Research-data management planning in the German mathematical community

Tobias Boege, René Fritze, Christiane Görgen, Jeroen Hanselman, Dorothea Iglezakis, Lars Kastner, Thomas Koprucki, Tabea H. Krause, Christoph Lehrenfeld, Silvia Polla, Marco Reidelbach, Christian Riedel, Jens Saak, Björn Schembera, Karsten Tabelow, Marcus Weber

In this paper we discuss the notion of research data for the field of mathematics and report on the status quo of research-data management and planning. A number of decentralized approaches are presented and compared to needs and challenges faced in three use cases from different mathematical subdisciplines. We highlight the importance of tailoring research-data management plans to mathematicians' research processes and discuss their usage all along the data life cycle.

1 Introduction

Scientific progress heavily relies on the reusability of previous results. This in turn is closely linked to reliability and reproducibility of research, and to the question whether another researcher would arrive at the same result with the same material. In mathematics proofs, together with references to definitions of mathematical objects and already verified theorems, traditionally contained all the information needed in order to verify results. However, the advent of computers has opened up new resources previously deemed impossible, while increasing the need for well-adapted research-data management (RDM). For example, algorithms are now implemented to arrive at new conclusions. The size of examples has exploded several orders in magnitude. And some proofs have become too complicated for even the brightest minds, such that software is consulted for thorough understanding and verification.[1] Studies [12, 13, 18, 19] from various fields of applied mathematics show that nowadays many results cannot be easily reproduced and hence verified.

As we outline in Section 2, there are research data in all subdisciplines of mathematics that need responsible organization and documentation in order to ensure they are handled according to the FAIR principles [25] for sustainable, reproducible, and reusable research. One way to achieve this is via a tailored research-data management plan (RDMP), describing the data life cycle over the course of a project [17] and providing guidance to fulfill funding requirements.[2] In mathematics, it is particularly important to treat the RDMP as a living document [4] because the mathematical research process is hardly projectable and does usually not follow a standardized collection–analysis–report procedure. In subfields with experience in using such documentation, three-fold reports – at the grant-application stage, as a working document, and as a final report – have proven useful. We discuss this in Sections 3 and 4, spotlighting examples from different subfields, and conclude this article by listing central topics for RDMPs in all areas of mathematics.

2 Mathematical research data

Following [9, p. 130], we define *research data* as all digital and analog objects that are generated or handled in the process of doing research.[3] In mathematics, research data thus include paper publications and proofs therein as well as computational results, code, software, and libraries of classifications of mathematical objects. A non-exhaustive list of possible formats and examples is presented in Table 1 and Section 4. The apparent diversity of mathematical research-data formats is also reflected in other characteristics, such as their storage size, longevity, and state of standardization [3, 8, 10, 22, 24], leading to RDM needs and challenges that are very specific to the discipline of mathematics.

One of the most apparent challenges is the question what metadata are sufficient for reusability. We will answer this question partially for the mathematical subfields presented in Section 4.

[1] See, e.g., the story outlined in https://xenaproject.wordpress.com/2020/12/05/liquid-tensor-experiment, solved with the lean project https://github.com/leanprover-community/lean-liquid.

[2] E.g., at the European level https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf, and at the German national level https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf.

[3] This is in line with the notions employed by the DFG https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/index.html, forschungsdaten.info https://www.forschungsdaten.info/themen/informieren-und-planen/was-sind-forschungsdaten, and the MPG https://rdm.mpdl.mpg.de/introduction/research-data-management, e.g.

Research-data type	Examples of data formats
Mathematical documents	PDF, LaTeX, XML, MathML
Literate programming sources	Maple Worksheets, Jupyter/Mathematica/Pluto Notebooks
Domain-specific research software packages and libraries	R for statistics, Octave, NumPy/SciPy or Julia for matrix computations, CPLEX, Gurobi, Mosel and SCIP for integer programming, or DUNE, deal.II and Trilinos for numerical simulation
Computer-algebra systems	SageMath, SINGULAR, Macaulay2, GAP, polymake, Pari/GP, Linbox, OSCAR, and their embedded data collections
Programs and scripts	written in the packages and systems above, in systems not developed within the mathematical community, input data for these systems (algorithmic parameters, meshes, mathematical objects stored in some collection, the definition of a deep neural network as a graph in machine learning)
Experimental and simulation data	usually series of states of representative snapshots of an observed system, discretized fields, more generally very large but structured datasets as simulation output or experimental output (simulation input and validation), stored in established data formats (i.e., HDF5) or in domain-specific formats, e.g., CT scans in neuroscience, material science or hydrology
Formalized mathematics	Coq, HOL, Isabelle, Lean, Mizar, NASA PVS library
Collections of mathematical objects	L-Functions and Modular Forms Database (LMFDB), Online Encyclopedia of Integer Sequences (OEIS), Class Group Database, ATLAS of Finite Group Representations, Manifold Atlas, GAP Small Groups Library
Descriptions of mathematical models in mathematical modeling languages	Modelica for component-oriented modeling of complex systems, Systems Biology Markup Language (SBML) for computational models of biological processes, SPICE for modeling of electronic circuits and devices, and AIMMS or LINGO as a modeling language for integer programming

