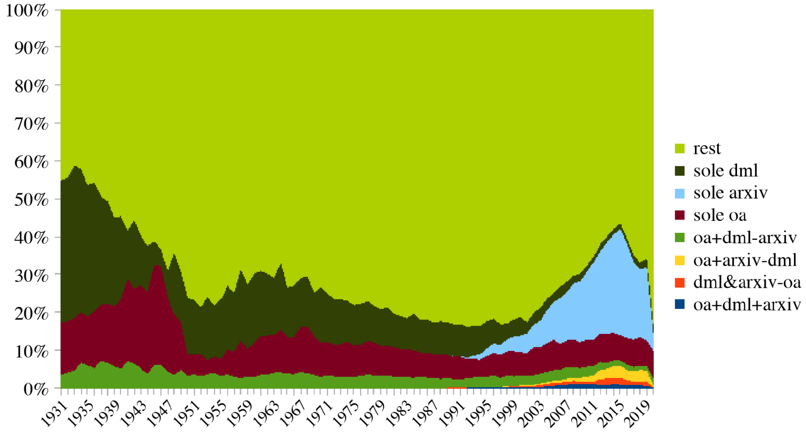


Figure 1. Share of different open access solutions/combinations in core math electronic journals (defined below) by publication years.

While the repositories mentioned above can be considered to be relatively stable, the argument can be made towards the usefulness of starting a repository to collect digital content found on sites deemed to be "unstable" (for instance a preprint on a homepage of an individual author). In such a scenario, zbMATH would take on responsibility to provide links and open access to preprints, or in some cases to digital content from journals, of selected articles. Beyond storage and digital repository software concerns, effort would have to be made to ensure that no digital content runs counter to copyright protections. Furthermore, keeping the repository up to date (replacing older versions of articles/preprints with the latest versions) is essential to good maintenance.

These concerns will be taken into consideration in zbMATH's endeavor to keep a digital repository of journal articles currently at EMIS,[1] the goal being to link to an internal repository, when possible and if necessary (for instance in the case there is no journal link to an article), of an article when called from zbmath.org.

In the end, however, linking to digital content, as well as linking to open access versions either via external or internal repositories greatly enhances zbMATH's efforts into providing and supporting the distribution of open access math content.

References

[1] D. Ehsani and O. Teschke, On the road to a comprehensive open digital mathematics library. *Eur. Math. Soc. Newsl.* **118** (2020), 76–78

[2] P. D. F. Ion and O. Teschke, Continuing toward a Global Digital Mathematics Library. Talk given at the *AMS Special Session on Mathematical Information in the Digital Age of Science at the Joint Mathematics Meetings*, San Diego, CA, 2018

[3] C. Neylon, D. M. Roberts, and M. C. Wilson, Results of a worldwide survey of mathematicians on journal reform. *Eur. Math. Soc. Newsl.* **103** (2017), 46–49

[4] O. Teschke, Green, gold, platinum, nickel: On the status of open access in mathematics. *Eur. Math. Soc. Newsl.* **110** (2018), 60–63

[1] https://www.emis.de

Chapter 7

Examination of the state of the art of mathematical formula search for zbMATH Open

Johannes Stegmüller, André Greiner-Petter, Petr Sojka, Olaf Teschke, and Moritz Schubotz

The service for abstracting and editing mathematical content zbMATH Open, which also offers a formula search, is constantly being developed. Since the beginning of 2021, zbMATH Open has been open for public access. To leverage the opportunities of some recent developments in zbMATH Open and for formula search, we examine the state of the art in math search engines and their applications. Also, based on our investigation, we present several proposals for improvements to the formula search of zbMATH Open.

1 Introduction

zbMATH Open[1] (shorter zbMATH, formerly Zentralblatt MATH) is an abstracting and reviewing service for mathematical content. At the time of writing, it contains 4.3 million bibliographic entries with publication years between 1826 and 2022. There have been 1,123,159 reviews since 1868 by the community of reviewers, which currently counts 7,677 active associates.

The publicly available web-interface of zbMATH offers specialised search opportunities for finding entries in the huge collection of mathematical publications. Entries can be found by specifying keywords that refer to information about the document, its author or its classification in the MSC2020 [5] as well as other attributes linked to the document. A significant attribute in the context of this work is the search of documents by specifying mathematical formulae. Currently, over 160 million formulae are indexed. The first prototype for a formula search in zbMATH was established in a research collaboration between FIZ Karlsruhe and the Jacobs University Bremen.[2]

Since recent developments in 2021, all zbMATH content is openly available for free to the public domain. Open interfaces enable the integration of other services, e.g., better search functions for full texts from free digital libraries such as arXiv and EuDML. Opening up the content offers another dimension of new applications by linking it to mathematical research data that has been largely isolated and inadequately tapped in the past [18].

[1] https://zbMATH.org

[2] https://zbMATH.org/formulae

Due to the current progress of zbMATH and to constantly improving the formula search for the mathematical community, we present the current work. In this work, we investigate various math search engines and formula-search-applications to outline the state-of-the-art in mathematical formula search. Furthermore, we check the results of our investigation for their applicability to the extension of formula search in zbMATH. On the foundation of our investigation, we present suggestions for several extensions.

This work is organized as follows. First, in Section 2 we provide an overview of math search engines and applications for formula search and point out their major attributes. At the start of the section, we provide a summary of the currently used math search engine, *MathWebSearch*.

In Section 3 we propose extensions for formula search in zbMATH based on the investigation of the search engines and applications in the previous section.

Section 4 renders a brief synopsis, concludes our paper and gives an outlook on the future.

2 Overview of math search applications and engines

In this section, we follow the example of [8, 19] and test real-world math search engines and applications related to formula search with functional demos. For math search engines without a functional demo, we consult literature that evaluates them.

2.1 MathWebSearch

2.1.1 Description. The MathWebSearch system (MWS)[3] is the math search engine currently utilized in zbMATH. MWS was actively developed by the KWARC group at Jacobs University, mostly by C. Prodescu, until the current latest release in 2014.[4]

2.1.2 Functionalities. MWS is a content-based full-text search engine that concentrates on low-latency query responses in interactive applications. It combines exact formula matching with full-text search capabilities for simultaneous search for keywords and formulae. For the full-text queries, it uses Elasticsearch [10].

For creating the search engine index, MWS reads *Content MathML* formatted formulae using a technique derived from automated theorem proving called substitution tree indexing [7].

Most scientific publications (e.g. the arXiv corpus) in STEM fields use LaTeX formula notation [9]. Since LaTeX denotation is quite common, it is used in the web-interface of zbMATH for the input of formulae. To create suitable Content MathML

[3] https://search.mathweb.org

[4] https://github.com/MathWebSearch/mws

queries and enable indexing for input documents, the LaTeXML converter developed by Bruce Miller[5] at NIST is utilized for LaTeX conversion to MathML [15].

MWS offers system components for multiple stages in the process of enabling formula search [10]: A MWS component enables parsing HTML and XHTML documents to annotated XML, which contains document metadata as well as the formula encoded as Content MathML. The annotated XML is read from a folder by the Formula Indexer and a formula search index is created. The Indexer also provides an RESTful API for formula query. For text-based document queries, Elasticsearch is used. For concurrent keyword and formula queries, Elasticsearch and the Indexer are prepended with a proxy. The proxy prioritizes the proportion of search hits in keyword and formula query responses in the final response.

2.1.3 Review. The core engine of MWS is not actively developed. The last MWS release was in December 2014 and since then mainly support for containerisation has been added to the codebase. The source code for MWS is publicly available on GitHub and licensed under GPLv3. In 2014 KWARC was participating with MWS 1.0 in the NTCIR-11 Math-2 Task, which is specially dedicated to information access to mathematical content. It scored above average for retrieval precision among eight task participants [2].

2.2 MathDeck

2.2.1 Description. MathDeck[6] is a math search related application which aims to offer simplified methods for entering formulae [6]. Entered formulae are rendered and can be exported as image files or their LaTeX notation can be forwarded to common math search engines.

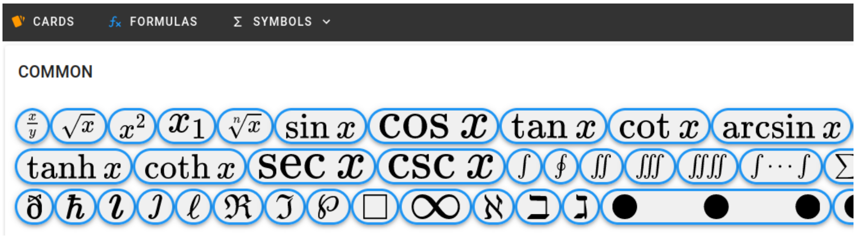


Figure 1. Symbol palette for well-known maths symbols in MathDeck.

[5] https://dlmf.nist.gov/LaTeXML

[6] https://mathdeck.org

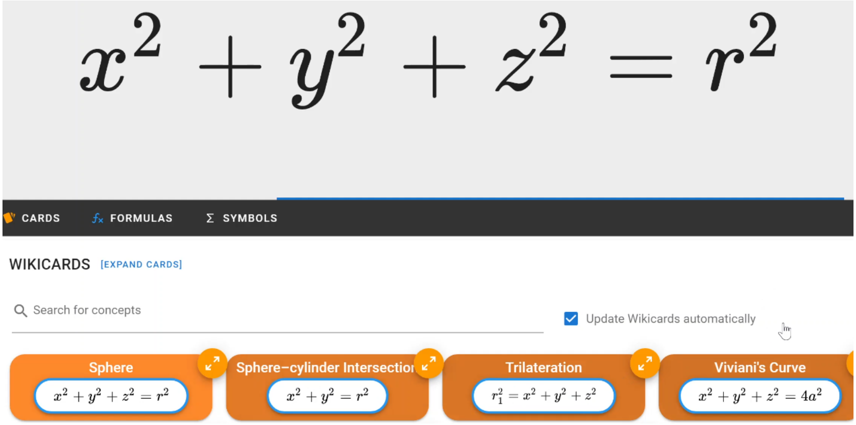


Figure 2. Annotating a formula with Wikidata-concepts in MathDeck.

2.2.2 Functionalities. MathDeck offers possibilities to draw handwritten formulae with a graphics editor or upload formulae in a picture, which then can be converted to LaTeX. It provides a symbol palette with a selection of well-known mathematical symbols to compose LaTeX-formatted formulae (see Figure 1). Also, with its *Wikicards* functionality, MathDeck can automatically link concepts from Wikipedia (via Wikidata) to well-known formulae (see Figure 2) to obtain a label and contextual information [6].

The MathDeck frontend is developed using *Vue.js*,[7] and makes use of a customized MathJax library for rendering math [6]. For obtaining the Wikicards suggestions, a modified version of Tangent-CFT [13] is used.

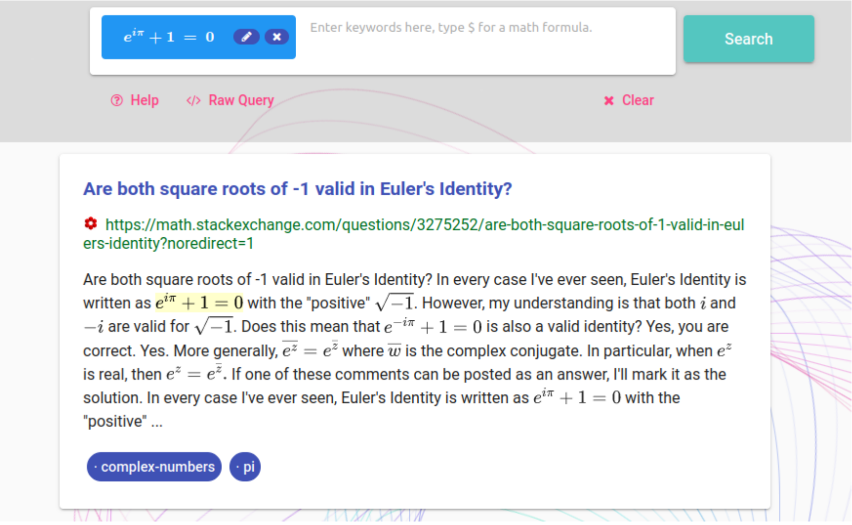
2.2.3 Review. MathDeck provides an example of a modern and advanced user interface for entering math formulae. Also linking, or even search for formulae as semantic concept might be a valuable addition for zbMATH users. MathDeck has been published recently [6] and contains proposals for further enhancements of the Wikicards. To our knowledge, the source code is not open-source. As of this writing, the Wikicards functionality did not suggest any cards for several well-known formulas in a test with multiple popular browsers.

2.3 Approach Zero

2.3.1 Description. Approach Zero[8](AZ) is a math search engine that can search for math expressions and keywords simultaneously. The referenced website also includes

[7] https://vuejs.org
[8] https://approach0.xyz

a demo-application. The search-engine, crawlers and more components are openly available on GitHub.[9]. The core of the search-engine is authorized under the MIT license. There have been no major changes to the core repository since the last release in 2016.



Are both square roots of -1 valid in Euler's Identity?

https://math.stackexchange.com/questions/3275252/are-both-square-roots-of-1-valid-in-eulers-identity?noredirect=1

Are both square roots of -1 valid in Euler's Identity? In every case I've ever seen, Euler's Identity is written as $e^{i\pi} + 1 = 0$ with the "positive" $\sqrt{-1}$. However, my understanding is that both i and $-i$ are valid for $\sqrt{-1}$. Does this mean that $e^{-i\pi} + 1 = 0$ is also a valid identity? Yes, you are correct. Yes. More generally, $\overline{e^z} = e^{\bar{z}}$ where $\overline{w}$ is the complex conjugate. In particular, when e^z is real, then $e^z = e^{\bar{z}}$. If one of these comments can be posted as an answer, I'll mark it as the solution. In every case I've ever seen, Euler's Identity is written as $e^{i\pi} + 1 = 0$ with the "positive" ...

Figure 3. Found formula highlighting in the query results of Approach Zero.

2.3.2 Functionalities. AZ currently indexes the websites theartofproblemsolving.org and math.stackexchange.com. As a formula input format for indexing it exclusively reads LaTeX [24]. AZ uses a special search engine called OPMES (Operation-tree Pruning based Math Expression Search). This parses a math expression into an operator tree. It then extracts leaf root paths from the tree to represent structural information [23]. OPMES was evaluated retroactively with the NTCIR-12 MathIR Wikipedia Formula Browsing Task which is a benchmark for isolated formula retrieval [24]. For most configurations, it reaches higher scores than MCAT and an improved version of Tangent 3 [4], called Tangent-S, in terms of retrieval performance. It also achieved the best results among the systems compared in ARQMath-2 Task 2 for formula retrieval [14].

The user interface of the application has a symbol palette for entering well-known math symbols. The UI also parses input LaTeX while typing and directly renders the formula within the user input. In the list of search results for a query, the found formula in the text is directly highlighted (see Figure 3).

[9] https://github.com/approach0

Further notable features are cache on-disk index and the option to specify the memory usage limit.[10]

2.3.3 Review. The formula search engine can be considered for further investigation for use in zbMATH due to its advanced functions, free licensing and comprehensive documentation. The search results in the demo application provide a very illustrative example for the highlighting of found formulae.

2.4 SearchOnMath

2.4.1 Description. SearchOnMath[11] is a formula search engine equipped with a publicly available web application. It enables to search for combinations of keywords and formulae.

2.4.2 Functionalities. Similar to Approach Zero, the frontend of SearchOnMath has one single-line input field with combined keyword and formula queries. The input format for formulae is LaTeX and MathJax is used for rendering query results. Found formulae in the full text are also highlighted in the query results. The web application indexes a list of popular websites which contain math (e.g. MathOverflow[12]) as well as arXiv. Preview search results are rendered using MathJax. The indexed data can also be queried through an OAS3-specified RESTful API.

2.4.3 Review. SearchOnMath was a research project until 2015, then it became a start-up [17]. Since then, the project is developing in a more commercial direction. To our knowledge, there is no public code repository. This complicates reusability in zbMATH.

2.5 MIaS and EuDML

2.5.1 Description. The Math Indexer and Searcher (MIaS) is a math-aware full-text search engine based on Apache Lucene [20]. The engine differs from conventional search engines in a way that it allows fuzzy formulae search on joint text and math inverted index. The European Digital Mathematics Library[13] (EuDML) is an online library with more than 26,000 indexed items, where MIaS is used as a formula search engine.

[10] https://approach0.xyz/docs/content/en/features.html

[11] https://www.searchonmath.com/about

[12] https://mathoverflow.net

[13] https://eudml.org/search

2.5.2 Functionalities. The search form of EuDML enables one to define a set of keywords and a formula denoted in LaTeX or MathML notation. The search terms can then be combined with boolean operators. The frontend has an element displaying a rendered live preview of the given formula input. In the keyword-input fields, suggestions for common keywords are provided. Also, it shows comprehensive statistics about the overall search-results for a query.

MIaS is openly available on GitHub[14] and it is licensed with Apache 2.0. Also, under the same license, a publicly available web interface (WebMIaS)[15] exists. In all NTCIR-11 Math-2 tasks, it exceeded in its math retrieval performance [2]. The group from Masaryk University also participated in the main tasks NTCIR-12 MathIR. At the time of evaluation with NCTIR-12, MIaS was more in the mid-range of results [22].

2.5.3 Review. Similar to Approach Zero, the MIaS formula search engine can be considered for further investigation for use in zbMATH due to its free licensing and open availability. Also, we consider making the EuDML available through the zbMATH capabilities.

