Mathematical Research Data – An Analysis Through zbMATH References

Klaus Hulek (Gottfried Wilhelm Leibniz Universität Hannover), Fabian Müller (FIZ Karlsruhe, Berlin, Germany), Moritz Schubotz (FIZ Karlsruhe, Berlin, Germany), Olaf Teschke (FIZ Karlsruhe, Berlin, Germany)

Are there mathematical research data?

In fitting with our data-driven age, research data have become an increasingly important aspect of scientific work. In Germany, the Federal Ministry for Education and Research has launched a program to build a national research data infrastructure. Correspondingly, the DFG issued a call to form consortia dealing with the management of such research data. Within the German mathematical community a proposal to establish MaRDI (Mathematical Research Data Initiative) is prepared [1]. One may initially wonder what the mathematical equivalent of the vast amount of LHC measurements or data from clinical trials might be. Indeed, as one of the driving forces of storing research data has been the reproducibility crisis in several fields, one may ask whether storing research data is relevant to our subject at all, since mathematical results usually come with an inherently much higher level of confirmability than those connected with empirical scientific methods.

However, reproducibility is just one aspect connected to research data, and perhaps not even the most important one in the future. Storing and sharing research data according to the FAIR principles (Findability, Accessibility, Interoperability, Reusability) generates several benefits for mathematicians (as for all scientists):

1. Improved citability: work that does not fit the classical format of journal articles or books should still be adequately acknowledged and cited when used as a basis for further work.
2. Better findability: appropriate data repositories (ideally, intrinsically cross-linked with each other, as well as the literature) would enable mathematicians to easily identify prior results on a different level rather than just entangled in the context of an article.
3. Confirmability: For appropriate peer review, computational results must be available to redo the computations, or provide a way to confirm the correctness of the results.
4. Reusability: research data should be available in a form that facilitates building upon these results in a manner that is as efficient as possible. This also prevents the unnecessary repetition of work and uses human resources and available publication space more efficiently.
5. Long-term preservation: storage of research data in a dedicated infrastructure framework ensures that its longevity is independent of individuals or institutions.

How urgent are these aspects for mathematics? Historically, our subject has been the origin of arguably the most frequently used data: from Babylonian multiplication tables to Greek and Indian tables for sine values to the logarithmic tables ubiquitous for calculations until the second half of the 20th century. While computers have made such tables obsolete, they also generate a vast landscape of new resources. Today, mathematical research data may still derive from tables like collections of special functions, algebraic representations or combinatorial data, but likewise exist as libraries of formalised mathematics or be generated by extensive computations involving computer algebra systems or numerical simulations. Based on zbMATH references, we will derive a rough heuristic of the current usage and discuss some examples.

A heuristic analysis of possible research data references

In this section, we report on the current status of our preliminary investigations. A more in-depth analysis is in preparation.

The zbMATH database [2] currently contains more than 30M references. Of those, currently 53.7% link back to other publications that are indexed in zbMATH. Other references are out of the scope: overall, 36.7% have a DOI and 10.9% have a DOI, but not one connected to a publication within zbMATH. One can estimate from this that more than 75% of references are connected to the published literature. Moreover, much of the rest consists of literature available at the arXiv, other repositories, or personal homepages.

We used the following heuristic to detect links to non-literature online resources. There are about 795,000 references containing a (‘http’, ‘www.’, ‘ftp’) link to a website. Excluding the most common patterns to literature repositories leaves us with about 161,000 links. Of these, 20,518 are links to mathematical software as identified in the swMATH database [3]. For the remaining 141,000 references, we identified 3 common link patterns: references to mathematical online compendia such as the Online Encyclopedia of Integer Sequences [4], references to normative data like standards or benchmarks, and references to community-maintained websites such as Wikipedia or MathOverflow. There is a large variety of different links included, and it becomes clear that there is an extremely long tail of specific data used in relatively few publications. Although we did not yet identify a suitable method to classify the links automatically into reasonable categories, the general structure of the sample analysed in [5] could be confirmed. To give an impression, we will present some examples in the following.

