The Transition of zbMATH Towards an Open Information Platform for Mathematics

Klaus Hulek (Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany) and Olaf Teschke (FIZ Karlsruhe, Berlin, Germany)

Introduction

Mathematics has been both a driving force and a beneficiary of digitisation from the very beginning. Computers entered mathematical institutes early on, enabling previously unattainable calculations, and mathematical research literature was among the first to be made available electronically – both through new electronic journals and extensive retro-digitisation efforts. Review databases have long provided an overview of the state of research, and extensive collections of mathematical objects such as simulation data, integer sequences or special functions have been available electronically for more than twenty years. Platforms are widely used for collaborative work and for rapid dissemination and discussion of results.

At the same time, mathematics depends more than almost any other science on the reliability and completeness of its body of knowledge. The intrinsic interdependence of mathematical concepts and theorems means that errors in the knowledge corpus might propagate dramatically. It is a characteristic feature of mathematics that results which date far back in time are often highly relevant. The half-life of mathematical knowledge is considerable and typically significantly longer than that of most other disciplines [KTW12, AT17, AHT17]. The growing quantity, and above all the diversity of results, pose a great challenge for safeguarding the standards established in the past. Clearly, a close connection of the various digital resources is of great value for mathematical research – it would both allow much more efficient work (e.g., by avoiding duplicated efforts or reducing routine tasks) and increase the chances for completely new insights. Due to traditional business models, many valuable resources have grown monolithically as closed systems and, as a result, many desirable links have not yet been created and are difficult to establish.

zbMATH[1] (formerly Zentralblatt für Mathematik) has long been a comprehensive service with information on mathematical publications, authors and references, as well as more recently, on mathematical software. However, the reusability of this data for networked services or research purposes has been severely restricted by the traditional licensing model. Following a recommendation of the 2017 evaluation of FIZ Karlsruhe, the Joint Science Conference of the German Federal Government and the German States in 2019 decided to support the transformation of zbMATH into an open access platform. As an immediate effect, the zbMATH database will become open access from 2021, but the benefits for the mathematical community will certainly go beyond this, facilitating openly available data and the resulting networked information. Here we want to outline the next steps in this transition, as well as possible long-term prospects.

Mathematical information systems: Status quo and needs

Mathematical information has traditionally been stored primarily in publications, but mathematical data is also available today in a variety of other forms, such as mathematical software and research data, community platforms, non-textual materials or educational content. Such information is characterised by a high degree of formalisation and longevity, whereby the linking of the information requires a high quality of indexing, not least due to the mathematical formulae. Indeed, publications are still the most important medium for exchanging and securing scientific results. With zbMATH and MathSciNet[2] there are two independent, high-quality reference systems, thus providing excellent access to the published literature. A growing proportion of mathematical publications are now available via open services like such as arXiv[3] or EuDML[4].

Classical publications are being supplemented by an increasing amount of research data, which, however, are still available in mostly isolated form and have hardly been indexed. Nevertheless, they are produced and used by many mathematicians. Examples are the number sequences in the On-Line Encyclopedia of Integer Sequences[5], functions in the NIST Digital Library of Mathematical Functions[6], Calabi–Yau data[7] or the L-Functions and Modular Forms Database[8]. Mathematical modelling and simulation require and generate terabytes of numerical data.

[1] https://zbmath.org

[2] https://mathscinet.ams.org/mathscinet/index.html
[3] https://arxiv.org
[4] https://eudml.org
[5] OEIS, https://www.oeis.org
[6] DLMF, https://dlmf. nist.gov
[7] http://hep.itp.tuwien.ac.at/~kreuzer/CY/
[8] LMFDB, https://www.lmfdb.org

Community platforms play an increasingly important role in the generation and exchange of information. An excellent example is the Q&A forum MathOverflow[9]. This is a source of information, especially for young researchers, that should not be underestimated. Another example are the PolyMath projects[10], in which mathematical questions are explored collaboratively. Wikipedia[11] and Imaginary[12] reach a broad audience beyond the field of mathematics and make an essential contribution to the dissemination of mathematical knowledge, the popularisation of mathematics and scientific education.