2.6 MCAT

2.6.1 Description. The MCAT group from the National Institute of Informatics in Tokyo participated in the NTCIR MathIR tasks 11 and 12 with an indexing scheme for mathematical expressions within an Apache Solr (Lucene) database. With this scheme, they enabled mathematical expressions searching using queries which contain both formulae and keywords [2, 22].

2.6.2 Functionalities. The method of MCAT reads Presentation as well as Content MathML and utilizes Apache Solr as a full-text search engine. Their search method obtains context window and description of formulae during the indexing process. It includes three levels of granularity for obtaining textual information (math, paragraph, and document levels). Also, it utilizes a dependency graph of mathematical expressions and a post-retrieval re-ranking method [11].

2.6.3 Review. MCAT has achieved excellent results in all tasks at NCTIR-12 [22]. The project is to our knowledge not publicly available, and this could make re-usage more difficult for zbMATH. Technological aspects from the publications describing the system could be extracted to build a custom system.

[14] https://github.com/MIR-MU/MIaS

[15] https://github.com/MIR-MU/WebMIaS

2.7 Tangent based search engines

2.7.1 Description. Tangent was introduced in 2015 as a method for indexing and retrieving mathematical expressions [21]. Since then, many methods have been introduced using Tangent as a baseline. The ARQMath labs, one and two, both contain a task for formula retrieval. These tasks have similarities in design to the NTCIR-12 Wikipedia Formula Browsing task, but differ in a way that relevance is defined contextually and evaluation is based on visually distinct formulae, rather than all formula instances [12].

2.7.2 Review. In the ARQMath formula retrieval tasks, modifications of Tangent were used by the participating teams from *Mathdowsers*[16] (Tangent-L) [16] and *DPRL* (Tangent-S) [12]. Both achieved comparable results.

2.8 Tangent-CFT

2.8.1 Description. Tangent Combined FastText (Tangent-CFT) is an embedding model for mathematical formulas which can be used for mathematical formula retrieval. It makes use of the SLT and OPT formula representations produced by the Tangent-S formula search engine. [13] Also, it utilizes the FastText n-gram embedding model [3]. The source code for Tangent-CFT and further model variations are publicly available[17].

2.8.2 Review. The *TanApp* introduced in [13] leverages linear combined retrieval scores from Tangent-CFT and Approach Zero. With this combination, the formula retrieval precision in the NTCIR-12 formula browsing task can be significantly boosted in comparison to the original Approach Zero. Also, in the same task, Tangent-CFT in its standalone application outperforms MCAT [13]. Tangent-CFT could be utilised as an additional method with MWS in zbMATH, to enhance the retrieval precision.

3 Extension proposal

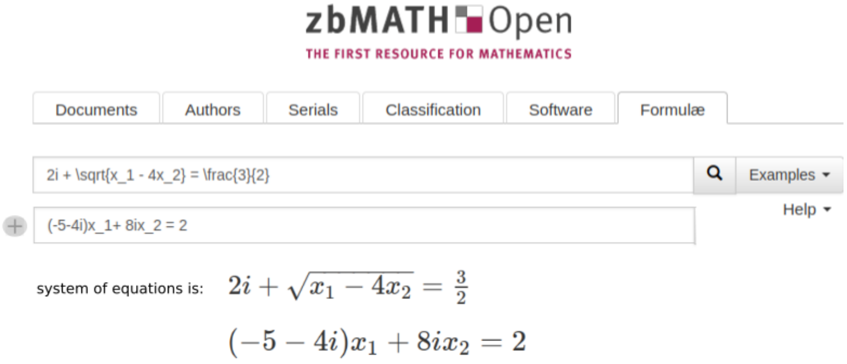
By examining the previously mentioned math engines and applications, several potential extensions for zbMATH were extracted. This section presents the proposals for the new functionalities.

3.1 Search function for systems of equations

In some cases (e.g. in systems of linear equations) a mathematical expression can consist of a collection of multiple equations involving the same set of variables. Traditional math search currently allows querying for a single search term at a time, and

[16] https://github.com/kiking0501/MathDowsers-ARQMath

[17] https://github.com/BehroozMansouri/TangentCFT



2i + \sqrt{x_1 - 4x_2} = \frac{3}{2}

(-5-4i)x_1+ 8ix_2 = 2

system of equations is: $2i + \sqrt{x_1 - 4x_2} = \frac{3}{2}$

$(-5-4i)x_1 + 8ix_2 = 2$

Figure 4. Design of a user input for systems of equations in zbMATH formula search.

in some cases for a Boolean combination of multiple terms. In our formula search, we propose to add a search input element that allows using a consistent set of variables across multiple search terms. Figure 4 shows what such user input for a system of equations could look like in zbMATH in the future.

3.2 Syntax verification

Users may not use the correct LaTeX syntax when entering the formula in the formula search input field. Therefore, we propose an automated syntax check that highlights the errors in the input field or its surroundings.

3.3 Math expression simplification

Mathematical expressions can take different forms while articulating the same functionality. Math search users may not have the simplest form of the formula they want to search for at their fingertips. This is a suggestion for introducing a term simplification within the formula search, which can be optionally activated. This term simplification will also provide the zbMATH user with the simplified input term as an additional output value.

3.4 Improving accessibility

This section proposes several minor features that we propose for improving the accessibility of zbMATH. One of them is the introduction of a symbol palette for the zbMATH math search, containing well-known and frequently used mathematical symbols. When the user selects the visual symbol, its LaTeX representation is added to the search bar. We also consider introducing a user input component for math search that allows drawing math symbols in an image editor embedded in the zbMATH web interface. Examples of both proposed changes can be seen in MathDeck.[18] Another proposed feature is to provide an auto-completion list for the current user input of LaTeX expressions in the zbMATH search bar.

3.5 Semantic concept annotation

Hereby, we propose the annotation of the given formula in the search box of zbMATH to semantic concepts in a knowledge graph (KG). With the human-readable labels to a formula extracted from the KG, keyword suggestions can be realised. Also, additional information about research data obtained from the KG can be linked to the search results. The MathDeck *Wikicards* (see Figure 2) can be viewed as an existing example of concept annotation. We consider annotating semantic concepts and obtaining additional information from the KG of the MaRDI Portal.[19] This KG is currently being set up by FIZ Karlsruhe together with the Zuse Institut Berlin. It will connect vast amounts of mathematical research data and formulae.

3.6 Rendered search results

We propose a highlighting for the found formulae in the query results similar to Approach Zero in zbMATH (see Figure 3). The found formula and the surrounding text from the summaries of the indexed publications will be displayed directly in the list of results. Also, we consider rendering the found formula in this preview.

3.7 Improving retrieval accuracy

Currently, the ranking of documents in the search results of a formula search in zbMATH is realised using similarity scores calculated by MWS. MWS uses traditional tree-based search engine indices. Following the example of [13], we propose improving the formula retrieval accuracy by computing a linear combination of the traditional MWS-based score and a newly introduced score based on formula embeddings such as Tangent-CFT.

3.8 Finding mathematical symbols by facetted search

When typing a formula in the input of math search, the names of mathematical symbols can be unknown or may differ between cultural and lingual contexts. While the

[18] https://www.mathdeck.org

[19] https://portal.mardi4nfdi.de

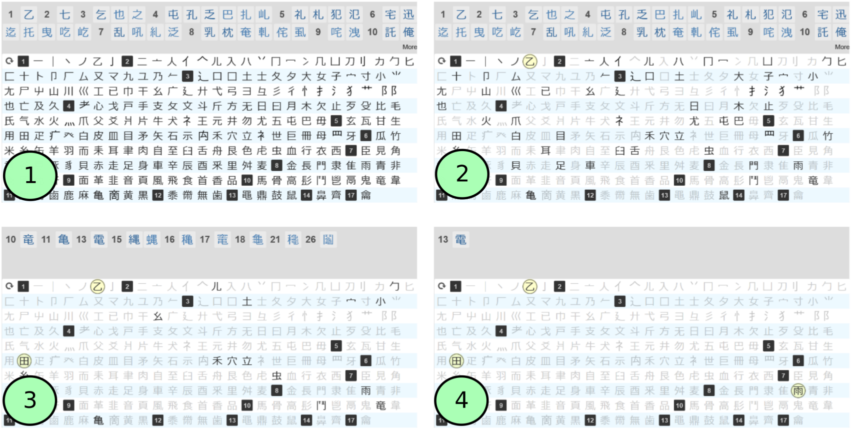


Figure 5. The four stages of Jisho's facetted search [1] to find the 13 stroke Kanji 電 ('*den*'). Selecting radicals in each step (2-4) narrows down the possible search results.

US or Germany uses $\geq$ to express a greater or equal relation, in Japan, the notation $\geqq$ is more common. Considering the sheer amount of different math notations, it might not be obvious to a student from Japan that $\geq$ and $\geqq$ refer to the same relation.

Consider an example case, where a Japanese student, reading a German math publication, encounters the visual appearance of a formula. Lacking the cultural context, the student does not know the meaning or the denomination of the formula since the explicit meaning of symbols is often not mentioned, even in educational literature.

To find an explanation for the formula and its meaning, the researcher could consider using a math search engine to find literature explaining the formula. In a conventional math search engine, the input notation will be difficult, since a written notation (e.g. LaTeX symbols) is not known to the researcher.

One solution to aid the search for mathematical symbols based on their visual appearance is using the aforementioned symbol palette. This can be rather confusing since there is a vast variety of mathematical symbols which will overload the symbol palette.

The problem of searching an unknown math symbol without knowing its name and meaning is significantly related to finding unknown words or morphemes in a natural language that uses logograms, such as *hanzi* in Chinese or *Kanji* in Japanese. In the Japanese writing system, a character (Kanji) can often be composed of one or more combinations of the 214 radicals.[20] The radicals represent smaller basic symbols of the Kanji and also can represent Kanji themselves. Japanese Kanji are often classified by the order of strokes and the radicals they consist of, which allows a

[20] 214 is the number of traditional radicals, but the exact number may vary

reader to find a Kanji fairly easy in a dictionary. The *Jisho* [1] digital library provides a facetted search system to find Japanese characters by their parts (see Figure 5).

To simplify the search for mathematical formula where neither the name nor the meaning is known to the user, we propose a new approach for defining the search input. The user input will be defined by adopting the Logogram classification of Chinese and Japanese characters. This can be realized by deriving a limited set of basic visual components of mathematical symbols. Like the mentioned radicals, these can be used as base symbols to implement a faceted search system similar to *Jisho*, which can find mathematical expressions instead of Kanji characters.

Engine	Input Formats	Availability	Technology	Evaluated	Applications
MathWebSearch [10]	CMML, LaTeX (with converter)	Public repo, GPLv3	Substitution tree indexing	NTCIR-11 and 12 ARQMATH-1 and 2	zbMATH Open, Mediawiki Extension
Approach Zero [24]	LaTeX	Public repo, MIT-license	Operation-tree pruning based math expression search	Retroactively NTCIR 12, ARQMATH-1 and 2	Approach Zero
SearchOnMath [17]	LaTeX	Public API, no code repo	Lexical analysis, degree of similarity	SearchOnMath has been used to evaluate hardware performance	SearchOnMath
Tangent CFT [13]	LaTeX	Public repo license N/A	Embedding model for mathematical formulae	Retroactively NTCIR-12 ARQMATH-1 and 2	—
MIaS [20]	CMML, PMML, LaTeX	Public repo Apache 2.0	Lucene index for text, variables and constants unification for formula ordering	NTCIR-11 and 12	EUDML
Notes: CMML stands for Content Math ML, PMML stands for Presentation Math ML					

Table 1. Overview of the core math search engines and applications investigated and their main attributes

4 Conclusion and outlook

For our investigation, we examined various math search engines and formula search related applications. A summary of the main engines and their applications and attributes can be found in Table 1. We found several suggestions for improvements to zbMATH. We have collected suggestions for extensions to the user interface that allow for increased accessibility, as well as for linking search terms and search results with additional information from knowledge graphs. Furthermore, from the investigation of several search engines, we proposed that retrieval performance can be improved by linear combination of traditional tree-based ranking scores with a scoring based on formula-embeddings.

The third ARQMath lab[21] announced for the middle of 2022 can be a source for future evaluations of formula search engines. Synergy effects in the implementation of extensions to zbMATH regarding the knowledge graph and data linking may result from the MaRDI portal and the associated KG, which have been under construction since the beginning of 2022.

References

[1] K. Ahlström, M. Ahlström, and A. Plummer, Jisho.org: Japanese dictionary. https://Jisho.org visited on 17 March 2024

[2] A. Aizawa, M. Kohlhase, I. Ounis, and M. Schubotz, NTCIR-11mmath-2 task overview. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, December 9–12, 2014*, edited by N. Kando, H. Joho, and K. Kishida, pp. 88–98, National Institute of Informatics (NII), 2014

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5** (2017), 135–146

[4] K. Davila, R. Zanibbi, A. Kane, and F. W. Tompa, Tangent-3 at the NTCIR-12 MathIR task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences, Tokyo, June 7–10, 2016*, edited by N. Kando, T. Sakai and M. Sanderson, pp. 338–345, National Institute of Informatics (NII), 2016

[5] MSC2020 Mathematics Subject Classification System. 2021 https://zbmath.org/static/msc2020.pdf visited on 17 March 2024

[6] Y. Diaz, G. Nishizawa, B. Mansouri, K. Davila, and R. Zanibbi, The MathDeck formula editor: Interactive formula entry combining LaTeX, structure editing, and search. In *CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing System*, edited by Y. Kitamura, A. Quigley, K. Isbister, and T. Igarashi, pp. 1–5, Association for Computing Machinery, New York, NY, 2021

[7] P. Graf, Substitution tree indexing. In International Conference on Rewriting Techniques and Applications RTA 1995: Rewriting Techniques and Applications, edited by J. Hsiang, pp. 117–131, Lecture Notes in Computer Science 914, Springer, Berlin, 1995

[21] https://www.cs.rit.edu/~dprl/ARQMath

[8] F. Guidi and C. Sacerdoti Coen, A survey on retrieval of mathematical knowledge. In *International Conference on Intelligent Computer Mathematics CICM 2015: Intelligent Computer Mathematics*, edited by M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, pp. 296–315, Lecture Notes in Computer Science 9150, Springer, Cham, 2015

[9] R. Hambasan and M. Kohlhase, Faceted search for mathematics. In *International Conference on Mathematical Aspects of Computer and Information Sciences MACIS 2015: Mathematical Aspects of Computer and Information Sciences*, edited by R. Hambasan and M. Kohlhase, pp. 406–420, Springer, Cham, 2016

[10] R. Hambasan, M. Kohlhase, and C.-C. Prodescu, nMathWebSearch at NTCIR-11. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, December 9–12, 2014*, edited by N. Kando, H. Joho, and K. Kishida, pp. 114–119, National Institute of Informatics (NII), 2014

[11] G. Y. Kristianto, G. Topić, and A. Aizawa, MCAT Math Retrieval System for NTCIR-12 MathIR Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, June 7–10, 2016*, edited by N. Kando, T. Sakai and M. Sanderson, pp. 323–330, National Institute of Informatics (NII), 2016

[12] B. Mansouri, D. Oard, and R. Zanibbi, DPRL systems in the CLEF 2020 ARQMath Lab. 2020. In *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, Thessaloniki, September 22–25, 2020, edited by L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol, article no. 223, CEUR Workshop Proceedings 2696, CEUR-WS.org, 2020*

[13] B. Mansouri, S. Rohatgi, D. W. Oard, J. Wu, C. L. Giles, and R. Zanibbi, Tangent-CFT: An embedding model for mathematical formulas. In *ICTIR '19: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, Santa Clara, CA, October 2–5, 2019, edited by Y. Fang, Y. Zhang, J. Allan, K. Balog, B. Carterette, and J Guo,* pp. 11–18, Association for Computing Machinery, New York, NY, 2019

[14] B. Mansouri, R. Zanibbi, D. W. Oard, and A. Agarwal, Overview of ARQMath-2 (2021): Second CLEF lab on answer retrieval for questions on math. In *Experimental IR meets multilinguality, multimodality, and interaction. 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, edited by K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, and N. Ferro, pp. 215–238, Springer, Cham, 2021

[15] F. Müller and O. Teschke, Full text formula search in zbMATH. *Eur. Math. Soc. Newsl.* **102** (2016), 51–51

[16] Y. K. Ng, D. J. Fraser, B. Kassaie, G. Labahn, M. S. Marzouk, and F. Wm. Tompa, and K. Wang, Dowsing for math answers with tangent-L. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, September 22–25, 2020*, edited by L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, article no. 16, CEUR Workshop Proceedings 2696, CEUR-WS.org, 2020

[17] R. M. Oliveira, F. B. Gonzaga, V. C. Barbosa, and G. B. Xexéo, A distributed system for SearchOnMath based on the Microsoft BizsPark program. In *2018: Proceedings of the 33rd Brazilian Symposium on Databases, edited by B. F. Lóscio, pp. 289–294, Sociedade Brasileira de Computação, 2018*

[18] M. Petrera, D. Trautwein, I. Beckenbach, D. Ehsani, F. Müller, O. Teschke, B. Gipp, and M. Schubotz, zbMATH Open: API solutions and research challenges. In *DISCO 2021: Digital Infrastructures for Scholarly Content Objects 2021*, edited by W.-T. Balke, A. de Waard, Y. Fu, B. Hua, J. Schneider, N. Song, X. Wang, pp. 4–13, CEUR Workshop Proceedings 2976, CEUR-WS.org, 2021

[19] M. Schubotz, *Augmenting mathematical formulae for more effective querying & efficient presentation*. Ph.D. thesis, Technical University of Berlin, 2017, Epubli, Berlin, 2017

[20] P. Sojka and M. Líška, The art of mathematics retrieval. In *DocEng '11: Proceedings of the 11th ACM symposium on document engineering*, edited by M. Hardy, pp. 57–60, Association of Computing Machinery, Mountain View, CA, 2011

[21] D. Stalnaker and R. Zanibbi, Math expression retrieval using an inverted index over symbol pairs. In *Document recognition and retrieval XXII*, edited by E. K. Ringger and B. Lamiroy, pp. 34–45, Proceedings Vol. 9402, International Society for Optics and Photonics, SPIE, 2015

[22] R. Zanibbi, A. Aizawa, M. Kohlhase, I. Ounis, G. Topić, and K. Davila, NTCIR-12 MathIR task overview. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, June 7–10, 2016*, edited by N. Kando, T. Sakai and M. Sanderson, pp. 299–308, National Institute of Informatics (NII), 2016

[23] W. Zhong and H. Fang, OPMES: A similarity search engine for mathematical content. In *European Conference on Information Retrieval ECIR 2016: Advances in Information Retrieval*, edited by N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Nunzio, C. Hauff, and G. Silvello, pp. 849–852, Lecture Notes in Computer Science 9626, Springer, Cham, 2016

[24] W. Zhong and R. Zanibbi, Structural similarity search for formulas using leaf-root paths in operator subtrees. In *European Conference on Information Retrieval ECIR 2019: Advances in Information Retrieval*, edited by L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, pp. 116–129, Lecture Notes in Computer Science 11437, Springer, Cham, 2019

Chapter 8

Author identification through and for interconnectivity: A brief history of author identification at zbMATH Open

Nicolas D. Roy

1 Introduction

Author identification, or *author disambiguation*, is the process of matching an *authorship record*, i.e. an author name string in a publication, with an entity in a database of persons. The person entity is usually defined by a set of metadata, like for example name, birthdate, affiliation, email, etc., but also by a collection of other publications authored by the person. The link between authorship record and person entity is called an *authorship assignment*.