Table 1. Mathematical research data come in a variety of data formats. Updated table based on [24, pp. 26–27].

However, as the authors of [3] note, 'the meaning and provenance of [mathematical research] data must usually be given in the form of complex mathematical data themselves.' It is thus not surprising that there is no common, standardized metadata format yet. A search in the RDA Metadata Standards Catalog[4] at the time of writing reveals five hits, four from a subfield of statistics and one from economics, none of which could encode information about, say, a computer-algebra experiment. This lack of standardization is in contrast to other disciplines such as the life sciences, where the OBO Foundry[5] hosts more than one hundred interoperable ontologies to describe and link research results, including common naming conventions [1, 20].

Another important aspect of mathematical research data is its particular data life cycle. Again in contrast, for instance, to the life sciences, where older results can be overruled by new evidence, mathematical results that have been proven true remain true indefinitely. Since they cater for other disciplines such as the physical, social, health or life sciences [24, Fig. 1 and discussion], mathematics has a particular responsibility to science to preserve their results in a sustainable manner. We discuss this aspect and how mathematics can be embedded in interdisciplinary research pipelines in more detail in Sections 4.1 and 4.2. Section 4.3 stresses the role thorough documentation plays in this context, using classifications as an example.

3 Status quo of RDMPs in mathematics

In the narrower sense of data (rather than research data), it is a common claim in the community at the time of writing that mathematics rarely produces data[6] and that the few data available need no particular management.[7] This is often based on an interpretation of data being something computational, and mathematics

[4] https://rdamsc.bath.ac.uk/subject/Mathematics%20and%20statistics

[5] https://obofoundry.org/principles/fp-000-summary.html

[6] Usually, only statistics is mentioned as a data-producing subdiscipline, see, e.g., https://wissenschaftliche-integritaet.de/kommentare/software-entwicklung-und-umgang-mit-forschungsdaten-in-der-mathematik.

[7] See, e.g., the unofficial document https://www.math.harvard.edu/media/DataManagement.pdf.

being a discipline which is very much paper rather than computer based. Anecdotal evidence suggests that this view is widely established, that there is little knowledge about general RDM, that existing local facilities are hardly used, and that RDMPs are not a standard tool at any stage of the research process.[8] The proposal [24] has identified the need to build common infrastructures for all subdisciplines of mathematics, and mathematics-specific DFG guidelines for FAIR research data will be developed in the foreseeable future.[9]

Now, the question of what these guidelines should be is not trivial. A large number of questions from a general RDMP catalogue,[10] are irrelevant for a community which produces foremostly theoretical results. For instance, for mathematicians the cost of producing data is rarely relevant – unlike, e.g., in the life sciences where data might have to be collected in the field. In the same vein, ethical or data-protection questions most often do not play a role, save for, for instance, industry collaborations or studies conducted in didactics. Large parts of the community have little training in legal aspects as, for example, formulae cannot be assigned proprietary rights. In order to avoid the impression that thus all general RDMP questions apply only to sciences different than mathematics, it is imperative to design bespoke catalogues of questions. These should (a) use unambiguous language, for instance, using the term 'research data' rather than the more specific 'data' which many mathematicians do not handle in their research, and (b) avoid superfluous topics while at the same time including sufficient detail, for instance, for mathematics' metadata and preservation needs identified in the previous section. Now, rather than endeavoring to find a one-size-fits-all solution, in the subsequent section we identify important RDM questions for a number of use cases which are known to the authors – focusing on metadata, software, data formats and size, versioning, and storage – and provide those with what we consider to be sensible answers. We use the remainder of this section to report on two RDM solutions implemented in DFG-funded Collaborative Research Centers (CRC).

