MathBases: Indexing and Creating Mathematical Databases

Katja Berčič[1,2][0000−0002−6678−8975], Edgar Costa[3][0000−0003−1367−7785], and David Roe[3][0000−0003−3585−7525]

[1] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[2] Institute of Mathematics, Physics, and Mechanics, Ljubljana, Slovenia
katja.bercic@fmf.uni-lj.si
[3] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
edgarc@mit.edu, roed@mit.edu

Abstract. MathBases (mathbases.org) is an open-source platform with two functions: *exposure* for existing mathematical databases and *education* to aid in the creation of new ones. It provides a searchable index of over 130 existing databases, annotated with structured metadata (area, access method, license, authors), as well as resources aimed at the creation of new projects. We describe the site in detail and raise a set of open questions that we hope will catalyze discussions about the state and direction of the community.

Keywords: mathematical databases · data discovery · open data

1 Introduction

In recent decades, easy access to powerful computers and advances in computer algebra systems have led to the creation of mathematical databases across many different disciplines. These resources range from the On-Line Encyclopedia of Integer Sequences [19] in combinatorics and the L-functions and modular forms database [17] in number theory to the House of Graphs [10] in graph theory and the π-Base [22] in point-set topology. Each of these websites provides invaluable examples that can be used for refining conjectures and analyzing statistics of the objects included, and represents years of work on the part of communities of experts. However, a much broader ecosystem of data exists, spread across personal websites, GitHub repositories, and appendices to journal articles. Most are difficult to discover, and their creators receive little academic credit, especially when compared to the substantial amount of work required to assemble and curate these community resources [5,2].

MathBases[4] aims to improve the ecosystem of mathematical data in two complementary ways: *exposure* and *education*. A user's initial view of the site is

[4] Live at mathbases.org, archived on Internet Archive's Wayback Machine on April 9, 2026.

as a searchable index of mathematical databases, organized by area. We hope that this index will help mathematicians find data that will aid in their research and also inspire some to create new resources that will be of broad utility. The second half of MathBases consists of a collection of tutorials, templates, and practical articles aimed at equipping new database builders with an accumulation of wisdom from those who have gone before. Together, we believe that these two ingredients will both increase the reach of existing datasets and accelerate the development of new ones.

Our goal in this paper is to further MathBases' mission of exposing mathematicians to the wide variety of mathematical data available online, and to engage the broader community in a discussion of metadata standards, data reliability, credit for database work, and identification of gaps in coverage. The remainder of this paper is structured as follows. Section 2 briefly surveys the landscape of mathematical collections and the challenges of navigating it. Section 3 describes the MathBases platform. Section 4 covers community and governance. Section 5 concludes with a call for participation.

2 The Landscape of Mathematical Collections

Most mathematical resources organize knowledge around statements, definitions, and proofs. Journal articles, encyclopedias, and wikis are primarily textual, and general-purpose web search works reasonably well for navigating them. To some extent, this is true even for libraries of formalized mathematics. MathBases is instead concerned with resources focused on examples, which range from small collections of counterexamples to large databases of objects generated by computation. Their utility depends more on the user's ability to search by mathematical properties than by keywords, so they require a different kind of infrastructure [12].

Collections of mathematical objects take many forms. At one end of the spectrum are *curated* collections, which gather examples selected for their interest or noteworthiness. The On-Line Encyclopedia of Integer Sequences (OEIS) [19], with almost 400,000 sequences, is the paradigmatic example: entries are contributed and reviewed by a large volunteer community and are accompanied by formulas, code, and bibliographic references. The House of Graphs [10] similarly collects graphs that have appeared as counterexamples or extremal cases in the literature. The π-Base [22] catalogs topological spaces serving as counterexamples in general topology, together with their properties and the theorems connecting them. The NIST Digital Library of Mathematical Functions [20] provides a curated reference for special functions, with formulas and graphs.