The task of attributing a given publication to a certain author based only on person name is notoriously difficult because of the following reasons:

- *Incompleteness*. The name contained in an authorship record may be abbreviated (often the given name) or even missing (e.g. a second/middle name). In the last centuries it was also not uncommon to publish under the family name only. On the other hand, it is almost impossible to completely avoid any kind of data corruption, and that could lead to an erroneous author name in a publication.
- *Synonymy*. Different names might refer to the same individual. A name change after marriage is a common source of such name variability, but zbMATH Open provides many other examples due to the use of different historical transliteration rules, e.g. in the Russian literature.[1]
- *Homonymy*. Different persons might bear the same name. The most famous examples come probably from the Chinese language, since just the top three surnames *Wang* (王), *Li* (李), and *Zhang* (張; 张)[2] cover more than 20% of the population [2]. But European languages provide also examples like *Peter Müller*[3] or *Andrzej Nowak*[4].

The last decades have seen the emergence of various person-centered databases. As a result, interconnectivity has become a staple feature of many online services and

[1] The case of https://zbmath.org/authors/chebyshev.p-1 is a particularly illustrative example of this phenomenon

[2] See for example https://zbmath.org/authors/?q=wang.wei

[3] https://zbmath.org/authors/?q=müller, peter

[4] https://zbmath.org/authors/?q=nowak,andrzej

information databases. In the domain of disambiguation of publication authorships, interconnectivity is on the one hand a very valuable by-product of the author identification, since well-identified author entities allow for a reliable matching with other similar person-based databases. But on the other hand, interconnectivity itself can constitute a powerful component of any author disambiguation system, through the harvesting of additional data relevant to author disambiguation (biographical, bibliographic, ...) from other related services, as soon as a reliable matching between both databases can be established.

2 Author identification workflow at zbMATH Open

As of October 2021, zbMATH Open indexes approximately 4.3 million documents, corresponding to about 8 million authorship records that are linked to the author database, containing currently more than 1.1 million items.

The author disambiguation at zbMATH Open is the result of a complex interaction between several algorithmic tools and user interfaces, where each module enriches the capabilities of the others.

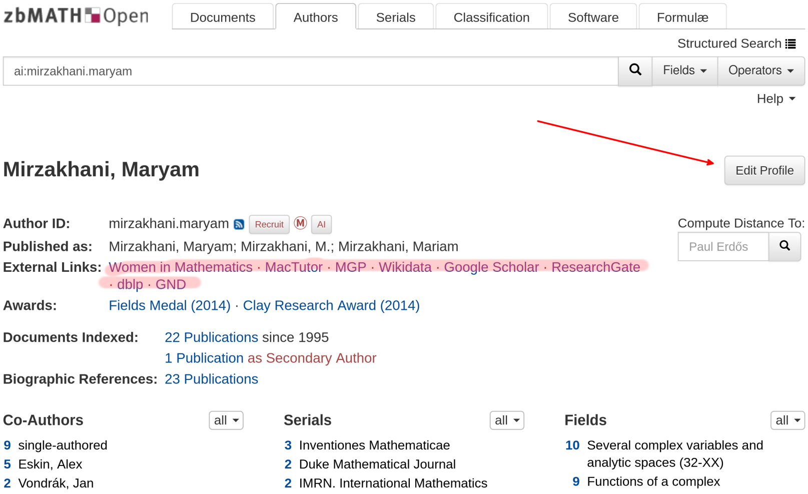
Automatic disambiguation

Approximately 78% of the authorship records are handled by an algorithm mainly based on a name-similarity feature, which is well fitted to the particularities of the zbMATH Open dataset (taking into account for example the variability in transliterations from Cyrillic script). This name similarity is refined by time and coauthor similarity features. The time-similarity feature relies heavily on the presence of biographical metadata associated to the author entities. This biographical data is for a big part harvested from external services, as described in the next section. The automatic identification process runs daily in order to handle the newly indexed documents (ca. 600 per day), but also because, in principle, the disambiguation of a given document might have some influence on the disambiguation of other documents on the next day.

Community Author Identification Interface

Like every automatic process, the author disambiguation algorithm produces various errors: authorship records wrongly attributed to another author, publications of a given person incorrectly split into several author profiles (as a typical result of the above-mentioned ‘synonymy’ issue), or publications of several persons incorrectly mixed into one single profile (as a typical result of the above-mentioned ‘homonymy’ issue).

To circumvent that, zbMATH Open has been offering since summer 2014 an Author Identification Interface [1], accessible through the button *Edit Profile* at the top right of any author profile, and allowing every zbMATH Open user to send some correction requests. Since then, more than 8,000 user requests have been sent with this tool. Besides the possibility to correct authorship assignments or to merge together several wrongly split author profiles, the interface allows also for adding *external links* to other related services, like e.g. Wikidata, ORCID, Mathematics Genealogy Project (MGP), etc. (see Figure 1).



Mirzakhani, Maryam

Author ID: mirzakhani.maryam

Published as: Mirzakhani, Maryam; Mirzakhani, M.; Mirzakhani, Mariam

External Links: Women in Mathematics · MacTutor · MGP · Wikidata · Google Scholar · ResearchGate · dblp · GND

Awards: Fields Medal (2014) · Clay Research Award (2014)

Compute Distance To:

Documents Indexed: 22 Publications since 1995
1 Publication as Secondary Author

Biographic References: 23 Publications

Co-Authors	Serials	Fields
9 single-authored	3 Inventiones Mathematicae	10 Several complex variables and analytic spaces (32-XX)
5 Eskin, Alex	2 Duke Mathematical Journal	9 Functions of a complex
2 Vondrák, Jan	2 IMRN. International Mathematics	

Figure 1. Maryam Mirzakhani's author profile showing several *external links* as well as the button to enter the Author Disambiguation Interface.

Every correction request is examined by the zbMATH Open Author Identification team and is visible at zbmath.org usually within 1 or 2 days.

3 Interaction with the scientific publication landscape

Matching processes

The external links provided by community users through the public author identification interface are complemented by several matching tools, based on name similarity but also on various other features adapted to the data available in the considered services. For example, the ORCID matching (ca. 30,000 matched items) highly relies on DOIs, while the matching with MGP (ca. 50,000 matched items) is based on a *collaborator similarity* between the publication coauthors on the zbMATH Open side and the PhD advisors and students on the MGP side.

Data harvesting and spreading tools

The connection with other relevant services[5] not only increases the visibility of the respective services and the Internet presence of the authors,[6] but also allows for very valuable mutual data enrichment and data quality improvement. This is achieved through several automatic data harvesting and spreading tools:

- *Data spreading tools*. An automatic process checks every night the presence of new wikidata IDs in the zbMATH Open author database, and incorporates the involved zbMATH Open author IDs into the corresponding Wikidata profile when necessary.
- *Data harvesting tools*. Every night, the external services linked to any of the zbMATH Open author entities are queried for the presence of any new relevant information, like biographical data (birth year, PhD year, etc.), scientific information (awards, etc.), or other external links.

Integration of harvested data into author disambiguation

The additional data automatically harvested from partner services is then incorporated into the corresponding zbMATH Open author entity, and it is subsequently used to improve the performance of the algorithmic author identification (particularly the time-similarity feature).

Figure 2 is a hypothetical but realistic example of how the combination of matching processes and data harvesting can lead to major changes in the author disambiguation, as described below:

- Assume that at the beginning, many papers authored by a person named 'María López' published in the years 1978, 1983, 1990, 2001, 2011, 2013, and 2017, are grouped together by the author identification algorithm in a profile with id `lopez.maria`.
- Suppose that in Step 1 a matching with the ORCID database would return an ORCID entity with ID `1234-5678-9876-5432` because of name similarity and DOI concordance with the later papers (2011–2017).
- The data harvesting tool would then in Step 2 query the Wikidata API and may find a suitable Wikidata person entity with ID `Q12345678`, together with a MGP entity with ID 343434.

[5] The currently supported services are: http://www.mathnet.ru, https://mathscinet.ams.org, https://dblp.uni-trier.de, https://www.wikidata.org, https://orcid.org, https://portal.dnb.de, https://www.idref.fr, https://www.mathgenealogy.org, https://www.researchgate.net, https://scholar.google.com, https://arxiv.org, https://mathoverflow.net, https://celebratio.org, https://mathshistory.st-andrews.ac.uk, https://www.agnesscott.edu/lriddle/women/alpha.htm,

[6] Currently there are more than 220,000 external links in zbMATH Open author profiles

- In Step 3, the MGP database would be queried and the PhD year 2012 of this entity 343434 would be fetched and incorporated into the zbMATH Open author profile `lopez.maria`.
- This important biographical data would then be exploited in the next run of the author disambiguation algorithm, and would lead to the exclusion of the earlier papers (1978–2001) from the profile `lopez.maria` for time incompatibility. This would result in a splitting of the original author profile into two separate author entities with same name ‘María López’, where the first one (`lopez.maria.1`) would contain the later papers (2011–2017), the ORCID, Wikidata and MGP IDs and the PhD year, whereas the second author entity (`lopez.maria.2`) would gather the older papers (1978–2001).

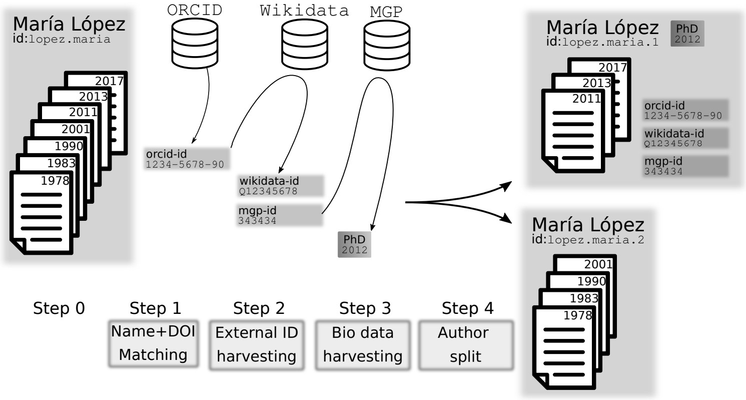


Figure 2. Integration of different harvested data into the author disambiguation.

4 Future developments

The current implementation of the author disambiguation algorithm at zbMATH Open has somehow reached its limit. In particular, the constant adjustment and fine-tuning of the name-similarity procedure to the peculiarities of the zbMATH Open dataset (e.g. the high name variability occurring in transliterations from Russian), has led to a very heuristic and possibly over-fitted algorithm, which is difficult to maintain and improve.

Acknowledging this observation(s) was the starting point of the project *ScAD* (Scalable Author Disambiguation for Bibliographic Databases)[7] in cooperation with *Schloß Dagstuhl* and the *Heidelberg Institute for Theoretical Studies*, launched in

[7] https://www.dagstuhl.de/en/about-dagstuhl/projects/author-disambiguation/

2015, whose legacy is the development of a new framework for author disambiguation (called *KafkAdam*), fully parallelized, highly modular and scalable. It will allow to improve the quality of the author disambiguation through the easy integration of new similarity features (topic analysis, citations, etc.), the possible use of machine learning, and a more efficient and more automated interconnection with other services.

References

[1] H. Mihaljević-Brandt and N. Roy, zbMATH author profiles: Open up for user participation. *Eur. Math. Soc. Newsl.* **93** (2014), 53–55

[2] Wikipedia, Chinese name. https://en.wikipedia.org/wiki/Chinese_name visited on 14 March 2024

Chapter 9

zbMATH Open and community platforms

Isabel Beckenbach

Since the 1st of January 2021, zbMATH is an open access platform called zbMATH Open. It allows every mathematician to freely access zbMATH Open from anywhere in the world. The transition to an open access platform does not only mean free access but should give further benefit to the mathematical community by connecting zbMATH data with information systems of research data, collaborative community platforms, funding agencies and so on. This article focuses on the connection of zbMATH Open to MathOverflow and arXiv, which are the two most used community platforms in mathematics.

1 MathOverflow

MathOverflow describes itself as "a question and answer site for professional mathematics."[1] It is mainly used for asking questions on mathematical research but also for literature or reference requests, questions on the history of mathematics or about mathematical publishing, and many more. Some MathOverflow questions even inspire mathematical research.[2] An example is the question "Does every polyomino tile $\mathbb{R}^n$ for some n?"[3] Vytautas Gruslys, Imre Leader, and Ta Sheng Tan prove in their article "Tiling with arbitrary tiles" that this is indeed the case and even cite this MathOverflow question in the reference section, see [1].

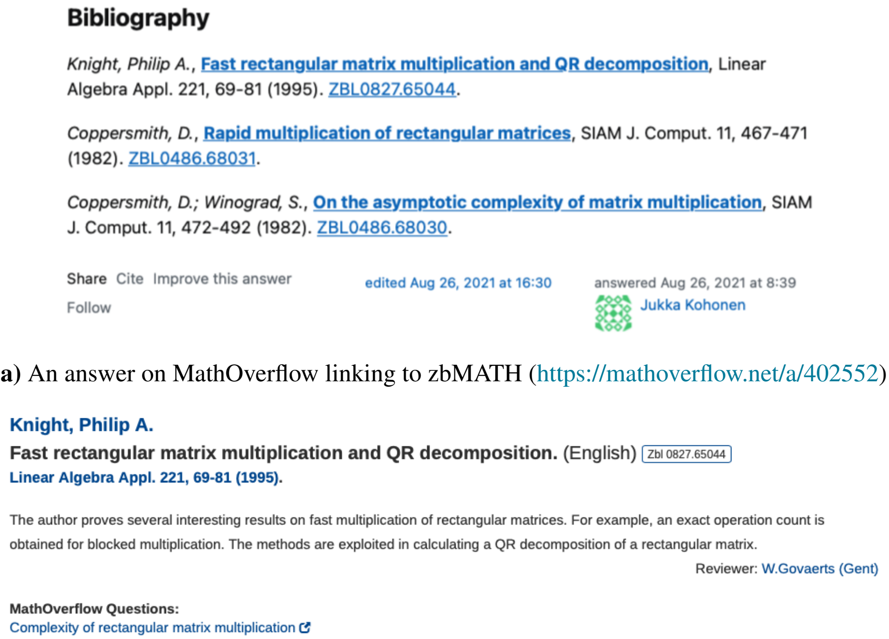
In a joint project with MathOverflow we added the possibility to cite zbMATH records directly in a MathOverflow post using an "Insert Citation" button. One starts typing a reference and the most similar zbMATH records are generated from which the user can choose the best matching one. More details are given in [2].

Figure 1 shows an example of a linking between zbMATH and MathOverflow. The zbMATH citations on the MathOverflow website are linked to the corresponding zbMATH record. On the zbMATH side, we use the StackExchange API to generate links to MathOverflow posts citing a zbMATH record. We also find links added manually and not by the "Insert Citation" functionality.

[1] https://mathoverflow.net/tour

[2] see https://meta.mathoverflow.net/q/617

[3] https://mathoverflow.net/q/49915



Bibliography

Knight, Philip A., Fast rectangular matrix multiplication and QR decomposition, Linear Algebra Appl. 221, 69-81 (1995). ZBL0827.65044.

Coppersmith, D., Rapid multiplication of rectangular matrices, SIAM J. Comput. 11, 467-471 (1982). ZBL0486.68031.

Coppersmith, D.; Winograd, S., On the asymptotic complexity of matrix multiplication, SIAM J. Comput. 11, 472-492 (1982). ZBL0486.68030.