The CRC 1456 'Mathematics of Experiment' includes 17 scientific projects in applied mathematics, computer science, and natural sciences such as biophysics and astronomy, aiming to improve the analysis of experimental data. The research data here are extremely diverse (e.g., mathematical documents, notebooks, programs, simulation data or experimental measurements) and their handling is supported by the CRC's dedicated infrastructure project. In regular RDM meetings, four themes are recurrent. First, reusage scenarios: especially in interdisciplinary research the same datasets may be processed or used by different groups; documentation, curation, and publication should be tailored to those groups' needs. Second, reproducibility, both computationally and practically in data recreation. Third, metadata: finding accurate descriptors to help the user understand cross-scientific research data. And fourth, visibility: receiving recognition for stand-alone research data beyond a journal publication is hard. This last topic is usually not part of a standard set of RDMP questions but aims to provide an incentive to increase the effort in research-data creation, publication, and curation.

The CRC 1294 'Data Assimilation' includes 15 interdisciplinary research projects focusing on the development and integration of algorithms, e.g., in earthquake prediction, medication dosing, or cell-shape dynamics. Researchers are thus confronted both with diverse research data and varying cultural data-handling habits. A central project supports their RDM, and IT infrastructure to facilitate collaborative work and knowledge perpetuation to advance good scientific practice is provided. In particular, the CRC designed an RDMP template in collaboration with the University of Potsdam's research-data group. This covers policies and guidelines, legal and ethical considerations, documentation, and dataset-specific aspects. A vital component of the training is then the classification of the digital objects that are reused and created by the individual researchers. This helps them to develop tailored strategies to improve the quality and reproducibility of published results and to sensitize their research-data handling throughout the data life cycle.

4 Use cases

We now consider four very different mathematical use cases and discuss their particular research-data needs. These use cases have been identified by the different mathematical subfields in [24] as particularly representative for the research community. Central in these expositions for us is to find out how, using RDMPs, we can provide the best, case-specific guidance to make a project reusable.

4.1 Applied and interdisciplinary mathematics

In numerous scientific fields real-world problems are simplified, e.g., to experiments, and subsequently described in abstract ways using mathematical models. If a model is combined with input data, it forms a concrete instance of such a problem. With the help of algorithms, the input data are then transformed into output data. Following validations, the interpretation of outputs provides the solution of the initial problem in a so-called Modeling–Simulation–Optimization workflow [24, p. 77]. For complete RDM,

[8] In fact, RDMPs became compulsory in DFG-funding applications only in March 2022, and there are no statistics available on how many mathematics proposals included such a document. See also https://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_25/index.html.

[9] https://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_25/index.html

[10] For instance, the current questionnaire supplied by the DFG-funded research-data management organiser RDMO https://github.com/rdmorganiser/rdmo-catalog/releases/tag/1.1.0-rdmo-1.6.0.

such workflows should be documented in detail as part of an RDMP.

A standard RDMP questionnaire includes some guidance for the documentation of workflows, such as the main research question, involved disciplines, tools, software, technologies, processes, research-data aspects, and reproducibility. Using this as a template, a tailored questionnaire is currently being developed within the framework of MaRDI[11] to document workflows in detail. This is divided into four sections dealing with the problem statement (object of research, data streams), the model (discretization, variables), the process information (process steps, applied methods), and reproducibility. It is aimed at all disciplines and differs only slightly in whether a theoretical or experimental workflow is documented. The central element of the questionnaire is to establish connections between different steps of the research process in order to improve interoperability of research data. The description of an individual process step, for example, requires the assignment of the relevant input and output data, the method and the (software) environment. At the same time, the documentation of the methods, software, input and output data requires persistent identifiers (e.g., Wikidata, swMATH, DOI) in addition to topic-dependent information.