At the other end of the spectrum are *enumeration* databases, which aim to list all objects of a given type. Since many interesting classes of mathematical objects are infinite, a common approach is to enumerate them up to a size or complexity bound. The LMFDB [17] comprises billions of arithmetic objects (L-functions, modular forms, elliptic curves, number fields, and more) generated by large-scale computation. The Small Groups Library [3] provides complete catalogs of groups of small order (up to isomorphism), encoding over 4×10^8 groups in roughly 30 MB through efficient encodings of group families.

Collections of examples differ widely in scope, format, access method, and the type of objects they contain (see Section 3.3 for a discussion of the latter). Some offer web interfaces; others are downloadable datasets or packages integrated into computer algebra systems.

Licensing is complicated by the legal status of mathematical data. Individual mathematical facts and formulas are generally not copyrightable, but a database as a whole can be protected when the selection or arrangement of its contents constitutes an original intellectual creation; both the European Union's Database Directive [14] and United States copyright law [23] recognize this principle. A curated collection or a completeness result may therefore carry copyright protection even when the individual entries do not. In practice, licensing across the landscape ranges from explicit open licenses to no license at all.

A shared index helps with several challenges in the mathematical data ecosystem. Because mathematical databases are rarely recognized as scholarly contributions on their own, much mathematical data remains "in somebody's bottom-left drawer" [5]: produced as a byproduct of research, never publicly released or announced. Even when data are public, they may be difficult to find because they are hosted on personal pages. The "dark long tail" of science [15] is especially pronounced in mathematics, which has historically produced little infrastructure for data sharing [5]. Even an incomplete index helps researchers see more of what already exists and notice gaps where further exploration would be valuable. Collections maintained by a single researcher or small group are also vulnerable to link rot and abandonment; long-term stewardship requires community awareness and shared norms, both of which presuppose discoverability.

These challenges are not new. Tables and handbooks of mathematical data have circulated for millennia [7], and many of today's datasets have predecessors that predate the internet by decades; the same difficulties of visibility, maintenance, and credit have arisen throughout. They are becoming more acute, however, as the mathematical community increasingly relies on data-driven methods, and as initiatives such as the FAIR principles [24] and the German National Research Data Infrastructure (NFDI) and its mathematics consortium MaRDI [18] raise expectations for data accessibility and reuse.

3 The MathBases Platform

3.1 Index of databases

The most visible component of MathBases is its index of mathematical databases. The metadata recorded for each entry can include:

- the mathematical area(s) covered and finer-grained topic tags;
- the primary URL of the database and, when available, related URLs for code, downloads, or further documentation;

- the access method (e.g., web interface, API, downloadable dataset, or computer algebra system package);
- the license under which the data are available;
- the number of objects, the number of datasets, and the approximate size of the collection;
- author and maintainer contact information;
- bibliographic references (e.g., DOI, arXiv identifier, or URL);
- a free-text description highlighting the scope and notable features of the database.

As an example, the LMFDB entry records roughly 2.9 billion objects, lists 117 contributors, and describes more than 20 datasets.

Entries are stored as Markdown files with YAML front matter in the public Git repository github.com/MathBases/MathBases. We provide a standard entry template.[5] This format is both human-readable and machine-parsable: the YAML header carries the structured metadata, while the Markdown body provides the free-text description. The repository is built into the website using Jekyll,[6] a static-site generator. Contributors add or update entries through pull requests, and every change is version-controlled.

All metadata are published under a Creative Commons Attribution-ShareAlike (BY-SA) license, making it freely reusable.

Browse and search The MathBases homepage presents the index as an interactive table listing all entries at once. Columns show the database name, with the title linking to the database itself, together with author information, references, mathematical areas, and topic tags; each row also includes a link to a dedicated MathBases entry page with the full metadata and description. Entries are ordered alphabetically by name and can be reordered by clicking sortable column headers.

Users can type free-text queries into a search box to filter the table. Searches are matched against both the visible metadata and hidden search fields such as the short description, so a query such as "graph" returns databases whose name, areas, tags, or descriptions mention graphs.