Share Cite Improve this answer
Follow
edited Aug 26, 2021 at 16:30
answered Aug 26, 2021 at 8:39
Jukka Kohonen

(a) An answer on MathOverflow linking to zbMATH (https://mathoverflow.net/a/402552)

Knight, Philip A.

Fast rectangular matrix multiplication and QR decomposition. (English) Zbl 0827.65044

Linear Algebra Appl. 221, 69-81 (1995).

The author proves several interesting results on fast multiplication of rectangular matrices. For example, an exact operation count is obtained for blocked multiplication. The methods are exploited in calculating a QR decomposition of a rectangular matrix.

Reviewer: W.Govaerts (Gent)

MathOverflow Questions:
Complexity of rectangular matrix multiplication

(b) Backlink on zbMATH to the MathOverflow answer above (https://zbmath.org/0827.65044)

Figure 1. Example of a bidirectional linking between zbMATH and MathOverflow.

Furthermore, on the author profile page we display links to the author's MathOverflow user page. Currently 304 author profiles have links to their respective MathOverflow user profiles, which were all added manually. Everyone can edit the information on a zbMATH author profile via a public interface (click on the "Edit Profile" button at an author page). The suggested changes are checked, and applied if they are correct.

The ongoing development of several APIs will give rise to new possibilities in cooperation with MathOverflow. Several ideas are discussed in [3]. For example, one could compare the tags used on MathOverflow and the curated keywords used at zbMATH Open. It might be possible to recommend useful citations based on the tags given in a MathOverflow post or to generate tags automatically. Another idea would be to give users the possibility to connect their MathOverflow profile with the one on zbMATH Open. This would allow to display the publications and reviews of a user on its MathOverflow page.

2 arXiv

arXiv is a preprint server for mathematical articles and related fields as physics, computer science and economics. It is widely used and accepted in the mathematical community. Even some “arXiv overlay journals” exist which do not publish articles themselves but just link to the corresponding arXiv preprint. The refereeing process is similar to the one for non-overlay journals and is carried out by an editorial board. An example is the journal “Discrete Analysis” which has some well-known and respected mathematicians in its editorial board, e.g. the Fields Medal winners Timothy Gowers and Terence Tao.[4] This shows the wide acceptance and importance of arXiv in the mathematical community.

Some articles indexed at zbMATH Open already contain links to their corresponding arXiv preprint. These links were added manually or thanks to information provided by the publishers. However, many arXiv preprints are still missing, which should be changed in the future. Therefore, we developed an algorithm which finds an arXiv preprint for a given zbMATH article if one exists.

It is already possible to search for an arXiv identifier on zbMATH Open, using the syntax `en:<arXiv-id>`, where `<arxiv-id>` might contain the prefix “arxiv:”. If there exists a corresponding zbMATH article linking to `<arxiv-id>`, then the search will return this article.

We are working on adding other open access full texts using information from unpaywall.[5] Having open access to the full-text of an article is not only important for mathematicians who might not have a subscription to some journals, but it also gives new possibilities, for example full-text search including formulae. Right now, the formula search[6] at zbMATH Open only searches in the abstracts and reviews. In the future we plan to expand the search to full-texts.

3 JabRef

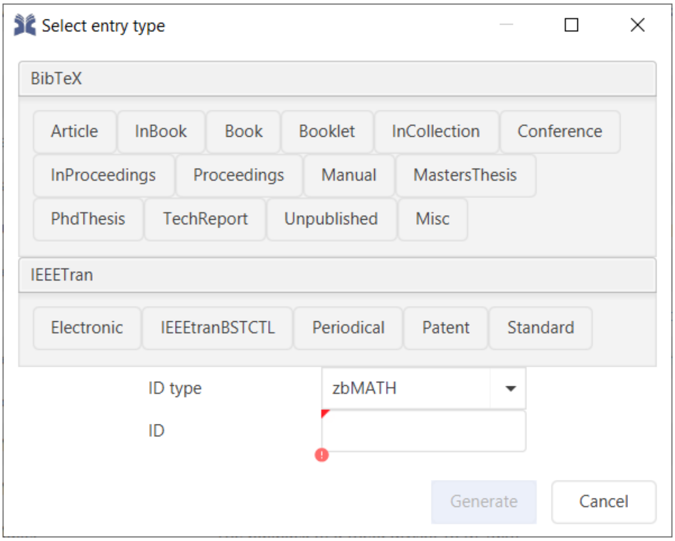
JabRef[7] is an open source scientific reference management system which manages BibTeX files. It offers the possibility to import references from many online scientific catalogues. One of them is zbMATH Open. Since zbMATH became open, it is possible to fetch bibliographic data from zbMATH without subscription. JabRef supports three different possibilities to get bibliographic information from zbMATH Open:

[4] https://discreteanalysisjournal.com

[5] https://unpaywall.org

[6] https://zbmath.org/formulae

[7] https://www.jabref.org



ID type: zbMATH

ID:

Figure 2. Adding bibliographic information via a zbMATH identifier in JabRef.

- Fetch the BibTeX file for a given article by its Zbl number (see Figure 2).
- Use a structured search to fetch all results of that search.
- Enrich an existing BibTeX file with bibliographic information from zbMATH.

In particular, the second option is very useful. An easy example would be to get all references for articles written by a given author. However, one can create more complex search queries. JabRef supports most of the query syntax of zbMATH, however, there are some differences. For example, one has to use `author:<name>` instead of `au:<name>` to search in the authors field. We refer to the documentation of JabRef for details of the query syntax.[8]

4 Conclusion

zbMATH Open incorporates information from community platforms such as MathOverflow. On the other hand, information from zbMATH is used in the open source project JabRef. In the future there will be much more possibilities to integrate data from zbMATH Open with data from further external partners. The OAI-PMH API[9]

[8] https://docs.jabref.org/collect/import-using-online-bibliographic-database

[9] https://oai.zbmath.org

for zbMATH Open already provides a subset of the zbMATH data under the CC-BY-SA 4.0 license which enables diverse use cases.

We are looking forward learning about the ideas and needs of the mathematical community for developing useful tools for researchers in mathematics.

References

[1] V. Gruslys, I. Leader, and T. S. Tan, Tiling with arbitrary tiles. *Proc. Lond. Math. Soc. (3)* **112** (2016), no. 6, 1019–1039

[2] 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** (2019), 50–52

[3] M. Petrera, D. Trautwein, I. Beckenbach, D. Ehsani, F. Müller, O. Teschke, B. Gipp, and M. Schubotz, zbMATH Open: API solutions and research challenges. In *DISCO 2021: Digital Infrastructures for Scholarly Content Objects 2021, edited by W.-T. Balke, A. de Waard, Y. Fu, B. Hua, J. Schneider, N. Song, X. Wang, pp. 4–13, CEUR Workshop Proceedings 2976, CEUR-WS.org,* 2021

Chapter 10

API solutions at zbMATH Open

Matteo Petrera, Fabian Müller, Moritz Schubotz, and Olaf Teschke

1 Introduction

Since January 2021 zbMATH Open[1] is open for public access. For the mathematics community this means open access to the available literature from anywhere in the world without subscription or authentication. Additionally, we are spending efforts to connect the open data of zbMATH Open with other information systems, collaborative platforms, and funding agencies, as outlined in [3, 6]. We expect that our commitment in disseminating mathematics research literature will considerably increase both the range of our target audience and the visibility of mathematics.

Recently we developed Application Programming Interface (API) solutions to facilitate and optimize the open access to mathematical research data. The main purpose of this contribution is to provide some details about these developments, thus extending our previous publications [5, 6].

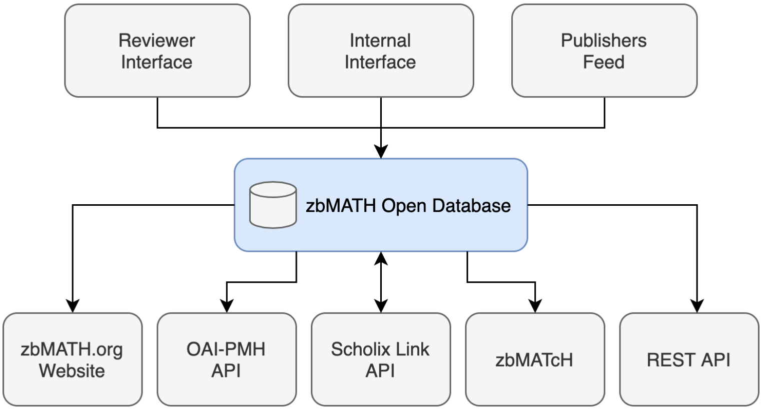


Figure 1. Overview of the zbMATH Open database and its associated data flows.

It is worth to sketch in Figure 1 a conceptual overview of our services in order to illustrate the current and future state of zbMATH Open. This will help the reader in understanding the overall structure of both zbMATH Open and the paper itself. In

[1] https://zbmath.org

Figure 1 we represent the data entering and leaving our database. The boxes called 'Reviewer Interface', 'Internal Interface', 'Publishers Feed', and 'zbMATH.org Website' show well-established components of zbMATH Open and are outside the scope of this paper. The box called 'OAI-PMH API' refers to the harvesting API that was released in April 2021. This will be shortly discussed in Section 5 and we refer to [6] for further and more technical details. The box 'Scholix Link API' refers to an API that is in the staging phase and will be deployed very soon. This will be discussed in Section 2, but we mention that a preliminary version of it has been already presented in [5]. Section 4 is devoted to the 'zbMATH Citation Matcher' (labelled 'zbMATcH' in Figure 1), that consists of an HTML interface designed for manual use, as well as an API. A part of the zbMATH Citation Matcher is a MathOverflow endpoint, discussed in Section 3. Finally, an open 'REST API' is currently in the planning stage. Some information about it will be provided in Section 6.

The motivation behind the recent implementation of APIs at zbMATH Open is twofold. On the one hand, we want to provide the community with efficient tools to benefit from the 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 digital resources. Both motivations offer an opening for potential research opportunities, both on our part and on the part of any institutions interested in our data.

The potential users of our API endpoints may be clustered into five categories:

(1) Bibliographic consumers (MathOverflow, Wikimedia, arXiv, etc.) displaying references to scientific publications;

(2) Aggregators (research data infrastructures, OpenAIRE, SemanticScholar, etc.) extracting data to be then standardized for specific data models;

(3) Archives (research/software digital archives, etc.) typically interested to digitally preserve optimised and standardised flows of data;

(4) Search engines, e.g., Google, implementing the OpenSearch standard;

(5) Researchers interested in the literature for research purposes.

Let us remark that, before zbMATH became an open access web service, the main category of users interested in our product was represented by (5), namely researchers needing access to the literature. It is clear that with zbMATH becoming open the target audience has expanded incredibly.

To conclude, our main goals for the future can be summarised as follows:

- To be a modern and open reference tool for research data in mathematics;
- To promote a functional connection with external information systems of research data;
- To maximise the visibility and discoverability of research in mathematics.

2 Scholix Link API

The Scholix Link API is currently in the staging phase and it will be deployed very soon. A beta version of it is available for public testing[2] and we recently presented it in [5] in occasion of the DISCO2021 workshop at the ACM/IEEE Joint Conference on Digital Libraries.

The main purpose of this API is to document the interconnections (more specifically, *links*) between zbMATH Open and external platforms (called *partners*) which display and use documents indexed in the zbMATH Open database. Potential partners are (see Figure 2):

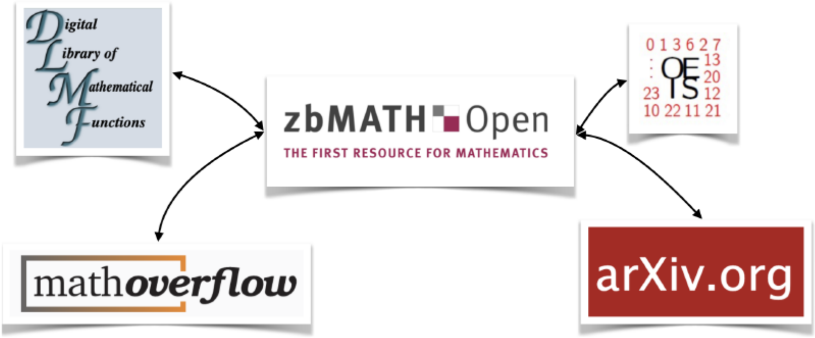

Digital Library of Mathematical Functions
zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS
OEIS
mathoverflow
arXiv.org

Figure 2. zbMATH Open and some of its partners.

- MathOverflow.[3] This is a question-and-answer platform for mathematics that is part of the StackExchange Network.[4] In a previous collaboration, zbMATH Open and MathOverflow added the possibility to cite entries of zbMATH Open in a MathOverflow post directly; see [4] and Section 3;
- arXiv.[5] arXiv is one of the most used open-access repositories of electronic preprints in mathematics. Roughly 250,000 zbMATH Open records contain links to specific arXiv preprints that were added manually, matched algorithmically, or provided by the publishers;
- Online Encyclopedia of Integer Sequences.[6] This a renowned online database of sequences of numbers launched in November 2010. It currently contains more

[2] https://zblink.formulasearchengine.com/links_api

[3] https://mathoverflow.net

[4] https://stackexchange.com

[5] https://arxiv.org

[6] https://oeis.org

than 340,000 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, etc.;

- Digital Library of Mathematical Functions.[7] Please see Section 2.1 for further details.

Search engines or researchers from mathematics or the field of bibliometric research can use this API to present and use the search results. Furthermore, the source code of our API has been released in the form of a public Python package,[8] 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 users.

2.1 DLMF as a zbMATH Open partner

Among the possible platforms that interact with zbMATH Open, we selected the 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 therefore has been 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,’[9] edited by M. Abramowitz and I. A. Stegun in 1964, into a modern and functional digital library. As the title of the original book 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 standardised notation, together with tables, figures, and references to peer-reviewed papers and books. It was published online in May 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 almost 3,000 references[10], out of which about 75% are directly linked to zbMATH Open.[11]

[7] https://dlmf.nist.gov

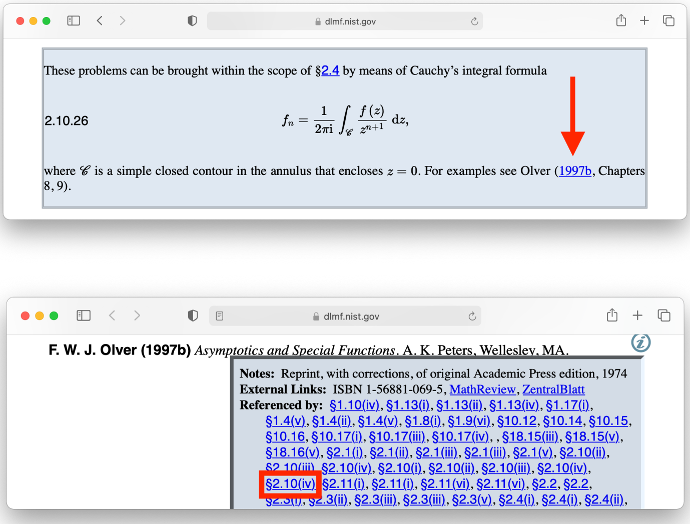
[8] https://github.com/zbMATHOpen/linksApi

[9] https://zbmath.org/0171.38503

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

[11] The remaining 25% of publications not linked to zbMATH Open refer to documents not indexed in the zbMATH Open database.

Before providing more details about our Scholix Link API, let 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 Open. For example, the book 'Asymptotics and special functions' by F. W. J. Olver (Reprint, 1997; Zbl 0982.41018)[12] is referenced 332 times. Each citation uniquely defines a link to zbMATH Open. An example of one of these links is: https://dlmf.nist.gov/2.10#iv.p2 (see Figure 3). In this case, Olver's book is referenced in Part 2 of § 2.10 (iv) with title 'Taylor and Laurent Coefficients: Darboux's Method.' In Figure 3, we also see that the § 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 3.



These problems can be brought within the scope of §2.4 by means of Cauchy's integral formula

2.10.26
$$f_n = \frac{1}{2\pi i}\int_{\mathscr{C}} \frac{f(z)}{z^{n+1}}\,dz,$$

where $\mathscr{C}$ is a simple closed contour in the annulus that encloses $z = 0$. For examples see Olver (1997b, Chapters 8, 9).

F. W. J. Olver (1997b) *Asymptotics and Special Functions*. A. K. Peters, Wellesley, MA.

Notes: Reprint, with corrections, of original Academic Press edition, 1974
External Links: ISBN 1-56881-069-5, MathReview, ZentralBlatt
Referenced by: §1.10(iv), §1.13(i), §1.13(ii), §1.13(iv), §1.17(i), §1.4(v), §1.4(ii), §1.4(v), §1.8(i), §1.9(vi), §10.12, §10.14, §10.15, §10.16, §10.17(i), §10.17(iii), §10.17(iv), , §18.15(iii), §18.15(v), §18.16(v), §2.1(i), §2.1(ii), §2.1(iii), §2.1(iii), §2.1(v), §2.10(ii), §2.10(iii), §2.10(iv), §2.10(i), §2.10(ii), §2.10(iii), §2.10(iv), §2.10(iv), §2.11(i), §2.11(i), §2.11(vi), §2.11(vi), §2.2, §2.2, §2.3(i), §2.3(ii), §2.3(iii), §2.3(iii), §2.3(v), §2.4(i), §2.4(i), §2.4(ii),

Figure 3. 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).