We consider the documentation of a concrete workflow combining archaeology and mathematics as an example. This is based on [11], was created by Margarita Kostre independently afterwards, and described in personal communication as 'very helpful for the reflection of the own work.' The author commented that she will use workflow documentation in the future again, as she believes it facilitates interdisciplinary communication, e.g., about the status of a project, its goal, and data transfer, it provides better clarity in larger collaborations and allows colleagues to enter a project more easily. The aim of this work is to understand the Romanization of Northern Africa using a susceptible infectious epidemic model. On the process level, the workflow starts with data preparation, e.g., collecting, discretizing, and reducing archaeological data. Once a suitable epidemics model is found, the inverse problem is solved to determine contact networks and spreading-rate functions. Subsequent analysis allows the identification of three different possibilities of the Romanization of Northern Africa. The detailed documentation can be found on the MaRDI Portal.[12]

4.2 Scientific computing

While research in pure mathematics strives to determine an ultimate truth, applied or computational mathematics in majority need to deal with approximations to reality: models are usually expressed in terms of real or complex numbers and only finite subsets of these can actually be implemented on computer hardware. Consequently, the result of a computation depends on the format of the finite-precision numbers used and on the specific hardware executing the computations, making a detailed documentation of the computer-based experiment crucial and reusability of code a must-have [5]. Thus, the input data and results of a computer experiment and also the precise implementation (code, software, and hardware) of the algorithms used constitute important research data.

Absence of such details in documentation makes applied mathematics face the same reproducibility issues (e.g., [2]) as other scientific fields. Still mathematical algorithms make up the foundation of many computational experiments, for instance as solvers for linear systems of equations, eigenvalue problems, or optimization problems, and are thus at the heart of science today. This responsibility calls for rigorous RDM and documentation in RDMPs.

The main difficulty in establishing RDMPs in scientific computing seems to be in creating incentives to adhere to common standards. In case of a single multi-author paper within a larger project cluster, there are two levels to this question: the funding context and local RDM. Regarding the first, incentives should clearly address reporting requirements and incorporate rewards for sustainable RDM, rather than merely counting publications and citations, to ensure the cluster can *stand on the shoulders of giants* instead of *building on quicksand*. The beneficiaries here are other researchers in the project and world-wide. Consequently, global RDM needs to answer what is reported where and why. Regarding the local context, for the collaborative work of the authors, incentives are far more evident. Thorough RDM, documented in a living RDMP, not only accelerates the paper writing, it also improves the reusability of information for future endeavors of the individual authors. Questions center around 'When is the code/data provided? Where in the (local) infrastructure is it stored? By whom? Who is processing it next?'

Consequently, RDMPs should be modularized to enable the single modules to change at their appropriate pace. While the global management rules of a project cluster may not change at all, or at best very slowly, the findings in a single work package may alter the RDMP and thus RDM needs high agility to react to changes. For the software pipeline of an example paper that means: a task-based RDMP, updated as the pipeline evolves, needs to fulfill the requirements of [5], while for the project cluster sustainable handover, following (e.g., [6]), needs to be addressed in the overarching RDMP.

4.3 Computer algebra and theoretical statistics

Large parts of the German mathematical community consider themselves as not doing applied work. This includes fields such as geometry, topology, algebra, analysis or number theory, and also mathematical statistics, for instance. However, these researchers increasingly use computers, too, to explore the viability of proof strategies, test their own conjectures or refute established ones.

[11] https://www.mardi4nfdi.de

[12] https://portal.mardi4nfdi.de/wiki/Romanization_spreading_on_historical_interregional_networks_in_Northern_Tunisia

As a consequence, *classifications*, the systematic and complete tabulation of all objects with a given property, grow wildly in size and complexity. They give a complete picture for some aspect of a theory and may be used in many ways, from the search of (counter)examples over building blocks for constructive proofs to benchmark problems.

For instance, the L-Functions and Modular Forms Database (LMFDB) [23] contains over 4.8 TB of data relating objects conjectured to have strong connections by the Langlands program: number fields, elliptic curves, modular forms, L-functions, Galois representations. It includes tens of millions of individual objects and stores the relations between these. Entries contain detailed information on reliability, completeness, and several versions of the code needed to compute them. The database has a public reporting system which allows all users to have visibility of any issues or errors.[13]

Computing mathematical objects for classification can often be algorithmically hard and time-consuming. But once computed, results are final and independent of the software used. With larger computer clusters and better algorithms, it is unreasonable and unsustainable to expect researchers who want to build on existing research to repeat individual computations. This expected reuse increases the need for responsible RDM and triggers challenges which need to be addressed in an RDMP. In particular, four themes are central in this regard. First, how can researchers ensure that their research data are correct and complete? Is the connection of mathematical theory and code sound? Second, how can other researchers access, understand, and reuse the research data? Third, how can one ensure longevity of their research data? And fourth, how can researchers report errors/corrections and upload new versions of research data if necessary?