The last two examples in particular are also an excellent illustration of the strength of open platforms: information of a different nature and difficulty can be integrated collaboratively, linked, and made available to the general public. In general, the mathematical community has been committed to open access to knowledge from the very beginning, for example through extensive use of the arXiv, early establishment of open journals or initiatives such as TheCostofKnowledge[13]. As the survey [NRW] shows, open access to mathematical knowledge continues to have a very high status in the community – only (and by a clear margin) quality assurance is considered more important. This is also important with respect to the change of the publication landscape, in particular to both secure openness and to avoid quality problems caused by commercial incentives. Above all, the outstanding importance of assured results for mathematical research can only be realised with the engagement of the community, which expects to have the results of their efforts openly available in return.

In order to meet these needs, the broadest possible free access to the various quality-assured, linked and tapped resources will be an indispensable basis for mathematical research in the future. A broad vision of such an infrastructure has been formulated by the IMU as the Global Digital Mathematics Library [GDML14]. Such a service must support mathematicians in their search for and evaluation of information, and enable them to recognise and analyse connections. The special nature of mathematical information must also be taken into account: mathematics has developed its own language in the form of mathematical formulae, which in a very compact form ensure a high degree of precision. Often the information contained in formulae cannot be found in any other form. This applies to all mathematically relevant data, and especially to research data, which is essentially formula-based and therefore requires a formula search for indexing, and further tools tailored to the management of mathematical knowledge in the longer term.

What obstacles have blocked the realisation of such an interlinked information system?

An important task for the future is to connect the resources described above with each other, develop their content and build a coherent, comprehensive, open and sustainable platform. This platform must contain information of different levels in order to guarantee a comprehensive knowledge transfer. A standardised evaluation of research data increases its reproducibility, which further serves to assure the quality of publications. Additionally, a systematic citation of research data must be made possible, which in turn ensures the recognition of the research work contained within it. It is also essential to make highly formalised mathematical content such as formulae accessible. Even though these developments are still in their infancy, they indicate the great potential of modern information technology for mathematics.

Free access to interconnected and annotated resources will be an indispensable basis of mathematical research in the future. This will help mathematicians to search for and evaluate this information; connections can be clarified and analysed. The high quality of all available information remains an essential prerequisite. However, a large part of the available information described above is currently not freely available or is not or only poorly developed. As a result, information is sometimes not findable or is lost due to the lack of interconnection.

Functions and contents of an open platform must go far beyond the current possibilities of closed systems such as zbMATH, MathSciNet, Google Scholar[14], ResearchGate[15] or WolframAlpha[16]. Currently, a large part of the existing information described above is neither available via open access nor as open data, nor is it sufficiently indexed or findable. As a result, much information is lost due to the lack of interconnection. Each individual information service provides individual building blocks of information which, if correctly combined and analysed, will lead to new findings. Figure 1 (see next page) shows a selection of possible different sources of information, each of which must be linked within the individual segments and also with each other.

These links are only the beginning: open interfaces give the community the opportunity to use existing data for their own research and to subsequently make research results retrievable. These can be publications or research data as well as tools for using or analysing the data. In this way, the community itself makes innovative contributions to the development of a comprehensive mathematical research system. All experience so far shows that this would uncover considerable and often unpredictable potential for innovation.