The underlying dataset of the API has been generated by scraping the DLMF bibliography. For this purpose we developed an auxiliary Python open package.[13] This package is supposed to work for any zbMATH Open partner hosted in the Scholix Link API and has two main functionalities:

[12] https://zbmath.org/0982.41018
[13] https://github.com/zbMATHOpen/Update_Links

(1) Initialise the database of the API with data of a given partner. For those partners for which datasets need to be created from scratch, we included the corresponding scraping scripts;

(2) Update the initial database, thus add new links, delete links that no longer exist and edit links that have been changed.

In the case of DLMF, our auxiliary package creates a dataset containing about 2,000 references (indexed at zbMATH Open) and almost 7,000 distinct links. In this framework, the links are objects belonging to the *source* (of a given partner; DLMF in the present case), and records of zbMATH Open are objects belonging to the *target*.

2.2 Endpoints and response body

The current version of the API offers twelve endpoints:

- `GET /link`. It retrieves links for given zbMATH Open objects.
- `DELETE /link/item`. It deletes a link from the database.
- `POST /link/item`. It creates a new link related to a zbMATH Open object.
- `GET /link/item`. It checks existing relations between a given link and a given zbMATH Open object.
- `PATCH /link/item`. It edits an existing link.
- `GET /link/item/{doc_id}`. It retrieves links for a given zbMATH Open object.
- `GET /partner`. It retrieves data of a given zbMATH Open partner.
- `PUT /partner`. It edits data of a given zbMATH Open partner.
- `POST /partner`. It creates a new partner related to zbMATH Open.
- `GET /source`. It produces a list of all links of a given zbMATH Open partner.
- `GET /statistics/msc`. It shows the occurrence of primary MSC codes[14] (2-digit level) of zbMATH Open objects in the set of links of a given partner.
- `GET /statistics/year`. It shows the occurrence of years of publication of zbMATH Open objects in the set of links of a given partner.

Our JSON response body is modeled on the Scholix metadata schema.[15] This also explains the reason of the name ‘Scholix Link API’ for this service. 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

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

[15] https://github.com/scholix/schema/releases/tag/3.0

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.

2.3 Analysis of DLMF data

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

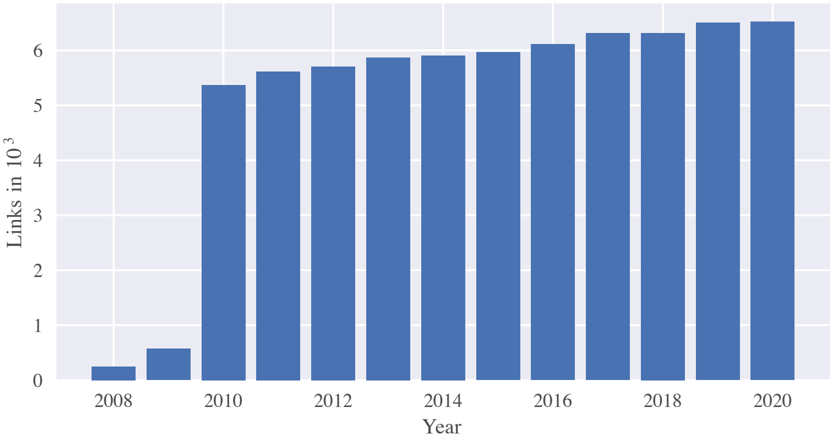


Figure 4. Growth of the links between DLMF and zbMATH Open.

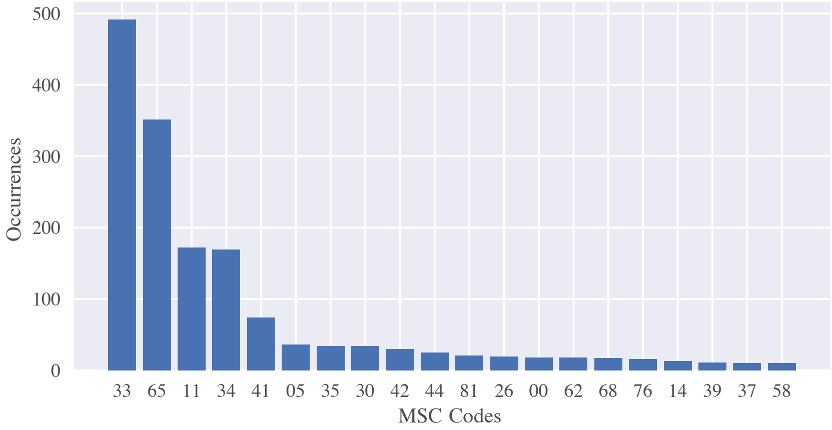


Figure 5. Distribution of primary 2-digit MSC codes in the DLMF dataset.

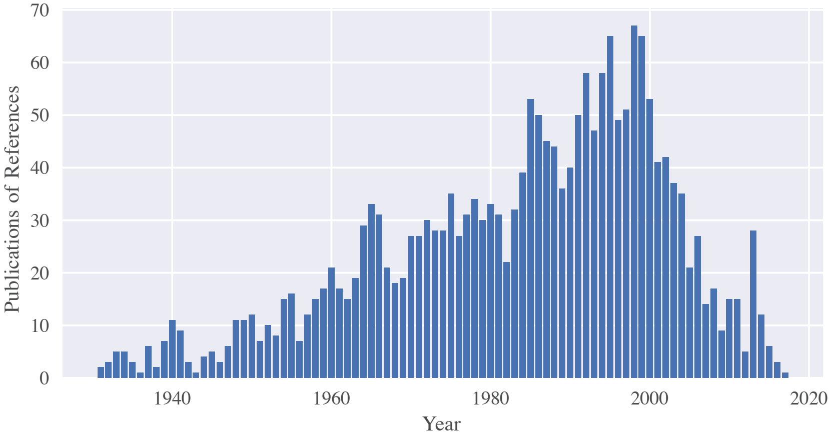


Figure 6. Distribution of years of publication of references in the DLMF dataset.

- In the JSON response body of our `GET` 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 and found the growth numbers depicted in Figure 4. 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 (see Figure 5 for more details):

  - 33 (Special functions), with 491 documents;
  - 65 (Numerical analysis), with 351 documents;
  - 11 (Number theory), with 172 documents.

  A byproduct of this simple analysis confirms the consistency of our MSC tagging system over time.

  On the other hand, the most frequent years of publication of the cited references in the dataset are (see Figure 6 for more details):

  - 1998, with 67 documents;
  - 1999, with 65 documents;
  - 1995, with 65 documents.

Looking at Figure 6 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)[16]: 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)[17]: 118 citations;
  - A. Erdélyi et al., Higher transcendental functions. Vol. I. New York: McGraw-Hill Book Co. (1953; Zbl 0051.30303)[18]: 110 citations.

  In Figure 7 one can see the references, identified by Zbl code, with more than 50 citations.

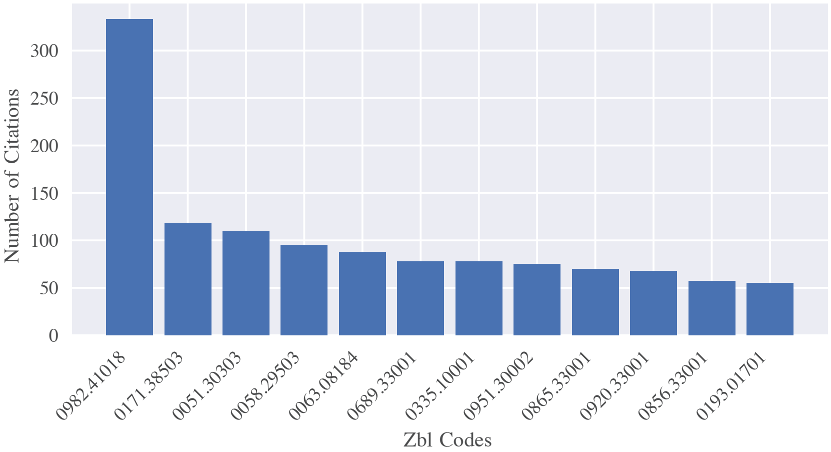


Figure 7. References (identified by Zbl code) in the DLMF dataset cited more than 50 times.

2.4 Usage

The Scholix Link API with its first partner DLMF represents a tool that can be used in various ways and contains many features that help the research process. Here, we

[16] https://zbmath.org/0982.41018
[17] https://zbmath.org/0171.38503
[18] https://zbmath.org/0051.30303

present concrete usage instances where a user of either DLMF or zbMATH Open can benefit from the service:

- A DLMF user can access all bibliographic resources indexed at zbMATH Open 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 Open 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 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 a complete and flexible bibliographic searching tool.

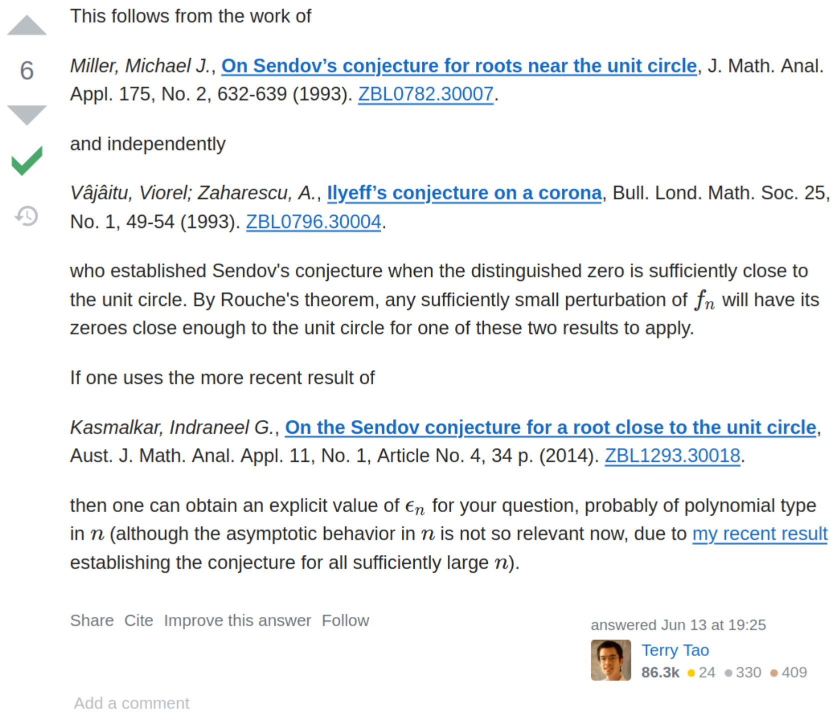
3 MathOverflow endpoint

Over four years ago, a fruitful collaboration between MathOverflow and zbMATH has led to the establishment of a new button labelled "Insert Citation" on the MathOverflow website[19]. The button appears when adding any question or answer and enables users to insert a properly formatted citation to a research article or book. The user can enter a few words of the title or names of some authors and will be presented with a short list of matching papers. If they click on one, a citation to the document will be inserted into the MathOverflow post. This citation includes a link to the respective zbMATH Open entry. The user-facing side of this process is described in [2], here we will focus more on the technical implementation.

To make this suggestion process possible, the well-known MathOverflow user Scott Morrison[20] added some client-side code which calls the MathOverflow API on

[19] https://mathoverflow.net

[20] https://mathoverflow.net/users/3



This follows from the work of

Miller, Michael J., **On Sendov’s conjecture for roots near the unit circle**, J. Math. Anal. Appl. 175, No. 2, 632-639 (1993). ZBL0782.30007.

and independently

Vâjâitu, Viorel; Zaharescu, A., **Ilyeff’s conjecture on a corona**, Bull. Lond. Math. Soc. 25, No. 1, 49-54 (1993). ZBL0796.30004.

who established Sendov's conjecture when the distinguished zero is sufficiently close to the unit circle. By Rouche's theorem, any sufficiently small perturbation of f_n will have its zeroes close enough to the unit circle for one of these two results to apply.

If one uses the more recent result of

Kasmalkar, Indraneel G., **On the Sendov conjecture for a root close to the unit circle**, Aust. J. Math. Anal. Appl. 11, No. 1, Article No. 4, 34 p. (2014). ZBL1293.30018.

then one can obtain an explicit value of ϵ_n for your question, probably of polynomial type in n (although the asymptotic behavior in n is not so relevant now, due to my recent result establishing the conjecture for all sufficiently large n).

answered Jun 13 at 19:25
Terry Tao
86.3k ● 24 ● 330 ● 409

Figure 8. The MathOverflow user Terry Tao citing some articles using the “Insert Citation” feature (post available at https://mathoverflow.net/a/395248).

the side of zbMATH Open and presents the results to the user in a readable format. The API is actually a part of the citation matcher described in more detail in Section 4. The text entered by the user is matched against an Elasticsearch[21] index, which contains all data from the following fields:

- document title;
- original title (in the case of non-English literature);
- author names;
- journal source;
- pagination;
- year of publication.

[21] https://elastic.co

Matching is then done using a standard TF/IDF algorithm:[22] A document is scored higher the more often a given term appears in it (TF, or term frequency). However, the more documents a given term appears in (IDF, or inverse document frequency), the lower its impact in boosting the score is. Thus less weight is given to common terms that appear in a lot of documents. The documents are ranked by the resulting score in descending order, and the top three are returned. The data is exchanged between the browser and the backend using a previously agreed JSON format. More details are presented in Section 4 below.

4 Citation matching

4.1 History

For services within the mathematical infrastructure community, it is often beneficial to be able to interconnect resources by adding links to external services. This includes links to article fulltexts (via DOIs or directly at open-access locations), or to arXiv preprints, but also to reviews at zbMATH Open. The ability to find such links easily is beneficial to publishers, providers of repositories and other services, and thus ultimately to mathematicians using such services. Six years ago zbMATH has therefore started to offer an automated link-finding service called the “zbMATH Citation Matcher”[23] (affectionately labelled “zbMATcH”). It consists of an HTML interface designed for manual use, as well as an API meant for automated access via script. A detailed documentation for the latter is available upon request.

4.2 Algorithm

Like the MathOverflow search described in Section 3, the Citation Matcher works by searching for the terms supplied by the user inside an Elasticsearch index. Here, however, the search is done in a more structured fashion, with dedicated fields for title, author, etc.

Both the HTML interface and the machine API accept input as an unstructured citation text as well as input split up into the respective fields. Thus one can search directly for a citation string like, e.g., “X. Chen, Rational curves on K3 surfaces, J. Algebraic Geom. 8 (1999), 245–278”, or manually enter each relevant part of the citation into the respective input field. In the latter case, matching is done directly, while in the former, the citation string has to be split up and tagged correctly first.

For splitting and tagging the citation string, we use the open source machine learning software GROBID.[24] It takes as input an unstructured string and returns

[22] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[23] https://zbmath.org/citationmatching/

[24] https://github.com/kermitt2/grobid

an XML-encoded response that, according to its best guess, tags which part of the string is an author name, title, publication year, etc. For many commonly used citation formats this works quite well, though in more exotic cases it can give wrong answers. Moreover, the models that the software is shipped with have not been trained specifically on citations as used commonly within mathematical journals, so any formats or citation styles that are specific to the mathematical community can lead to less accurate results. It would be ideal to use a custom model that has been trained on citation data specific to mathematics. However, doing so would need a large corpus of manually tagged citation strings from mathematical publications, hence this approach has not been implemented yet due to lack of resources.

Once the unstructured data has been split up and tagged, a search request can be made to the Elasticsearch index. The index used is filled with the data from zbMATH Open and contains the most important fields for searching citation data. As in the MathOverflow case, a TD/IDF score is computed for each document, and results are ranked by this score in descending order. The topmost results are returned if their score exceeds a certain threshold (set to 5.0 by default, but adjustable in the API).

How high this threshold score should be is a non-trivial question, especially when using the results of the API to add links to zbMATH Open automatically, i.e., without human supervision: The lower the threshold score, the more frequently a result is returned, but the proportion of false matches will also go up. If on the other hand one sets it too high, the retrieved results will be more likely to be correct, however it will also be more often the case that a citation is not matched even though it is contained in zbMATH Open. In automated applications, a somewhat conservative threshold score of 8.0 is recommended to keep the number of errors of the first kind to an acceptably low level.

4.3 Future developments

Even though the zbMATcH algorithm has worked well for several years, it does have its share of problems and things left to be desired.

- Chief of these is that the TF/IDF score computed by Elasticsearch is only designed to rank the results of a single query amongst each other, not to compare results of different queries. Of course, this is exactly what we are doing if we postulate a global threshold score and only accept results when their score exceeds this threshold.
- Secondly, using the current algorithm it is not easy to add new features or criteria to the matching, or to measure the effect such features have if they are introduced.
- And finally, the current implementation is not able to take into account additional information besides the content of the citation string. For example, most citation

strings do not materialise out of thin air, but are instead contained in the references section of some publication. Now it is quite unlikely that the cited article has a publication year that lies in the future of the citing document (it can happen occasionally, for example due to preprints, reprints, or delays in publication, but in general it is quite safe to assume that cited articles have been published prior to the ones citing them). Hence, if one has the additional information of the publication year of the citing document, one can take that information into account when scoring candidate results of the matching process. Other ways of incorporation additional metainformation are also possible, e.g., coauthor networks or MSC classification (articles are more likely to cite within their own field, and we have data on which MSC classes are frequently cited from which others).

We are therefore implementing a new algorithm that is designed to overcome these weaknesses and provide better matching results while being at the same time more modular and hence easier to modify and evaluate. Figure 9 gives an overview of the structure of the new algorithm. As before, GROBID is used to structure text citations, however it is still possible to query using structured data directly.