3.2 Resources for database creation

We hope that the many examples of mathematical databases in the index portion of MathBases will inspire mathematicians to create new ones. To that end, we provide a collection of tutorials, templates, and guides so that mathematicians who are new to such an endeavor can benefit from the experience of those who have worked in the area for years.

Many of the same practical questions arise for any database project: how to compute, store, and distribute data at scale; what to consider when selecting

[5] github.com/MathBases/MathBases/blob/main/template.md
[6] `https://jekyllrb.com`

a license; the importance of defining a schema; how to present data usefully to users; and how to ensure the reliability and long-term survival of the data.

At the moment, these tutorials are still rudimentary, and we encourage contributions from experienced database developers. We are also working to create templates, offering several models for code to present a database on a website. Since modifying existing code is easier than creating an application from scratch, we hope that these templates will ease the process of getting a new project off the ground.

3.3 Questions of scope

MathBases currently focuses on resources that clearly fit its scope. This focus is what keeps the index useful: a visitor should be able to find what they need without browsing through hundreds of entries. Many valuable resources sit near the edges of that scope, and editorial discussions have surfaced a few recurring categories where accommodating them well would first require better searching and filtering tools.

One concerns the definition of *object*. The Complexity Zoo [1] catalogs complexity classes, which seem more like meta-objects than examples of mathematical structures; the Quantum Algorithm Zoo [16] similarly lists algorithms rather than instances.

A second concerns the form in which content is *organized*. Cantor's Attic [8] is a wiki on large cardinals, and the Dictionary of Algorithms and Data Structures [4] is arranged as a dictionary; both are built around definitions and cross-references rather than enumerated examples, even though the underlying content is not obviously out of scope. The NIST Digital Library of Mathematical Functions [20] sits at a similar edge; we include it both because of its standing as a long-established reference and because an index of mathematical resources that omitted it would strike most readers as incomplete.

A third is the *boundary between a tool and a database*. KnotPlot [21], and Group Explorer [9] sit closer to interactive tools than to catalogs; the Combinatorial Object Server [11] sits close to the line, described in editorial notes as "more like a combinatorics CAS in a browser". PHOEG [6,13][7] sits on the same boundary but is included, both for the underlying store of graphs and extremal invariants that users can query independently of the conjecture-exploration interface, and because of its shared authorship and infrastructure with the House of Graphs.

A fourth is *scale*. Small catalogs appended to individual papers raise the question of the minimum extent at which an entry becomes broadly useful, though historical importance, or entries following a consistent structured format, can weigh in favor of inclusion.

A fifth is the *intended audience*: MathBases has so far oriented itself towards resources useful in research mathematics, while some collections serve primarily educational, recreational, or expository purposes.

[7] Available at phoeg.umons.ac.be/phoeg/

Broadening the index to accommodate more of these cases is something we would like to support in the future, but it would first require richer filtering and search, so that a larger catalog remains at least as useful as the current one.

3.4 History

MathBases succeeds the earlier MathDB project, from which it inherits the initial catalog and core motivation.

In late 2018, Berčič began surveying existing collections of mathematical objects as groundwork for MathDataHub[8], a planned platform for hosting mathematical datasets. She compiled a catalog at `mathdb.mathhub.info` and contacted the authors of roughly twenty collections by email, asking about their motivations, licensing practices, and needs. About half responded. The responses confirmed that mathematical collections differ not only in mathematical content but also in infrastructure, community, and intent. Some authors had built searchable web interfaces backed by large collaborations; others maintained static files on a personal homepage. Motivations ranged from providing test cases for algorithmic experiments to publishing classification results as reusable data. Attitudes toward licensing varied widely: one respondent held that mathematical data cannot be licensed; others used formal open-source licenses; several had not considered the question at all. A concern shared by nearly all respondents was the effort required to maintain a collection over time and the risk that data are lost when a maintainer retires or changes institutions. Several wished for long-term storage independent of individual employment and for standard formats that would reduce the cost of making data available. The resulting catalog, MathDB, grew to list over a hundred databases but was not actively maintained for several years before being revived and expanded as MathBases.