Examples

Singular

While it is still debated whether software code should be considered as research data, its output certainly is. Here we will take the example of the computer algebra system SINGULAR [6], which is widely used and has been frequently cited in mathematical papers throughout the last two decades (http://purl.org/zb/1). Here, as for other mathematical software, we can employ the swMATH database to track its usage in mathematical papers, although it is frequently referenced in a rather diverse form, ranging from the direct weblink or the manual to the related book [7] (see [8] for the current status of standardisation for software citations). An analysis of these publications reveals that the involved computational results almost never exist in a fully FAIR form, although the initial additional effort would likely pay off greatly in the long term.

This appears to be a general issue for computational results: The recent article [9] demands (emphasizing the reproducibility aspect) that results should be reproducible in identical, and comparable to runs in varied, settings. For long-running computations, this involves in particular the explicit saving of intermediate states (checkpoints). This involves among other things an exact specification of the computing environment used (software, libraries, versions, etc.) and the possibility for the full publication of all relevant entities (i.e. code/algorithms together with input datasets and results). Overall, while mathematics already enjoys an appropriate service to interlink information on the used software via swMATH, the task of adequately documenting the computational output still needs to be addressed.

DLMF

The NIST Digital Library of Mathematical Functions (DLMF, [10]) is among the most frequently cited collections identified through the above approach. It is the successor of the *Handbook of mathematical functions with formulas, graphs and mathematical tables* [11], which is currently the most cited document in zbMATH (http://purl.org/zb/2) with about 10,000 citations gathered by its five different editions. In comparison, there are still much fewer references (about 1,500) to the electronic version recorded by the DLMF entry (http://swmath.org/software/4968), although referencing to a function or formula can be done much more precisely within the DLMF, as in the handbook. The attitudes to citing such data appear to be changing slowly, but steadily; the ratio of DLMF citations has increased in recent years. This is also confirmed by a recent study by the NIST library [12] based on citation data from the Web of science dataset, which obtained a similar pattern (cf., Fig. 1). According to the NIST data analysis and the assumption of a linear growth model, the DLMF will be cited more often than the printed book as early as 2028. As depicted in Figure 1, our data from swMATH confirm a linear growth of the citation rates of the DLMF project. Despite the prominent link to citation instructions (http://purl.org/zb/4) only fewer than 1% of all citations that DLMF received in the zbMATH database use a deep link to a chapter, formula or section.

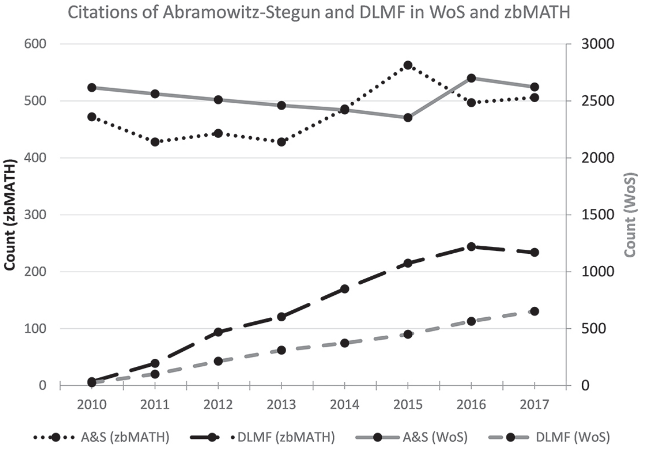

Fig. 1. Citations counts of the Handbook of mathematical functions and DLMF inWoS (according to the NIST library) and zbMATH. The citation counts of the online versions growwith a constant factor in contrast to the citation counts of the printed version

This example illustrates that although the heuristic above may be helpful in identifying interesting datasets, the distinction "literature" vs. "data" may be extremely misleading, since many literature references may in fact be research data in disguise (a pattern that we also already noted in the relation between software and related publications). This can also be seen by the next example.

OEIS

The On-Line Encyclopedia of Integer Sequences (OEIS [4]) is a browsable and searchable online resource launched in 1996 that grew out of N.J.A. Sloane's 1973 book *A Handbook of Integer Sequences* [13]. Starting in 1994, there are 2,752 references to it in zbMATH (http://purl.org/zb/5).