What follows is a list of modules called *candidate generators*. Their task is to loosely select a set of documents, according to appropriate criteria, which have at least a non-negligible chance of matching the input citation. Their purpose is mainly to ensure efficiency, so that the matching score does not need to be computed for every single document contained in the database, but rather for this preselected set only. Several ways of generating candidates are conceivable, for example an Elasticsearch query as in the old algorithm, or a selection just by the numbers occurring in the citation string. The latter is helpful in particular for citation styles that do not contain a title. Moreover, collections of numbers, like volume, issue, page numbers, and publication year, tend to exhibit a high degree of uniqueness and hence of specificity. Using more than one candidate generator helps to ensure that no relevant target document is accidentally left out of the matching process.

Next comes the core of the new algorithm: The so-called *featurizers*, which for each candidate compute a numerical feature that encodes a certain degree of similarity to the input citation. There can be as many of these as needed. Each focuses on one specific property of the input/candidate pair. They can, for example, compare certain structured fields using appropriate methods (for example, allowing for $\LaTeX$ encoding in titles, or author names using initials only), or work with the raw citation text (e.g., by comparing substrings). If supplied, featurizers can also make use of metadata of the document containing the citation, for example by comparing publication years as explained above.

The output of this step is a list of numerical feature scores, which can then be fed into almost any kind of machine learning algorithm. Our example uses a support vector machine (SVM), which essentially computes a (weighted) linear combination

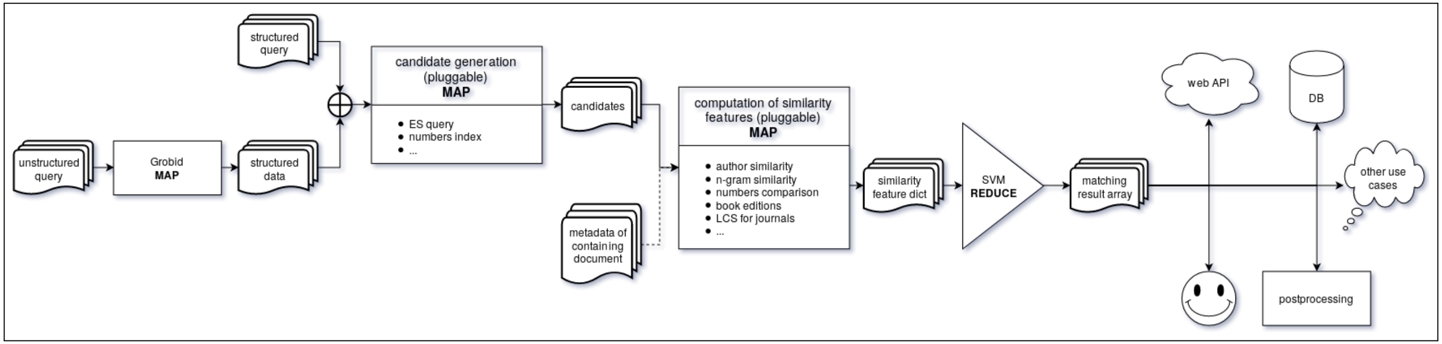


Figure 9. Schematic workflow of the new citation matching algorithm.

of the features. The weights have to be learned by training this algorithm, for which a set of so-called *gold data* is needed, i.e., citations where the correct matching document is known. For this, we use a set of references where a DOI is included (and the document referred to is part of the zbMATH Open corpus). By comparing the weights after the training is finished, it is even possible to determine to what extent each featurizer contributes to the final score.

As before, the final output of the algorithm is a list of scored candidates, ranked by score in descending order. The output fields of the new algorithm is a superset of the ones returned by the old, and likewise for the input. Hence existing tools can use the new one as a drop-in replacement without any changes. However, it is of course hoped that by making use of the new features the quality of the matching is greatly improved.

5 OAI-PMH API

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) protocol is widely used for metadata harvesting. With our OAI-PMH API,[25] one can harvest the entire zbMATH Open dataset or some specific subsets of it. In this section, we present an overview of the implemented endpoints and our custom extensions based on our previous publication [6].

5.1 Endpoints and response body

As required by the protocol, our API offers six endpoints:

- Endpoint 1 (“Identify”) helps aggregators and archives to discover the API fully unsupervised. Further it identifies the version of the OAI-PMH standard used;
- Endpoint 2 (“ListMetadataFormats”) lists the formats that we use to expose the data of zbMATH Open. We implemented two flavours, the standardized Dublin Core[26] metadata format (which is required by the standard) and a second format, that is closer to the internal data model of zbMATH Open. The content generated by zbMATH Open, such as reviews, classifications, software, or author disambiguation data are distributed under the CC-BY-SA 4.0 license.[27] 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

[25] https://oai.zbmath.org

[26] https://dublincore.org

[27] https://creativecommons.org/licenses/by-sa/4.0

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. On the other hand, we expect that for people used to our website, our own format is more appealing;

- Endpoint 3 (“ListSets”) lists the subsets of the zbMATH Open dataset, i.e., one set for each primary MSC label and one set for the articles originating from the ‘Jahrbuch über die Fortschritte der Mathematik.’
- Endpoints 4 (“ListIdentifiers”) and 5 (“ListRecords”) list the current identifiers and records, respectively. This endpoint is intended to provide a dump of all public data contained in zbMATH Open;
- Endpoint 6 (“GetRecord”) gets specific entries of zbMATH Open.

5.2 Extensions to the standard

While the endpoints defined in the OAI-PMH schema are useful for retrieving large fractions of the zbMATH Open dataset, the search capability for specific articles is limited. Therefore, we extended the OAI-PMH standard with custom endpoints without breaking the compatibility with the leading protocol. In particular, we have designed a simple query language that allows filtering based on the following properties: document type, year, document author, classification, keyword, document language, author variation, author reference, biographic reference, software, review type, review language, reviewer, serial publisher. All those fields can be combined with the boolean operators ‘and’, ‘or’ and ‘not’. We chose the operators in a way that they are outside the alphabet for set names. By doing so no extra escaping or confusion between operators and sets is possible.

6 zbMATH Open REST API

As outlined in Section 5.2, we immediately realised that the OAI-PMH standard is not an optimal fit for all use case scenarios and does not optimally match the requirements of the five user groups outlined in Section 1. With the implemented extensions, we can retrieve more specific subsets of our dataset. However, we are still bound to the OAI-PHM metadata format and protocol. For example, the result format must be XML, which is hard to process for less experienced developers. Moreover, the search capabilities are very limited and write operations are not defined by the standard. In addition, defining metadata schema definitions in XML is connected with significant overhead and makes changes to the API more difficult. For example, correcting mistakes is not the purpose of that standard. Therefore, we plan to develop a custom

REST API tailored explicitly to the zbMATH Open dataset, providing different results formats, including JSON and XML. Eventually, we want to take our API development to the next abstraction level, making all information visible on the website machine readable. This would allow future user interfaces to run on top of the API without accessing internal data sources directly. This will facilitate the development of alternative frontends such as clients for mobile devices. By doing so, we follow the example of DataCite[28] and others that provide different APIs to access the bibliographic content. Eventually, different APIs that present the data in different formats for various purposes will contribute to the vision of interoperable research graphs [1].

While the OAI-PMH API is designed for harvesting data, not for updating data (note that also the current zbMATH Open website performs read-only access to the zbMATH Open database), we will in the future allow write operations via APIs. To ensure high data quality, we will require user authentication and double-check all incoming data before processing it further to ensure the reliability of zbMATH Open. This was also a crucial point in the development of the Scholix Link API and will be subject to discussion and interaction with the communities in order to find a good balance between high quality and high volume of data available at zbMATH Open.

7 Concluding remarks

The main purpose of this contribution is to provide a broad and complete scenario of the recent digital innovations in zbMATH Open.

Having made the data freely accessible has obviously offered a wide range of new ideas and resources in order to optimise the usability of our service. Some of these ideas have already been worked out, as discussed in this article, others will come soon.

We see two great challenges for the future: on the one hand to improve and solidify what we have already built in the recent past, on the other to frame our digital services in a universal scheme. The latter is undoubtedly the most difficult and exciting test for us. This scheme must contain in a functional and organic way all the various services discussed in this article in order to make zbMATH Open a solid tool for the community, avoiding the risk of offering disconnected and non-interoperating services.

[28] https://support.datacite.org/docs

References

[1] A. Aryani, M. Fenner, P. Manghi, A. Mannocci, and M. Stocker, Open science graphs must interoperate! In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Lyon, France*, edited by L. Bellatreche, M. Bieliková, O. Boussaïd, B. Catania, J. Darmont, E. Demidova, F. Duchateau, M. Hall, T. Merčun, B. Novikov, C. Papatheodorou, T. Risse, O. Romero, L. Sautot, G. Talens, R. Wrembel, and M. Žumer, pp. 195–206, Communications in Computer and Information Science 1260, Springer Cham, 2020

[2] I. Beckenbach, zbMATH Open and community platforms. *This volume*, pp. 49–53, EMS Press, 2024

[3] K. Hulek and O. Teschke, The transition of zbMATH towards an open information platform for mathematics. *Eur. Math. Soc. Newsl.* **116** (2020), 44–47

[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** (2019), 50–52

[5] M. Petrera, D. Trautwein, I. Beckenbach, D. Ehsani, F. Müller, O. Teschke, B. Gipp, and M. Schubotz, zbMATH Open: API solutions and research challenges. In *DISCO 2021: Digital Infrastructures for Scholarly Content Objects 2021*, edited by W.-T. Balke, A. de Waard, Y. Fu, B. Hua, J. Schneider, N. Song, X. Wang, pp. 4–13, CEUR Workshop Proceedings 2976, CEUR-WS.org, 2021

[6] M. Schubotz and O. Teschke, zbMATH Open: Towards standardized machine interfaces to expose bibliographic metadata. *Eur. Math. Soc. Newsl.* **119** (2021), 50–53

Chapter 11

zbMATH Open as a tool for bibliographical studies

Klaus Hulek, Olaf Teschke

1 Introduction

Evaluations and rankings, be it of individuals or institutions, have become part of academic reality. These evaluations range from career-defining assessments of individuals to worldwide university rankings. Although the methodology of many of these evaluations has often been criticised, they remain ubiquitous with often extraordinary effects. The impact on individual careers, and hence lives, can be decisive. On a more global level these figures not only contribute significantly to the reputation of universities, but also affect the choices of perspective students and staff.

Various parameters are used to evaluate research performance, with bibliometric data playing an important role in (almost) all evaluations. Both, generating these data, as well as interpreting them, constitutes a major challenge. Therefore, it is important to understand the technical aspects, as well as the different parameters and perspectives, that go into bibliometric data.

The aim of this contribution is to show how zbMATH Open data provide insight into how mathematics is being published, with an emphasis to reveal how the genuine specifics of the mathematical literature render traditional subject-blind bibliometric approaches and measures inapplicable. Since most of zbMATH Open data – especially those relevant for bibliometric analysis – are openly available by a CC-BY-SA license [11] through the zbMATH Open REST API[1] [10], the following observations can not just be easily reproduced, but can serve as the basis for further, more sophisticated analysis.

2 Time line of mathematical references

It is a fundamental characteristic of mathematics that a theorem, once proved, remains valid forever. Nevertheless, scientific progress often leads to stronger and more general results which thus supersede earlier work. Hence the question about the relevance of older results, measured by the average time interval between publication and citation, is highly nontrivial. Other disciplines like biology, chemistry, physics, or medicine have recently seen a faster decline in citations [8] of a given paper, indicating that the half-life of publications might be decreasing.

[1] https://api.zbmath.org

With currently almost 50 million references available for a total of 2 million documents, the zbMATH Open citation database constitutes the largest curated citation database for mathematics.

To investigate the reference time line, it is not necessary to match the references to the database, which is only possibly for about 60% — the remaining 40% maybe unpublished work, or outside the scope of zbMATH Open. For this it is sufficient to extract the publication year from the string. This is the basis of the diagram in Figure 1, which shows the development of citation distances over time. It shows the average difference between the publication years of cited works and the publication year of the citing work, depending on the latter. An average was taken across all subject areas and all forms of publication.

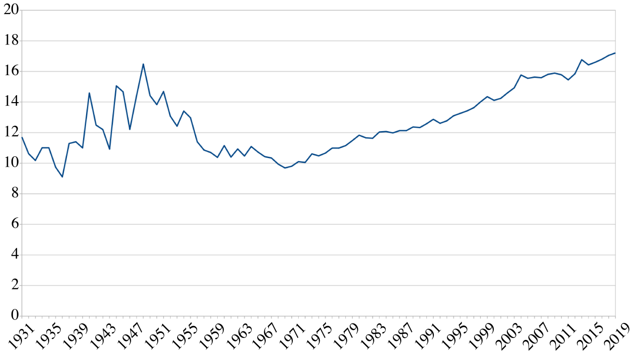


Figure 1. Average time interval between publication and references.

The striking result is that the average age of cited papers has actually grown constantly, and is now almost 18 years. To interpret the graph, various aspects must be taken into account. Reference data are currently available for only around 4% of publications before 1945. Therefore, this part is subject to increased uncertainty and shows a correspondingly erratic course. However, the effects of the two world wars are visible in the graphic, which led to a decrease in the interval between publication and citation in the following decade.

A plausible explanation for this is that publications from the war years, which would normally be cited more widely after a few years, are missing here and, on the other hand, reference is mainly made to more recent literature, especially since many working groups and networks had to be rebuilt. After WWII, this effect dominated until 1968. Since then, the diagram shows a continuous growth of this interval. This period of about 18 years after the end of WWII, before the citation distance starts to grow again, is a further indication that two decades represent a natural lower limit for average citation distances (excluding war effects), at least for the period in which extensive data are available.

3 Is there a half-life of mathematical results?

Vice versa, one can ask also how long a given work is cited. By investigating this with zbMATH Open data, we must keep in mind that, in contrast to the previous section, scope effects come into play – only the indexed citing documents contribute to the data.

The general concept of research impact suggests that research which is cited for a long time after its publication typically represents more significant contributions. However, this approach cannot be used to identify all outstanding publications. Important theorems often become so much part of common knowledge that a reference is no longer given. In other cases, more accessible version or survey articles are cited instead of the original work.

Instead, we consider mainly the question of longevity of references, i.e., the temporal distribution of references for documents published in a fixed year. One would expect that in general the number of citations increases sharply immediately after the publication year, but would show a steady decline. However, it turns out that the growth of the published literature has the strongest influence here, leading effectively to an unlimited growth of references to a set of documents with fixed publication year. Hence, in the diagram shown in Figure 2, the number of references to a given publication year is normalised by the overall number of references for the publication year of the citing documents. Moreover, for a better impression of the structure, the figures are also normalised with respect to the maximum of this figure.

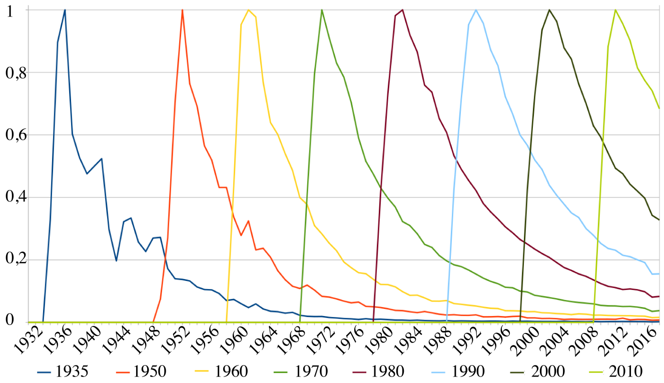


Figure 2. Timeline of relative citations to mathematical papers for different fixed publication years.

The x-axis shows publication years; the y-axis the number of citing articles relative to the overall references in the citing year and the year of maximum citations.

With these normalizations, the figures match more closely the expected shape. However, the relative maximum is usually obtained only after three years (or, taking the results of Section 5 into account, in average five years after it has been posted on the arXiv for core mathematics papers). Moreover, the decline remains relatively smooth, with more than half of the relative citations being generated more than eight years after publication. It should also be noted that the aggregations provide no information on the share of highly cited publications, which, based on the analysis of some samples, appears to grow over time (especially for books). Such an analysis would be beyond the scope of this note, but readers are invited to explore this effect by using zbMATH Open data.

Taking all this into account, it becomes clear that the usually applied short-term bibliometric measures (such as three- or five-year impact factors) miss the crucial part of the relevant citation information. Vice versa, assuming the usual timeframes in scientific careers, there seems no meaningful way to include into decision-making measures which only have a chance to become relevant about a decade after their underlying idea went public.

Another caveat would be that this diagram aggregates publications from all mathematical areas. However, both citation behaviour and publication growth depends heavily on the subject, so it seems natural to take subject specifics into account.

4 One step further: Subject specifics

One might wonder whether it is possible to differentiate this general picture further by taking mathematical subjects into account. Matching citations to zbMATH Open provides MSC information and raises the natural question what the interdependence between mathematical fields and citation networks is.

Figure 3 shows that there is indeed a strong concentration along the diagonal (which means that the bulk of references point to papers with the same MSC), although there obviously exist further cross-references which might be worth investigating in a more detailed analysis.

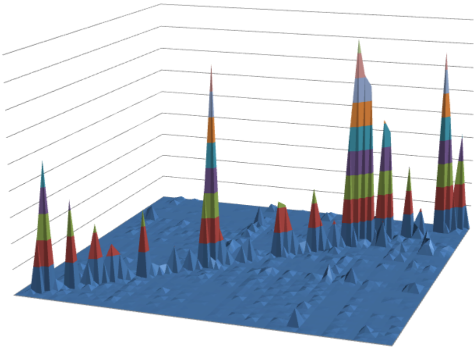


Figure 3. Cross-MSC citation map.