Roe ran across MathDB in spring 2022 and Costa met Berčič at a Dagstuhl workshop in fall of 2023. Both are editors for the L-functions and modular forms database, and were intrigued by the benefits of an index that could bring different mathematical communities together based on a common interest in data. Together with the other MathBases editors, the authors initiated the metamorphosis of MathDB into MathBases at an AIM workshop[9] in December 2023. Since then, efforts to expand the index and add resources for database creators have been ongoing.

4 Invitation to Contribute

The quality and currency of the MathBases index depend on community participation.

The current scope guidelines, collected at mathbases.org/policies, are broad but not unlimited. MathBases focuses on publicly available collections of mathematical objects, including both large collaborative platforms and small specialist datasets, while purely expository resources or formal libraries lie outside its

[8] data.mathhub.info
[9] aimath.org/pastworkshops/cyberinfrastructure.html

scope. The site also welcomes articles and templates for database creation, ideas for future databases, and announcements of relevant conferences.

Detailed instructions for contributing are available at mathbases.org/contributing. Adding a database entry is intended to be straightforward: a contributor proposes a file following the standard template and submits it as a pull request to the public repository github.com/MathBases/MathBases. Contributors who do not want to work directly with GitHub can instead open an issue or reach out through mathbases.org/contact. Each proposed entry is reviewed by the editorial board before publication. In addition to improving the website, the extent of such contributions also provides an objective measure of community engagement, since it is correlated with overall reach.

Discussion about entries, metadata standards, and the development of the platform takes place on the MathBases Zulip stream.

5 Questions for the Wider Community

A shared index is a prerequisite for everything else: avoiding duplication, identifying genuine gaps, and building the shared standards and interoperability that initiatives like FAIR [24] and MaRDI [18] call for. Before the community can coordinate, it needs to see what it already has. With this map in place, several unresolved questions come into sharper focus.

We close with four questions that seem especially pressing.

- What metadata are most relevant for collections of mathematical objects? There is currently no common standard, and many databases record too little information to be reliably found, cited, or reused in alignment with broader initiatives like FAIR [24] and MaRDI [18].
- What should it mean for a mathematical database to be *correct*? Mathematical data carry a notion of correctness that is stronger than in many other scientific settings. How should projects test, certify, and communicate the reliability of their data?
- How should research credit be assigned for the creation and maintenance of mathematical databases? The current academic incentive structure rarely rewards the long-term editorial, computational, and technical work on which these resources depend.
- Which collections are missing? A better picture of the current landscape should make it easier to identify genuine gaps in coverage rather than simply rediscovering projects that already exist.

We hope that MathBases can serve both as a starting point for data discovery and as a common point of reference for the discussions above.