These questions are neatly addressed in the LMFDB mentioned above. To show how things can go wrong without proper RDM, we discuss a classification of all conditional independence structures on up to four discrete random variables, originally published in a series of papers [14–16]. Of the $2^{24} = 16\,777\,216$ *a priori* possible patterns of how four random variables can influence each other, only 18 478 ($\approx 0.11\%$) are realizable with a probability distribution. Šimeček, the author of [21], digitized this result and then left the field after his PhD in 2007. His research data was deleted in 2021 from his former institute's website – the only public place which ever held the database.[14] It was encoded in a packed binary format which is hard to read, search, and reuse. Some files supporting the correctness of the classification for binary distributions use an unspecified, compiler-specific binary serialization format for floating-point data.[15] The programs used for the creation and inspection of the database were written in a dialect of the Pascal programming language, which has not been maintained since 2006. The sparse documentation is in Czech.

This situation can only be fixed by recreating the database from scratch, including proofs. An RDMP for this project should emphasize the need to list and document each step of redoing the computations, the use of standard data formats with rich metadata for interoperability and searchability of the database, and ensure future reusability of Šimeček's results.

5 Discussion and outlook

The problem of reusability strongly relates to a phenomenon called *dark data*, which 'exists only in the bottom left-hand desk drawer of scientists on some media that are quickly aging' [7]. If research data are not available, they are of course neither traceable nor reusable or FAIR. This phenomenon extends from lost USB sticks and conflicting cloud-based collaboration tools like Dropbox[16]nd Overleaf[17]ithout local backup to papers containing very condensed complicated proofs that can only be taken up in future work if access to handwritten notes of the authors is also possible. A prime example of this is presented in Section 4.3 where unavailable research data is in stark contrast to the everlasting truth of mathematical results. RDMPs are a tool of choice against such issues, serving as a basic measure to organize the full data life cycle.

From the three case studies considered in Section 4, we derive that RDMPs in mathematics in particular (a) stimulate reflection, clarity, and interdisciplinary communication, (b) require flexibility and modularization as living RDMPs, and (c) facilitate the documentation of iterative computational processes by fostering research-data interoperability and reusability.

We further conclude that archiving and preservation is key in any mathematical subdiscipline. As a very first step to improve the status quo, all research results necessary for reusability (data, code, notes, …) should be stored in a sustainable and findable manner, using resources already documented in an RDMP before a project starts. Ideally, in a second step, citable repositories with persistent identifiers for these research data can be chosen, and, in a third step, these can be annotated with interlinked metadata, implemented via knowledge graphs. Because of the diversity of mathematical research data, the choice of metadata should be made carefully with possible reusage scenarios and interest groups in mind, also documented in an RDMP. If code is part of a publication, thoughts should be given to the detail of documentation

[13] Other classification databases targeted at specific audiences are listed at https://mathdb.mathhub.info.

[14] A backup is still available on the Internet Archive at http://web.archive.org/web/20190516145904/http://atrey.karlin.mff.cuni.cz/~simecek/skola/models/.

[15] A set of scripts for reading these files is available at https://github.com/taboege/simecek-tools.

[16] https://www.dropbox.com

[17] https://www.overleaf.com

and again appropriate citable long-term repositories. In addition, an RDMP should be used as a tool to identify legal constraints, like the compatibility of software licenses, before any actual work is conducted.

Acknowledgements. The authors are grateful to Margarita Kostre for retrospectively compiling an RDMP for her project [11] and to Tim Hasler for background and discussion regarding the MATH+ research-data management organizer.

Funding. René Fritze, Christiane Görgen, Jeroen Hanselman, Lars Kastner, Thomas Koprucki, Tabea Krause, Marco Reidelbach, Jens Saak, Björn Schembera, Karsten Tabelow and Marcus Weber are at the time of writing supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 'MaRDI – Mathematische Forschungsdaten-initiative.' Christian Riedel is supported by the DFG, project-ID 318763901 – SFB1294. Christoph Lehrenfeld is supported by the DFG, project-ID 432680300 – SFB1456.

The authors have no competing interests to declare.

All authors made significant contributions to the design of this review as well as drafting and revising the manuscript. All have approved this final version, agreed to be accountable and have approved of the inclusion of those in the list of authors.