There are several kinds of obstacles which currently delay the realisation of such an ecosystem. Firstly, business models which exist for historically good reasons (not least to create resources needed for ensuring scientific

[9] https://mathoverflow.net/
[10] https://polymathprojects.org, see also [P14]
[11] https://www.wikipedia.org
[12] https://imaginary.org; see also [GMM14]
[13] http://thecostofknowledge.com
[14] https://scholar.google.com/
[15] https://www.researchgate.net/
[16] https://www.wolframalpha.com/

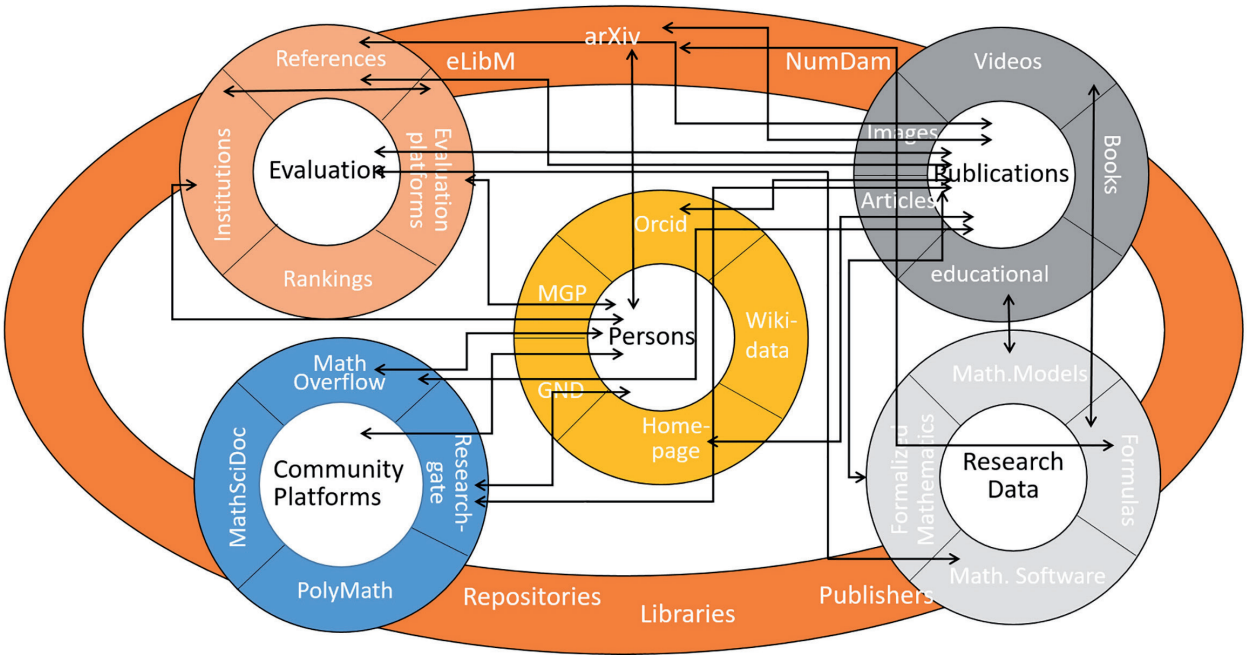

Figure 1. Components of an interlinked information system for mathematics.

quality) are in conflict with open solutions. Secondly, several techniques desirable for automated interlinking, indexing and knowledge management of digital mathematics are not yet mature or even existing for full-scale implementation (e.g., formula OCR, automated verification tools). Thirdly, many facets of such a networked, collaborative system would require extensive community engagement (like it has grown over decades for reviewing activities). Without doubt, this is the most critical, scarce and valuable resource, which requires manifest benefits to become activated.

What would the next steps be for zbMATH Open?

Due to the support of the German government, zbMATH will be able to offer its services open access from 2021 and thus overcome the previous limitations due to its traditional business model. Most of the data will become open via a CC-BY-SA license (with the exception of third-party content such as author summaries, which may be provided under a different license).

These open data will be freely available for independent research and development. As it is, there are already many projects on, for example, the history of mathematics, which make use of zbMATH data. Such projects and the resulting services will be much facilitated by open data. Another example is the use of zbMATH data in the ISC project Gender Gap In Science[17], led by the IMU. Part of the project results is the visualisation platform Gender Publication Gap[18]. In order to become valuable resources, these projects face the challenge that they require long-term maintenance. A key task will be to identify important project results involving zbMATH data, to integrate them into its services if possible, and to make them sustainably available.