Of these, more than 70% cite OEIS as a whole, while the remaining refer to one or, in about 5% of the cases, several actual entries of the database (with a single reference citing as many as 14 sequences in one case).

However, in contrast to the previous example, the references to the online service have quickly substituted those to the printed handbook (compare to http://purl.org/zb/6). The easy usability of the OEIS and its powerful search features (which benefit from the rather simple data shape of integer sequences) appear to be a crucial factor here, making it a model for highly findable, accessible, and reusable mathematical data. Nevertheless, interoperability remains an issue even for this resource. Currently, one can only dream of seamlessly cross-linking the generating functions of sequences in OEIS with respective entries in DLMF – a service which would open a whole new dimension of opportunities.

Calabi–Yau data

Lists of Calabi–Yau manifolds play a crucial role not just within mathematics, but due to their relation to string theory in physics. The data available at [14] are among the most prominent (although it is once more impossible to determine its real use, since related original publications are still as frequently cited as the data itself, see http://purl.org/zb/7). They also form a model case in the sense that both the software and the computational output were made available in a transparent, reusable form. However, this static page also illustrates an urgent issue manifest for many research data. Due to the untimely death of its creator, it has remained in a frozen state ever since, and its status with respect to sustainability is completely unclear (which is only underscored by several links to further Calabi–Yau sites which have partially ceased to exist). Many examples of such valuable resources in a potentially precarious state exist throughout the references and underscore the need for a more sustainable framework.

Further resources, big data vs. deep data, interdisciplinary issues

The reader is free to explore further examples by analysing our dataset of non-literature references generated by the procedure described above) available at github (http://purl.org/zb/8), e.g., by checking for entries collected in the catalogue of mathematical datasets [15]. As indicated by the discussed examples, mathematical research data are typically no "big data" of many terabytes (although there exists, e.g., the rather large collection of finite lattices [16]) but come along with highly diverse and sophisticated descriptional metadata, necessary to facilitate their FAIR usage. In this sense, mathematical metadata are rather "deep data" [19], which would require extensive semantic enrichment before they could be properly cross-linked with each other and the literature, finally leading to a framework from which a mathematician could benefit in everyday work. The vision of a Global Digital Mathematics Library [17] can be understood as such an infrastructure.

Another important aspect is, of course, interdisciplinarity. Mathematics, as the language of exact sciences, is naturally connected to other disciplines, which have their own collections of research data. These are often of a different nature, and are preserved according to the standards of the discipline. Large genome or medical datasets may also be of interest for mathematical work, but are associated with quite different legal and computational aspects. One may even ask whether a precise definition of mathematical research data is possible; certainly, the distinction is not always as clear as between Calabi–Yau data (mathematical) and LHC measurements (physical) in high-energy physics.

Mathematical modelling and simulation are now omnipresent in many sciences, and the related computations open up a whole new dimension of interdisciplinary research data [18]. Hence, a FAIR framework for mathematical research data would also require interfaces to application areas potentially dealing with them.

Conclusion and future work

Research data are widely used within mathematics, and their sustainable storage and FAIR availability will very likely become an important issue in the future. The requirement of an utmost level of confirmability for mathematical results in connection with the growing importance of computer aided computations and proofs will almost certainly be a driving force in establishing standards which should eventually lead to an interconnected, powerful infrastructure. However, the amount of work required to reach this goal is substantial: mathematical research data exists in very different forms, from small databases through to diverse software and its output to huge amounts of data, some of them created in collaboration with other sciences. Currently, they are not even always referenced in a transparent manner, but are often intrinsically connected to the literature. Building a framework that cross-links the various types of mathematical research data will require substantial metadata and semantic enrichment, enabling them to serve as "deep data" in an infrastructure facilitating new research dimensions, not just within mathematics but also its applications.

To achieve this goal, we at zbMATH are investigating diverse approaches: For one, we analyze citation data and mathematical formulae to identify similar (or even plagiarized) content [25]. Moreover, we connect our datasets to external datasets such as Wikidata or MathOverflow [23, 24]. Additionally, after having switched to a LaTeX the input format for zbMATH reviews [22], we are considering to allow for semantically enriched LaTeX dialects as used in the DLMF and DRMF [21] projects, or optional semantic annotations for mathematical formulae via graphical tools [20].