References

[1] R. Arp, B. Smith and A. D. Spear, *Building Ontologies with Basic Formal Ontology*. MIT Press, Cambridge, MA (2015)

[2] W. Bangerth and T. Heister, Quo vadis, scientific software? *SIAM News* **47**, 8–7 (2014)

[3] K. Berčič, M. Kohlhase and F. Rabe, (Deep) FAIR mathematics. *it – Information Technology* **62**, 7–17 (2020)

[4] J. Dierkes, 4.1 Planung, Beschreibung und Dokumentation von Forschungsdaten. In *Praxishandbuch Forschungsdatenmanagement*, pp. 303–325, De Gruyter Saur, Berlin, Boston (2021)

[5] J. Fehr, J. Heiland, C. Himpe and J. Saak, Best practices for replicability, reproducibility and reusability of computer-based experiments exemplified by model reduction software. *AIMS Mathematics* **1**, 261–281 (2016)

[6] J. Fehr, C. Himpe, S. Rave and J. Saak, Sustainable research software hand-over. *Journal of Open Research Software* **9**, article no. 5 (2021)

[7] P. B. Heidorn, Shedding light on the dark data in the long tail of science. *Library Trends* **57**, 280–299 (2008)

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

[9] M. Kindling and P. Schirmbacher, „Die digitale Forschungswelt" als Gegenstand der Forschung / Research on digital research / Recherche dans la domaine de la recherche numérique. *Information – Wissenschaft & Praxis* **64**, 127–136 (2013)

[10] T. Koprucki, K. Tabelow and I. Kleinod, Mathematical research data. *PAMM. Proc. Appl. Math. Mech.* **16**, 959–960 (2016)

[11] M. Kostre, V. Sunkara, C. Schütte and N. D. Conrad, Understanding the romanization spreading on historical interregional networks in Northern Tunisia. *Applied Network Science* **7**, article no. 53 (2022)

[12] M. S. Krafczyk, A. Shi, A. Bhaskar, D. Marinov and V. Stodden, Learning from reproducing computational results: introducing three principles and the *Reproduction Package*. *Philos. Trans. Roy. Soc. A* **379**, article no. 20200069 (2021)

[13] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, I. Di Marco, C. Draxl, M. Dułak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, N. A. W. Holzwarth, D. Iuşan, D. B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Küçükbenli, Y. O. Kvashnin, I. L. M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordström, T. Ozaki, L. Paulatto, C. J. Pickard, W. Poelmans, M. I. J. Probert, K. Refson, M. Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunström, A. Tkatchenko, M. Torrent, D. Vanderbilt, M. J. van Setten, V. Van Speybroeck, J. M. Wills, J. R. Yates, G.-X. Zhang and S. Cottenier, Reproducibility in density functional theory calculations of solids. *Science* **351**, article no. aad3000 (2016)

[14] F. Matúš, Conditional independences among four random variables. II. *Combin. Probab. Comput.* **4**, 407–417 (1995)

[15] F. Matúš, Conditional independences among four random variables. III. Final conclusion. *Combin. Probab. Comput.* **8**, 269–276 (1999)

[16] F. Matúš and M. Studený, Conditional independences among four random variables. I. *Combin. Probab. Comput.* **4**, 269–278 (1995)

[17] W. K. Michener, Ten simple rules for creating a good data management plan. *PLOS Computational Biology* **11**, article no. e1004525 (2015)

[18] C. Riedel, H. Geßner, A. Seegebrecht, S. I. Ayon, S. H. Chowdhury, R. Engbert and U. Lucke, Including data management in research culture increases the reproducibility of scientific results. In *Proceedings of INFORMATIK 2022*, Lecture Notes in Informatik P-326, pp. 1341–1352, Gesellschaft für Informatik, Bonn (2022)

[19] M. Schappals, A. Mecklenfeld, L. Kröger, V. Botan, A. Köster, S. Stephan, E. J. García, G. Rutkai, G. Raabe, P. Klein, K. Leonhard, C. W. Glass, J. Lenhard, J. Vrabec and H. Hasse, Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom. *J. Chem. Theory Comput.* **13**, 4270–4280 (2017)

[20] D. Schober, B. Smith, S. E. Lewis, W. Kusnierczyk, J. Lomax, C. Mungall, C. F. Taylor, P. Rocca-Serra and S.-A. Sansone, Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics* **10**, article no. 125 (2009)

[21] P. Šimeček, A short note on discrete representability of independence models. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models*, pp. 287–292, Action M Agency, Prague (2006)