Acknowledgments. Berčič was supported by the Air Force Office of Scientific Research (award number FA9550-21-1-0024), and by the Slovenian Research and Innovation Agency (program no. P1-0294). Berčič and Roe were supported by the AI For Math Fund, a program of Renaissance Philanthropy. Costa and Roe were supported by the Simons Foundation: Costa by SFI-MPS-Infrastructure-00008651 (AS), Roe by SFI-MPS-SSRFA-00005427 and both by 550033. MathBases was initiated at the American Institute of Mathematics workshop *Open-source cyberinfrastructure supporting mathematics research* (December 2023); we thank AIM and the organizers, Robert Beezer, Steven Clontz, and David Lowry-Duda. The authors also acknowledge the other editors of MathBases: Benjamin Hutz, David Lowry-Duda, Ben Spitz, and Adam Towsley.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aaronson, S., Complexity Zoo contributors: Complexity Zoo. `https://complexityzoo.net`
2. Berčič, K.: Data and Data Quality in Mathematics. In: Data Quality Matters - Best Practices for Integrity and Assurance. IntechOpen (Jan 2026). `https://doi.org/10.5772/intechopen.1013831`
3. Besche, H.U., Eick, B., O'Brien, E.A., Horn, M.: SmallGrp, the GAP Small Groups Library, Version 1.5.4. `https://gap-packages.github.io/smallgrp` (Jul 2024), gAP package
4. Black, P.E.: Dictionary of Algorithms and Data Structures. `https://xlinux.nist.gov/dads/`
5. Boege, T., Fritze, R., Görgen, C., Hanselman, J., Iglezakis, D., Kastner, L., Koprucki, T., Krause, T.H., Lehrenfeld, C., Polla, S., Reidelbach, M., Riedel, C., Saak, J., Schembera, B., Tabelow, K., Weber, M.: Research-data management planning in the German mathematical community. European Mathematical Society Magazine (130), 40–47 (Sep 2023). `https://doi.org/10.4171/mag/152`
6. Bonte, S., Devillez, G., Dusollier, V., Mélot, H.: PHOEG: an online tool for discovery and education in extremal graph theory. arXiv preprint arXiv:2603.27242 (2026)
7. Campbell-Kelly, M., Croarken, M., Flood, R., Robson, E.: The History of Mathematical Tables. Oxford University Press (Oct 2003). `https://doi.org/10.1093/acprof:oso/9780198508410.001.0001`
8. Cantor's Attic contributors: Cantor's Attic. `https://neugierde.github.io/cantors-attic/`
9. Carter, N., Ellis, R.: Group Explorer. `https://nathancarter.github.io/group-explorer/` (2019)
10. Coolsaet, K., D'hondt, S., Goedgebeur, J.: House of Graphs 2.0: A database of interesting graphs and more. Discrete Applied Mathematics **325**, 97–107 (2023). `https://doi.org/10.1016/j.dam.2022.10.013`
11. COS++. The Combinatorial Object Server. `https://combos.org` (2021)
12. Costa, E., Roe, D.: Zen and the art of database maintenance. In: Balakrishnan, J.S., Elkies, N., Hassett, B., Poonen, B., Sutherland, A.V., Voight, J. (eds.) Arithmetic Geometry, Number Theory, and Computation. pp. 277–292. Springer International Publishing, Cham (2021)
13. Devillez, G., Hauweele, P., Mélot, H.: PHOEG Helps to Obtain Extremal Graphs. In: Fortz, B., Labbé, M. (eds.) Operations Research Proceedings 2018, pp. 251–257. Springer International Publishing, Cham (2019). `https://doi.org/10.1007/978-3-030-18500-8_32`

14. European Parliament and Council of the European Union: Directive 96/9/EC on the legal protection of databases. https://eur-lex.europa.eu/EN/legal-content/summary/legal-protection-databases.html (1996)
15. Heidorn, P.B.: Shedding Light on the Dark Data in the Long Tail of Science. Library Trends **57**(2), 280–299 (2008)
16. Jordan, S.P.: Quantum Algorithm Zoo. https://quantumalgorithmzoo.org
17. LMFDB Collaboration, T.: The L-functions and modular forms database. https://www.lmfdb.org (2026), [Online; accessed 9 April 2026]
18. MaRDI Consortium: Mathematical research data initiative. https://www.mardi4nfdi.de, last accessed 2026/03/24
19. OEIS Foundation Inc.: The On-Line Encyclopedia of Integer Sequences. https://oeis.org (2026)
20. Olver, F.W.J., Olde Daalhuis, A.B., Lozier, D.W., Schneider, B.I., Boisvert, R.F., Clark, C.W., Miller, B.R., Saunders, B.V., Cohl, H.S., McClain, M.A.: NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov (2026), "F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds."
21. Scharein, R.: KnotPlot. https://knotplot.com
22. The π-Base Community: π-Base, a community database of topological counterexamples. https://topology.pi-base.org (2014)
23. United States Copyright Office: Report on legal protection for databases. https://www.copyright.gov/reports/appendix.pdf (1997)
24. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**(1), 160018 (Mar 2016). https://doi.org/10.1038/sdata.2016.18