The strong concentration on the diagonal (which is, by the way, an indication that the MSC actually depicts clusters of related work well) can serve as a justification that restricting to area-preserving citations serves well as a first approximation.

We employ here the following distribution into mathematical subdomains, as employed in [15, 17]:

- Gen: General Mathematics; History; Foundations. This corresponds to sections 00, 01, 03, 06, 08, and 18 of the Mathematics Subject Classification MSC
- Disc: Discrete Mathematics. Convex Geometry; MSC sections 05, 52
- NTAG: Number Theory. Algebra. Algebraic Geometry. Group theory; MSC sections 11, 12, 13, 14, 15, 16, 17, 19, 20
- Ana: Real and Complex Analysis; MSC sections 26, 28, 30, 31, 32, 33, 40, 41.
- OpTh: Harmonic and Functional Analysis; Operator Theory; MSC sections 42, 43, 44, 46, 47.
- DIEq: Differential and Integral equations; MSC sections 34, 35, 37, 39, 45.
- OptCS: Optimization. Numerical Analysis. Computer Science. Algorithms; MSC sections 49, 65, 68, 90, 93, 94.
- ProbStat: Probability Theory and Statistics. Applications to Economics, Biology and Medicine; MSC sections 60, 62, 91, 92.
- TopGeom: Topology and Geometry; MSC sections 22, 51, 53, 54, 57, 58.
- MaPh: Mathematical Physics; MSC sections 70, 74, 76, 78, 80, 81, 82, 83, 85, 86.

The aggregation over all publication years aims to eliminate the growth effects mentioned earlier. Figure 4 shows the relative distribution of references for these ten MSC clusters in relation to the interval between publication and citation (from 0 to 24 years). It is evident that a long-term decay for relative citation frequencies of subject-preserving citations exists, but there is also a significant long tail. A notable

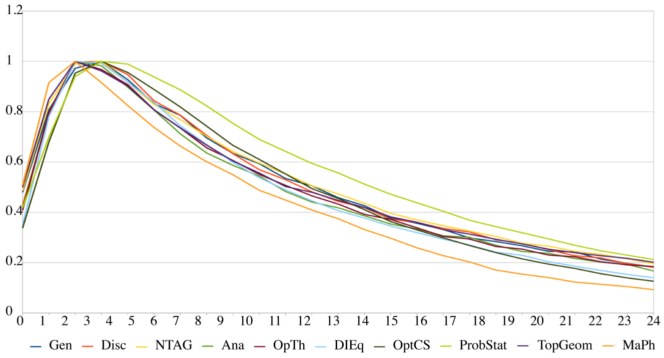


Figure 4. Relative time intervals for subject-preserving citations.

exception is mathematical physics, where the initial relative citation rate is much higher before descending much more quickly. For the remaining areas, the diagram confirms that citation metrics that only cover a short time interval can hardly have any significance for mathematics. With the observed distribution, it becomes obvious that any measure that will not omit the most relevant information must cover a span of at least five years.[2,3]

5 Effects of publication delay

Yet another prevalent effect which provides a strong argument against the use of short-term bibliometric measures in mathematics is the exceptionally long publication delay due to the rigorous, and hence often extensive, peer-review process. zbMATH Open data can be used surprisingly easily to determine its magnitude. This is done in the following way. For many years, the arXiv has established itself as the standard preprint repository for many areas in mathematics, often preceding the actual publication by several years. Since 2016, zbMATH matches mathematical publications to their arXiv versions. As shown in [16], the arXiv is rarely used for retrospective self-archiving, hence the difference between arXiv submission and publication date can serve as a proxy for publication delay.

The diagram in Figure 5 shows the distribution of articles with respect to publication year for various arXiv submission years. As it can be seen from zbMATH Open data, the average publication delay accounts for about 18 months, but may vary significantly depending on the journal, subject, or individual paper. The effect of the subject could again be explored further by an MSC-based analysis.

[2] In fact, the temporal development in Figure 4 does not seem to be consistent with the results in the previous section and Figure 2. However, the decline is due to two effects: On the one hand, the citations are summarised across all years, so that the effect of publication growth is leveled out. On the other hand, citations with a large time interval are more often cross-area and therefore not included in Figure 4.

[3] Another methodological artifact should be noted that could also influence the results of other statistical studies (such as [8]): studies are often limited to the so-called top 10 % papers (this refers to papers with high short-term citation numbers, whether justified or not). With such a selection, some areas would be over-represented, even based on the zbMATH Open data, and would suggest a faster relative decline in citations than justified.

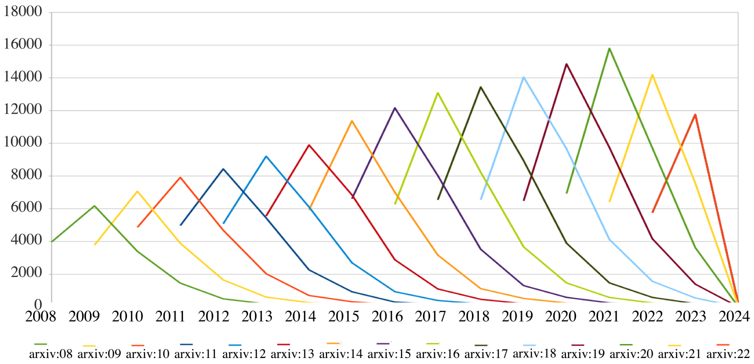


Figure 5. Publication timeline for several arXiv submission years.

With the same categories as in the previous section, the diagram in Figure 6 shows the distribution of the number of arXiv submission with respect to the average difference to the submission and publication year.

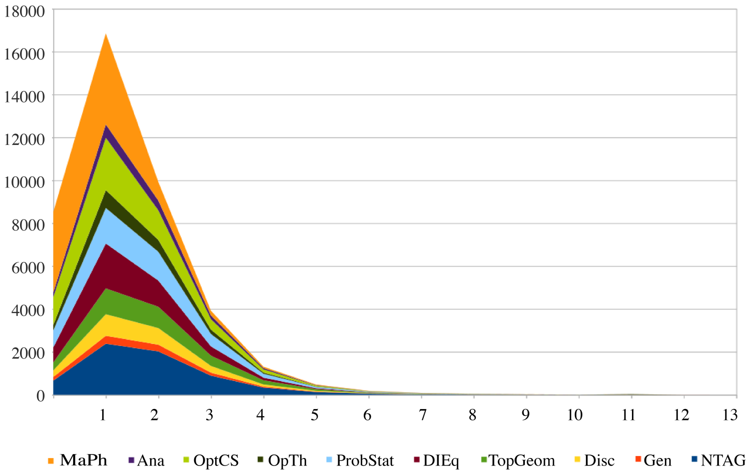


Figure 6. Publication delay based on arXiv submission dates for several mathematical areas.

One notices, e.g., that for mathematical physics (MaPh) the difference is much smaller than for core mathematics areas – e.g., for NTAG the average difference exceeds two years.

This adds further evidence that short-term bibliometric measures are inadequate for mathematics – indeed, the widely varying publication delay is a strong argument in itself that the two-year impact factor, which is often used in bibliometrics, is highly unreliable for mathematics journals [9].

6 Aggregated journal information

Citation data is often used in aggregated form, in particular summarised for journals, individuals or institutions. In this section we discuss the case of journals in more detail. Mathematical journals are characterised to varying degrees by the areas represented. In simple terms, one can differentiate between those for special topics from cross-field to general mathematics journals, although the definition of a general journal is not trivial and even in such cases the regional representation can vary widely [15].

In addition, the focus changes over time, there are changes in editors, and sometimes journals are renamed or produce spin-offs. The variance of citation measures is even greater between specialist journals to which the subject specifics considered in the previous section can be directly transferred.

The diagram in Figure 7 shows the total publication and citation numbers for four classes of journals. This is based on the zbMATH Open internal categorisation of journals. This classification is done less with the aim of a ranking than with a quick decision on priorities in the workflow and, ideally, a fair balance of specialist areas. It therefore differs in detail from other approaches (such as the Scandinavian or Australian ranking), but of course all highly relevant general mathematical journals (Acta Mathematica, Annals, Duke, Inventiones, JAMS, JEMS, Publ. IHES, ...) are represented in the 164 journals in the FAST TRACK category, as are the leading journals in the respective specialist areas. The other three categories distinguish further workflow priorities, with category 3 journals containing usually only a small fraction of research mathematics.

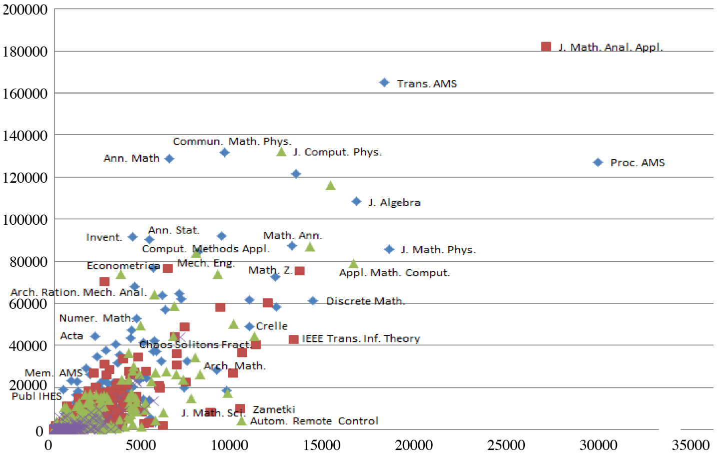


Figure 7. Publication- (x-axis) und citation (y-axis) figures of mathematical journals from four zbMATH Open categories: FAST TRACK (diamond), 1 (square), 2 (triangle) und 3 (cross).

In this diagram, the total number of all publications in the journal (x-axis) is related to the total number of all citations in this journal (y-axis). Accordingly, the slope of the origin line determined by the entry of a journal can be seen as a proxy for the average *impact factor*.

It is obvious that the spread is very wide within all categories. Indeed, the slopes vary very much with the mathematical specialties; in fact, they are strongly influenced by them. Although the average gradient in the FAST TRACK category is above category 1, it is apparently not significant, given the individual deviations. What is also striking is the often high increase in the next category 2, which is due to the fact that here a particularly large number of journals from mathematical physics or engineering are represented, so that the citation patterns of these areas dominate. In addition, there is an increased presence of journals in this category from countries (such as Iran), in which the evaluation of scientists is often very strictly linked to bibliometric values and thus a correspondingly adapted publication behaviour is enforced.

A possible conclusion is that aggregated citation information is primarily shaped by factors such as the area profile or the scientific environment – it is only after taking these dominant parameters into account that a noticeable correlation of a numerical citation indicator with the assessment made by experts can be observed. In order to analyse this in more detail, it is just as necessary to have this granular profile information available as well as to be aware of the influences of time delay and data availability and accuracy mentioned above.

7 Aggregated author information

While the previous analysis was mostly document-based, it is also worthwhile taking a more author-centred point of view when analysing publication behaviour. Such an analysis, however, requires extremely precise authorship data, since otherwise error propagation would disturb any derived quantities, making meaningful conclusions impossible. In this section, we take advantage of the significant progress of the zbMATH Open author disambiguation during the past years. Methods and progress on this matter have been amply described in [14, 18]. Nevertheless, we would like to mention that currently only roughly 3.5% of authorships are ambiguous (compared to 5% in 2018), despite the growing ratio of authorships involving Chinese names, which cause the most complicated disambiguation tasks. Most large clusters of Chinese names have now been successfully analysed (e.g, more than 1,500 documents involving the most frequent single name Wang, Wei have been distributed to currently 344 identities). The by now highly efficient author disambiguation will help to eliminate distortions in the subsequent analysis (which will take into account only the 96.5% of unambigious assignments).

We will first employ the zbMATH Open author database to derive figures on the number of actively publishing mathematicians in a given year. Some effects showing changing publication frequency and collaboration behaviour will become visible. With the assignment of MSC (Mathematical Subject Classification) classes since the 1970s, it is possible to analyse and compare these figures for different mathematical areas. For convenience (and to achieve some historical coherence by avoiding effects from the evolution of MSC) this is done for the set of ten clusters of main MSC classes which we already introduced above.

When one focuses on author counts, instead of publication numbers, one has to keep in mind that the distribution of papers is extremely biased. The median author has 2 publications, while the average publication number is about 7.9, with the maximal number of publication for a single author being 1769 (further data can easily derived from the zbMATh Open API).

There are many reasons why many authors are only connected with one paper. The obvious one is a short career in academia, often just a PhD thesis and one paper derived from this. Other people may have longer careers in research, but may switch to application areas where they drop out of the scope of zbMATH Open. In any case, this large percentage is the main reason for a large coincidence of the author and document count, as shown in Figure 8.

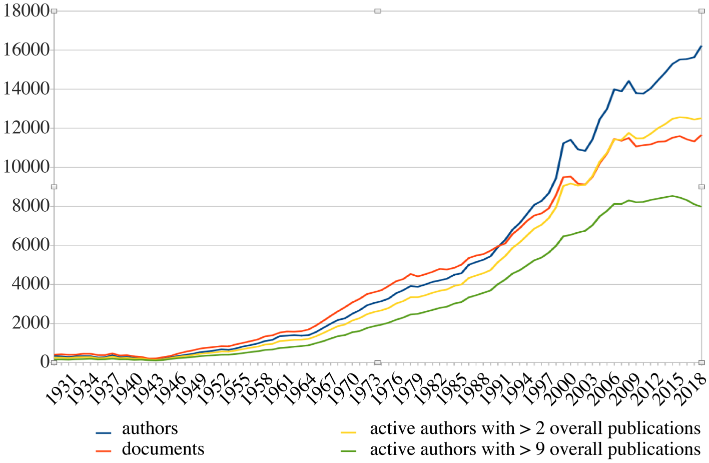


Figure 8. Actively publishing authors per calendar year, in relation to documents.

In spite of the possible methodological issues discussed above, two trends are clearly visible: (1) the number of active authors grows much quicker than that of the overall publications, and (2) the figure of established researchers with a larger number of papers grows much slower. Two main effects can conceivably play a role here

– the publication frequency and the collaborative behaviour. Due to the large number of authors with very few papers, a detailed analysis of the publication frequency is highly complicated, especially since it then seems appropriate to also involve an analysis of the length of the publications in such a study.

The overall length of publications has actually been decreasing. But this phenomenon is due to the shrinking role of books. Papers in journals have in fact become longer, at least in some areas [6]. Further effects here come from the replacement of printed by fully electronic versions and different journal policies. Again, this makes a more detailed analysis, which would also need to involve the journal status, as well as the area, quite demanding and is thus beyond the scope of this contribution. In other sciences a tendency to split results into least publishable units has been reported. At this stage our data do not allow us to draw substantiated conclusions on this for mathematics.

We will, however, see that the changing collaboration behaviour is likely to be a major factor in the increased growth of authors.

7.1 Collaboration behaviour and subject-based figures

Historically, mathematical publications were predominantly single-authored. Recently, this has changed significantly, following similar trends in other sciences. Though the overall effect is strongly driven by application areas, the phenomena are visible throughout mathematics. We employ the same categories as for the analysis of publication delay and obtain in Figure 9 a diagram of average authorships per publication for the calendar years.

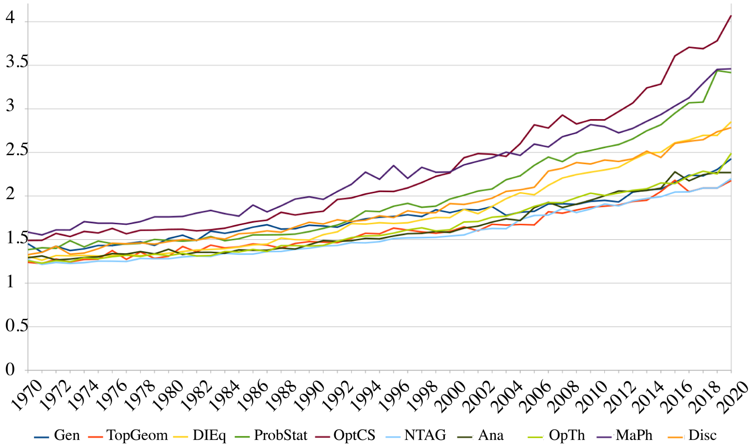


Figure 9. Average number of authors for a paper in clusters of ten mathematical areas.

There are significant differences between different clusters. Examples are given by OptCS (where the average now exceeds 4), MaPh, or Probstat (almost 3.5) and TopGeom or NTAG (about 2.2). In spite of this, however, the overall tendency is clear – collaboration has significantly increased in all fields. With mathematics being a very international enterprise, this seems to hold true globally, although samples indicate that figures may differ geographically, which may be explained both by area correlation or national science policies. However, such an analysis would again exceed the space of this article, and will be left to subsequent studies (again, the reader is encouraged to employ data available from the zbMATH Open API for a more sophisticated analysis).

Analogously, a breakdown can be made of the actively publishing mathematicians in each field; see Figure 10.