Publications are still the crucial core of mathematics research. In the last decades, a growing portion has become openly available (see [T18] for a qualitative and quantitative overview). Many developments accelerate this: the growing share of arXiv submissions [MT16B], implementation of moving wall policies, open retrodigitisation projects, transformative publisher agreements and initiatives like Plan S[19]. However, the technical level (as well as the used license) of the results is very different, ranging from plain scans to available LaTeX sources which already facilitate full-text formula search [MT16A]. Integration of these diverse resources with the help of open zbMATH data will greatly improve the current situation, and create a framework which will finally make full-text related services for indexing and retrieval available.

As outlined in [HMST19], mathematical research data will become an issue of strongly growing relevance. However, these data are very diverse by nature, and it is a great challenge to make them available sustainably according to the FAIR principles[20]. Again, their integration into a linked system involving open publication data will be a big step forward. This will enhance their retrieval and visibility and will allow mechanisms to be set up to ensure their quality. Some years ago, such an approach was started for mathematical software which evolved into the free swMATH service[21]. A similar service for the entirety of mathematical research data is a task that is far too complicated to be achieved within the zbMATH Open framework alone. Instead, we will pursue this for several tech-

[17] https://gender-gap-in-science.org/
[18] https://gender-publication-gap.f4.htw-berlin.de/

[19] https://www.scienceeurope.org/our-priorities/open-access
[20] Findability, Accessibility, Interoperability, and Reusability; see https://www.go-fair.org/fair-principles/
[21] https://swmath.org, see also [BGS13, CDNS17]

nology-ready resources, and make the results available to further networked activities like those envisioned in the currently formed MaRDI consortium[22] in the framework of the German National Research Data Infrastructure[23].

As in the case of mathematical software, efforts to create, develop and maintain research data collections are often not adequately appreciated. Such an interlinked platform would also greatly help its creators to obtain recognition for contributions not directly related to publications. When their widespread usage becomes visible in a linked system, it will be much easier for the mathematicians, as well as their institutions, to obtain the appropriate recognition and reputation for their efforts, and hopefully to obtain sufficient funding for services which are currently often pursued on a voluntary and therefore not always sustainable basis.

Likewise, the role of non-textual material such as videos has changed significantly – something we are experiencing right now when we try to maintain mathematics dissemination and education under lockdown circumstances. Even after the threat of SARS-CoV-2 is reduced, experiences from the forced distancing will probably have a lasting impact on communicating mathematics research. It seems unlikely (and maybe even undesirable, since, e.g, the much larger threat of global warming still remains) to quickly restore the same overfilled conference calendars as before. Instead, we will surely make further use of remotely communicating mathematics. Resources created in this course, however, will require the same efforts of integration, interlinking, sustainable preservation, and indexing as research data.

Similar work must be done for the vast amount of mathematical knowledge assembled by the already mentioned digital mathematical communities. While first steps have been taken, e.g., interlinking Wikidata information into zbMATH author profiles with the benefit of timeline creation, or the interlinking of publication with MathOverflow discussions [MST19], much more will be possible by empowering the creativity of mathematicians with appropriate tools. It is upon us all to define what are the most needed tools which make use of the open data and can be implemented as the next steps – your suggestions and feedback to editor@zbmath.org will be highly appreciated!

References

[AT17] A. Bannister, O. Teschke: An update on time lag in mathematical references, preprint relevance, and subject specifics. *Eur. Math. Soc. Newsl.* 106, 37–39 (2017; Zbl 1384.01099)

[ATH17] A. Bannister, K. Hulek, O. Teschke: Das Zitationsverhalten in mathematischen Arbeiten. Einige Anmerkungen. *Mitt. Dtsch. Math.-Ver.* 25, No. 4, 208–214 (2017; Zbl 1390.01104)

[BGS13] S. Bönisch, G.-M. Greuel, W. Sperber, The software information service swMATH – release of the first online prototype. *Eur. Math. Soc. Newsl.* 87, 48–49 (2013; Zbl 1296.68186).