Acknowledgements

We thank the DLMF editor Ronald Boisvert for sharing the DLMF citation analysis with us.

References

[1] MaRDI, https://www.wias-berlin.de/mardi

[2] zbMATH, https://zbmath.org

[3] swMATH, https://swmath.org

[4] The On-Line Encyclopedia of Integer Sequences (https://oeis.org)

[5] O. Teschke: Some heuristics about the ecosystem of mathematics research data. *PAMM* 16, No. 1, 963–964 (2016)

[6] https://www.singular.uni-kl.de/search.html

[7] G.-M. Greuel, G. Pfister: *A Singular introduction to commutative algebra*. Berlin: Springer. (2002).

[8] M. Kohlhase, W. Sperber: Software citations, information systems, and beyond. In: 10th int. conf. CICM2017, Edinburgh, UK. *LNCS* 10383, 99–114 (2017).

[9] M.A. Heroux, Trust Me. QED., SIAM News July 2019, https://sinews.siam.org/Details-Page/trust-me-qed

[10] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.23 of 2019-06-15. F. W. J. Olver, A. B. Olde Daalhuis, D.W. Lozier, B.I. Schneider, R.F. Boisvert, C.W. Clark, B.R. Miller, and B.V. Saunders, eds.

[11] M. Abramowitz, I. A. Stegun: *Handbook of mathematical functions with formulas, graphs and mathematical tables*. Washington: U.S. Department of Commerce. xiv, 1046 pp. (1964).

[12] K. Rapp: Citation Analysis for the NIST Handbook of Mathematical Functions, 2007–2017, Report of the Information Service Office, NIST (07-2018).

[13] N.J.A. Sloane: *A handbook of integer sequences*. New York-London: Academic Press, a subsidiary of Harcourt Brace Jovanovich, Publishers (1973).

[14] http://hep.itp.tuwien.ac.at/~kreuzer/CY/

[15] K. Berčič: Catalogue of Mathematical Datasets. https://mathdb.mathhub.info

[16] J. Kohonen: Lists of finite lattices (modular, semimodular, graded and geometric), doi:10.23728/b2share.dbb096da4e364b5e9e37b982431f41de

[17] Developing a 21st Century Global Library for Mathematics Research. Washington, DC: The National Academies Press. doi:10.17226/18619.

[18] T. Koprucki, K. Tabelow: Mathematical models: a research data category? In: Mathematical software – ICMS 2016. 5th int. conf., Berlin, Germany, July 11–14, 2016. *Lecture Notes in Computer Science* 9725, 423-428 (2016)

[19] M. Schubotz: Augmenting mathematical formulae for more effective querying & efficient presentation epubli 2017, ISBN 978-3-7450-6208-3, pp. 1-212 doi:10.14279/depositonce-6034

[20] M. Schubotz et ak.: VMEXT: A Visualization Tool for Mathematical Expression Trees, in proc. 10th int. conf., CICM 2017, vol. 10383, pp. 340–355. doi:10.1007/978-3-319-62075-6_24

[21] H. Cohl et al.: Growing the Digital Repository ofMathematical Formulae with Generic LaTeX Sources. In: Proc. Int. Conf. CICM2015, LNCS 9150, vol. 9150, doi:10.1007/978-3-319-20615-8_18

[22] M. Schubotz, O. Teschke: Four Decades of TeX at zbMATH. EMS Newsl. 6 (2019), 50-52. doi:10.4171/NEWS/112/15

[23] W Dalitz et al.: alsoMATH - A Database for Mathematical Algorithms and Software in Intelligent Computer Mathematics - 12th International Conference, CICM2019.Workshop on LargeMathematical Libraries

[24] J. Corneli andM. Schubotz: math.wikipedia.org: A vision for a collaborative semi-formal, language independent math(s) encyclopedia," in Proc. Int. Conf. on Artificial Intelligence and Theorem Proving, 2017.

[25] M. Schubotz et al.: Forms of Plagiarism in Digital Mathematical Libraries, in Proc. Int. Conf. 12th CICM

Pictures and CVs of the authors can be found in previous Newsletter issues.