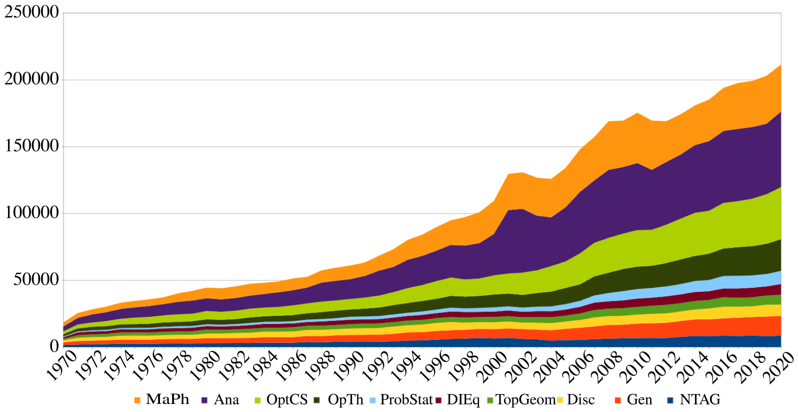


Figure 10. Actively publishing persons in ten clusters of math subjects.

There is a small caveat here – actively publishing mathematicians are evaluated separately for each area, so in the cumulative display, people active in several clusters may appear several times (the comparison with Figure 8 shows that this effect amounts to an about 20% increased height).

Summarizing, we can say that the publication behaviour has clearly changed throughout mathematics towards a more collaborative attitude, but the intensity with which this happens is somewhat different in different areas.

7.2 Citation and coauthor networks

Another aspect, which is relevant in connection with the observed increased collaboration, is the question as to how citations are distributed within the coauthor network. Although it is for many reasons clear that mathematical achievements cannot be compared on the basis of simple (especially, short-term) citation counts (cf. [1,2,4]), there is still a prevailing notion that some (possibly vaguely defined) impact is correlated with aggregated citations. For a better understanding of what citations reflect, we would here suggest a first step into an empirical analysis of their distribution in the collaboration network. Although there have been suggestions of a bibliometric index involving collaboration distances [3], it appears that such approaches have never been applied to real-world databases. One reason might be that such an analysis requires very precise authorship data, since otherwise the error propagation would lead to ever more unreliable results as the coauthor distance grows. In bibliometrics, the discussion is mostly restricted to the zero level (i.e., a possible exclusion of self-citations). This is unlikely to provide a comprehensive understanding.

The mathematics collaboration graph has been investigated frequently, especially in [13], based on zbMATH data. While the median distance in its large connected component is 5, the situation is different when one looks at the collaboration distance for citing authors.

Here one would naturally expect shorter collaboration distances. Since higher collaboration distances are linked to a higher error probability, we restrict our discussion to the ranges from 0 (self-citations), 1 (coauthor citations), 2, 3 and more than 3. The distribution shown in the diagram in Figure 11 indicates that these seem indeed the most significant categories.

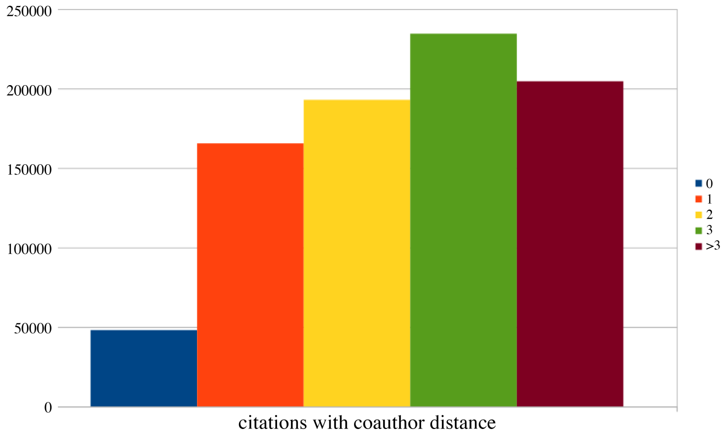


Figure 11. Minimal collaboration distance for citations of zbMATH Open authorships.

More precisely, we computed for each authorship in a paper cited in zbMATH Open the minimal collaboration distance to the citing paper (note that due to multiple authorships, the total number is larger than the overall number of matched references in the database). The figures show that both, the average and the median collaboration distance, is 3. The aggregation for authors, however, seems to indicate that the distribution is somewhat uneven; see Figure 12.

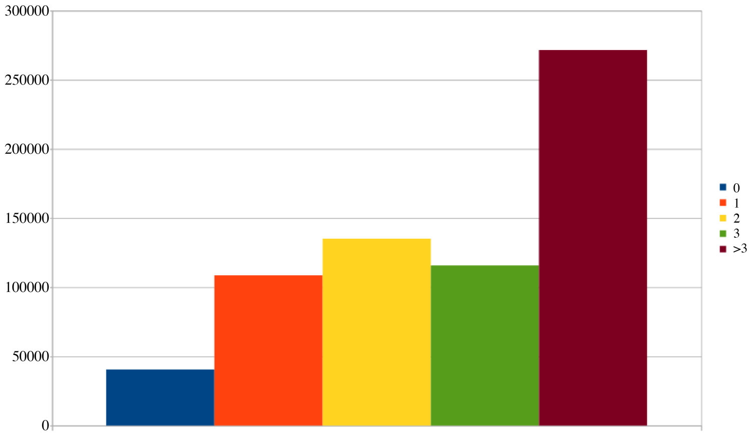


Figure 12. Number of authors in zbMATH Open with median collaboration distance n for their citations.

Of the 671,513 cited authors evaluated, most (271,435) have median collaboration > 3 distance for their citations, with a second maximum at distance 2. When we restrict this analysis to the top 15,000 cited authors in zbMATH Open (which account for more than half of all citations), the picture is, however, different; see Figure 13.

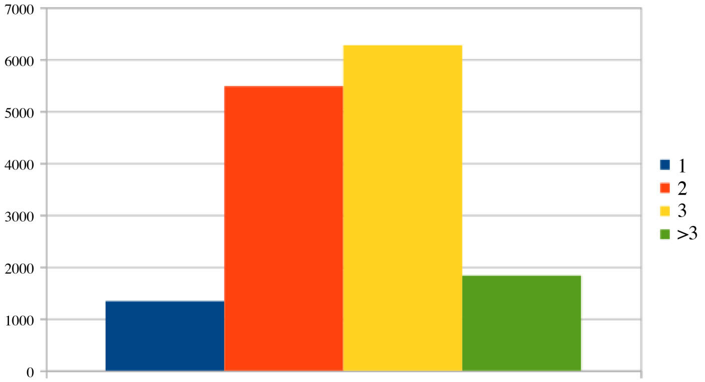


Figure 13. Number of top 15,000 cited authors in zbMATH Open with median collaboration distance n for their citations.

One sees that the distribution in Figure 12 derives from the large number of rarely cited (and thus presumably also rarely collaborating) authors, which therefore necessarily also have larger collaboration distances. For the 100 authors with most citations in zbMATH Open, the picture is even clearer; see Figure 14.

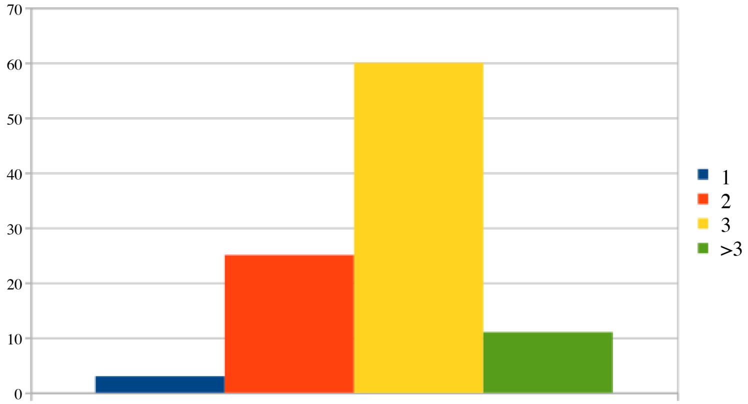


Figure 14. Number of top 100 cited authors in zbMATH Open with median collaboration distance n for their citations.

In the presence of a high number of citations, a median of three for the collaboration distance of citations seems indeed to be the default value, which is very much the standard for today's mathematical community. The larger value of four occurs almost exclusively for older mathematicians with fewer collaborations (e.g., Kolmogorov, Mac Lane, or Pólya), or in bordering areas for which collaboration paths may exist only outside the database (e.g, Barabási or Hawking). On the other hand, Erdős, who is obviously at a disadvantage due to his huge collaboration network, is almost the only elder famous mathematician with median 2; else, median 2 occurs mostly for younger mathematicians where the citations are more likely to derive from a narrower community. Especially, the rare cases of median 1 (i.e., most citations are self-citations or come from immediate coauthors) indicate almost invariably a very particular citation network.

Finally, we compare the collaboration distance (CD) distribution of zbMATH Open citations for the Fields Medalists (FM) and the highly cited researchers (HCR) in mathematics 2022 [5] of the Clarivate database:

CD	0	1	2	3	> 3
FM	7,129	37,576	117,667	193,372	130,562
HCR	29,893	139,980	164,290	175,220	81,515

The huge difference between the distribution in both series is obvious. Although the Clarivate HCR gather a much larger total citation number, only a relative small fraction affects collaboration distances ≥ 2, which usually accounts for most of the citations. By far most of HCR citations derive from the close coauthor network, and the median of two differs significantly from the corresponding figure of the most cited authors in zbMATH Open. Even as much as 10% of Clarivate HCR turn out to have an extreme collaboration median of two for their zbMATH Open citations, i.e. most of their citations are self- or coauthor citations. The difference of median citation distance for Clarivate HCR in comparison to highest cited zbMATH Open authors may indicate that the Clarivate database contains many more sources that involve large numbers of self- and coauthor citations. This adds evidence to the observation in [7] that citations for Clarivate HCR contain a significantly higher number of self-citations. Indeed, the difference exists not just at level zero, but becomes even more significant in the full distribution of citations with respect to the collaboration distance.

This indicates that the distribution of citations with respect to the collaboration distance provides a more meaningful impression of the "impact" reflected by citations. However, since it obviously depends heavily on both the age of the author and the size of the subject areas, it appears not advisable to derive yet another bibliometric measure from it. Rather, the distribution should be taken into account along with other information (such as age or subject specifics), to better understand what is usually hidden in total citation figures.

8 Conclusions

We have outlined how data available from zbMATH Open can be employed for a transparent investigation of publication and citation structures in mathematics. Even these few figures make it clear that common bibliometric measures appear to be ill-suited to reflect just only the formal bibliometric structure in mathematics publications, let alone can serve as proxies for scientific excellence. Throughout the note, we indicated several further questions which may deserve a more thorough investigation, for which data are available from the zbMATH Open API. The interested reader is encouraged to pursue a deeper analysis!

References

[1] T. Bouche and O. Teschke, An update on time lag in mathematical references, preprint relevance, and subject specifics. *Eur. Math. Soc. Newsl.* **106** (2017), 37–39

[2] A. Bannister, K. Hulek, O. Teschke, Das Zitationsverhalten in mathematischen Arbeiten. Einige Anmerkungen. *Mitt. Dtsch. Math.-Ver.* **25** (2017), no. 4, 208–214

[3] M. Bras-Amorós, J. Domingo-Ferrer, V. Torra, A bibliometric index based on the collaboration distance between cited and citing authors. *J. Informetrics* **5** (2011), no. 2, 248–264

[4] T. Bouche, O. Teschke, and K. Wojciechowski, Time lag in mathematical references. *Eur. Math. Soc. Newsl.* **86** (2012), 54–55

[5] Clarivate Highly Cited Researchers in mathematics 2023. https://clarivate.com/highly-cited-researchers visited on 14 March 2024

[6] E. Dunne, Are math papers getting longer? Blog article https://blogs.ams.org/beyondreviews/2021/10/14/are-math-papers-getting-longer/, Oct 14 (2021), visited on 14 March 2024

[7] E. Dunne, Don't count on it. *Notices Amer. Math. Soc.* **68** (2021), no. 1, 114–118

[8] P. Della Briotta Parolo, R. Kumar Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato, Attention decay in science. *Journal of Informetrics* **9** (2015), no. 4, 734–745

[9] A. Ferrer-Sapena, E. A. Sánchez-Pérez, F. Peset, L.-M. González, and R. Aleixandre-Benavent, The lack of stability of the impact factor of the mathematical journals. In *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, June 29 – July 3, 2015*, edited by A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, U. Al, pp. 415–416, Bogaziçi University Printhouse, Istanbul, 2015

[10] M. Fuhrmann and F. Müller, A REST API for zbMATH Open access. *Eur. Math. Soc. Mag.* **130** (2023), 63–65

[11] K. Hulek and O. Teschke, The transition of zbMATH towards an open information platform for mathematics. *Eur. Math. Soc. Newsl.* **116** (2020), 44–47

[12] K. Hulek and O. Teschke, How do mathematicians publish? – Some trends. *Eur. Math. Soc. Newsl.* **129** (2023), 36–41

[13] M. Jost, N. D. Roy, and O. Teschke, Another update on the collaboration graph. *Eur. Math. Soc. Newsl.* **100** (2016), 58–60

[14] H. Mihaljević-Brandt and N. Roy, zbMATH author profiles: open up for user participation. *Eur. Math. Soc. Newsl.* **93** (2014), 53–55

[15] H. Mihaljević-Brandt and O. Teschke, Journal profiles and beyond: what makes a mathematics journal "general"? *Eur. Math. Soc. Newsl.* **91** (2014), 55–56

[16] F. Müller and O. Teschke, Progress of self-archiving within the DML corpus, with a view toward community dynamics. In *Intelligent Computer Mathematics. CICM 2016. Lecture Notes in Computer Science* Vol. **9791**, edited by M. Kohlhase, M. Johansson, B. Miller, L. de Moura, and F. Tompa, pp. 63–74, Springer, Cham, 2016

[17] N. Schappacher, *Framing global mathematics—the International Mathematical Union between theorems and politics*. Springer, Cham, 2022

[18] O. Teschke and B. Wegner, On authors and entities. *Eur. Math. Soc. Newsl.* **71** (2011), 43–44

List of contributors

Isabel Beckenbach, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; isabel.beckenbach@fiz-karlsruhe.de

Hagen Chrapary, Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany; chrapary@zib.de

Howard S. Cohl, Applied and Computational Mathematics Division, National Institute of Standards and Technology, 100 Bureau Drive, Mail Stop 8910, Gaithersburg, MD 20899-8910, USA; howard.cohl@nist.gov

Wolfgang Dalitz, Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany; dalitz@zib.de

Dariush Ehsani, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; dariush.ehsani@fiz-karlsruhe.de

André Greiner-Petter, Institute of Computer Science, Scientific Information Analytics Group, Georg-August-Universität Göttingen, Papendiek 14, 37073 Göttingen, Germany; greinerpetter@gipplab.org

Klaus Hulek, Institut für Algebraische Geometrie, Fakultät für Mathematik und Physik, Leibniz Universität Hannover, Welfengarten 1, 30167 Hannover, Germany; hulek@math.uni-hannover.de

Volker Mehrmann, Fakultät II – Mathematik und Naturwissenschaften, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany; mehrmann@math.tu-berlin.de

Helena Mihaljević, Hochschule für Technik und Wirtschaft Berlin, Treskowallee 8, 10318 Berlin, Germany; helena.mihaljevic@htw-berlin.de

Fabian Müller, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; fabian.mueller@fiz-karlsruhe.de

Matteo Petrera, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; matteo.petrera@fiz-karlsruhe.de

Nicolas D. Roy, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; nicolas.roy@fiz-karlsruhe.de

Lucía Santamaría, Amazon, Germany; lucia.santamaria@ymail.com

Bernd Schneidmüller, Historisches Seminar, Heidelberg University, Grabengasse 3, 69117 Heidelberg, Germany; bernd.schneidmueller@zegk.uni-heidelberg.de

Moritz Schubotz, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; moritz.schubotz@fiz-karlsruhe.de

Petr Sojka, Masaryk University, Žerotínovo nám. 617/9, 601 77 Brno, Czech Republic; sojka@fi.muni.cz

Jan Philip Solovej, Department of Mathematics, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark; solovej@math.ku.dk

Wolfram Sperber, Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany; sperber@zib.de

Johannes Stegmüller, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; current address: Mannheim, Germany; johannesst@gmx.de

Olaf Teschke, Subject-specific services, Department of Mathematics, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH Berlin, Franklinstraße 11, 10587 Berlin, Germany; olaf.teschke@fiz-karlsruhe.de

Dirk Werner, Fachbereich Mathematik und Informatik, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany; werner@math.fu-berlin.de

Klaus Hulek, Octavio Paniagua Taboada, Olaf Teschke *(eds.)*

90 Years of zbMATH

zbMATH Open, the world's most comprehensive and longest-running abstracting and reviewing service in pure and applied mathematics was founded by Otto Neugebauer in 1931. It celebrated its 90th anniversary by becoming an open access database. In December 2019, the Joint Science Conference (Gemeinsame Wissenschaftskonferenz) agreed that the Federal and State Governments of Germany would support FIZ Karlsruhe in transforming zbMATH into an open platform. In future, zbMATH Open will link mathematical services and platforms so as to provide considerably more content for further research and collaborative work in mathematics and related fields.

This book presents how zbMATH Open has reacted to a rapidly changing digital era. Topics covered include: the linkage of zbMATH Open with different community platforms and digital maths libraries, the use of zbMATH Open as a bibliographical tool, API solutions, current advancements in author profiles, the indexing of mathematical software packages (swMATH), and issues concerning mathematical formula search in zbMATH Open. We also reflect on the gender publication gap in mathematics, and focus on one of the central pillars of zbMATH Open: the community of reviewers.

https://ems.press
ISBN 978-3-98547-073-0

EMS PRESS