[BTW12] T. Bouche, O. Teschke, K. Wojciechowski: Time lag in mathematical references. *Eur. Math. Soc. Newsl.* 86, 54–55 (2012; Zbl 1263.01012)

[CDNS17] H. Chrapary, W. Dalitz, W. Neun, W. Sperber: Design, concepts, and state of the art of the swMATH service. *Math. Comput. Sci.* 11, No. 3–4, 469–481 (2017; Zbl 1382.68320)

[GDML14] *NAS Committee on Planning a Global Library of the Mathematical Sciences: Developing a 21st Century Global Library for Mathematics Research* (2014), doi: 10.17226/18619

[GMM14] G.-M. Greuel, A. Matt, A. Mey: IMAGINARY – mathematics communication for the 21st century. *Eur. Math. Soc. Newsl.* 92, 3–6 (2014; Zbl 1302.00008)

[HMST19] K. Hulek, F. Müller, M. Schubotz, O. Teschke: Mathematical research data – an analysis through zbMATH references. *Eur. Math. Soc. Newsl.* 113, 54–57 (2019; Zbl 1422.01015)

[MST19] F. Müller, M. Schubotz, O. Teschke: References to research literature in QA forums – a case study of zbMATH links from MathOverflow. *Eur. Math. Soc. Newsl.* 114, 50–52 (2019; Zbl 1428.68357)

[MT16A] F. Müller, O. Teschke: Full text formula search in zbMATH. *Eur. Math. Soc. Newsl.* 102, 51 (2016; Zbl 1366.68355)

[MT16B] F. Müller, O. Teschke: Will all mathematics be on the arXiv (soon)? *Eur. Math. Soc. Newsl.* 99, 55–57 (2016; Zbl 1345.68267)

[NRW17] C. Neylon, D. M. Roberts, M. C. Wilson: Results of a worldwide survey of mathematicians on journal reform. *Eur. Math. Soc. Newsl.* 103, 46–49 (2017; Zbl 1366.01071)

[P14] D.H.J. Polymath et al.: The "Bounded gaps between primes" Polymath project – a retrospective analysis. *Eur. Math. Soc. Newsl.* 94, 13–23 (2014; Zbl 1315.01078)

[T18] O. Teschke: Green, gold, platinum, nickel: on the status of open access in mathematics. *Eur. Math. Soc. Newsl.* 110, 60–63 (2018; Zbl 1403.01045)

Photo: Sebastian Gerhard

Klaus Hulek studied mathematics at the University of Munich and the University of Oxford. He completed his PhD in Erlangen in 1979 and spent the academic year 1982/83 at Brown University, Providence, RI, USA. He received his habilitation in Erlangen in 1984. Klaus Hulek was a professor of mathematics in Bayreuth from 1985 to 1990, when he moved to Hanover. He held the position of Vice-President of Research of Leibniz Universität Hannover from 2005 to 2014, and from 2007 to 2014 he represented the German Rectors' Conference (HRK) at the Research Policy Working Group of the European University Association (EUA). Klaus Hulek was a member of the Institute for Advanced Study (IAS) in Princeton in 2015. He is currently Vice-President of the German Mathematical Society (DMV). His field of research is algebraic geometry.

Olaf Teschke studied mathematics at Humboldt University Berlin, and completed his PhD in algebraic geometry there. He moved to FIZ Karlsruhe in 2008, and has been working there since 2009 as the head of the Mathematics Department and Managing Editor of zbMATH (including a short intermediate term as Editor-in-Chief). Since 2017, he has been serving as the Vice Chair of the EMS Committee on Publications and Electronic Dissemination. His main occupation is information infrastructure for mathematics.

[22] Mathematical Research Data Initiative, https://wias-berlin.de/mardi/

[23] NFDI, https://www.dfg.de/foerderung/programme/nfdi/