[22] O. Teschke, Some heuristics about the ecosystem of mathematics research data. *PAMM. Proc. Appl. Math. Mech.* **16**, 963–964 (2016)

[23] The LMFDB Collaboration, The L-functions and modular forms database. http://www.lmfdb.org (2022)

[24] The MaRDI consortium, *MaRDI: Mathematical Research Data Initiative proposal*. Zenodo (2022)

[25] M. Wilkinson, M. Dumontier, IJ. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, article no. 160018 (2016)

Tobias Boege got his PhD from the Otto-von-Guericke-Universität Magdeburg working on conditional independence in algebraic statistics. After a postdoc position at the Max Planck Institute for Mathematics in the Sciences, he is currently at Aalto University.

post@taboege.de, orcid.org/0000-0001-7284-1827

René Fritze received his diploma in mathematics at the University of Münster. After working on various research projects, including MaRDI, he has joined the Digital Technology group at Arup as a Senior Software Engineer.

rene.fritze@wwu.de, orcid.org/0000-0002-9548-2238

Christiane Görgen (corresponding author) holds a PhD in statistics from Warwick University and has been doing research in algebraic statistics at the Max Planck Institute for Mathematics in the Sciences. Since 2021 she works at the University of Leipzig as MaRDI's mathematical data consultant.

goergen@math.uni-leipzig.de, orcid.org/0000-0002-6476-956X

Jeroen Hanselman did his PhD in mathematics at Ulm University and is currently working as a postdoc at the RPTU Kaiserslautern-Landau concerning himself with improving the software peer reviewing process for MaRDI. His area of research is computational arithmetic geometry, with a focus on Jacobians of curves.

hanselman@mathematik.uni-kl.de, orcid.org/0000-0002-1298-0961

Dorothea Iglezakis holds a diploma in psychology and a PhD in Computer Science. She is head of the research-data management team of the University of Stuttgart. Dorothea is mainly interested in (semantically enriched) metadata, the automation of research-data management processes and the interlinking of different research outputs, actors, and concepts in a global knowledge graph.

dorothea.iglezakis@ub.uni-stuttgart.de, orcid.org/0000-0002-8524-0569

Lars Kastner holds a PhD in mathematics from Freie Universität Berlin. His main research area lies at the intersection of algebraic geometry and combinatorics, with a focus on computational aspects. In 2022, he joined the MaRDI task area on computer algebra at TU Berlin.

kastner@math.tu-berlin.de, orcid.org/0000-0001-9224-7761

Thomas Koprucki holds a Diploma degree in physics and a PhD degree in mathematics. He works in the field of mathematical modeling and numerical simulation in nano- and opto-electronics at WIAS Berlin.

thomas.koprucki@wias-berlin.de, orcid.org/0000-0001-6235-9412

Tabea H. Krause holds a degree in mathematics and logic from the University of Leipzig. Since 2022 she works as MaRDI's consortia contact at Leipzig University.

tabea.krause@math.uni-leipzig.de, orcid.org/0000-0001-7275-5830

Christoph Lehrenfeld holds a PhD in mathematics from RWTH Aachen University, and his research focuses on numerical methods for partial differential equations. He has been a professor at the Georg-August-Universität Göttingen since 2016.

lehrenfeld@math.uni-goettingen.de, orcid.org/0000-0003-0170-8468

Silvia Polla (PhD in archaeology, University of Siena) works since 2021 as a research data steward and library manager at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS).

silvia.polla@wias-berlin.de, orcid.org/0000-0002-2395-2448

Marco Reidelbach holds a PhD in bioinformatics from Freie Universität Berlin and works in the field of molecular modeling and simulation. Since 2021 he works for MaRDI at Zuse Institute Berlin focusing on mathematics in an interdisciplinary context.

reidelbach@zib.de, orcid.org/0000-0002-1919-1834

Christian Riedel graduated in geoinformation and obtained a PhD in planetary sciences. He currently works at Potsdam University in the research-data management of an interdisciplinary Collaborative Research Center on mathematical data assimilation. His work involves research on data and software-based procedures in geosciences and the sustainable provision of interdisciplinary research data.

christian.riedel@uni-potsdam.de, orcid.org/0000-0001-5154-4153

Jens Saak holds a PhD in applied mathematics from TU Chemnitz. Since 2010, he has been a team leader at the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg. His research covers various aspects of industrial and applied mathematics, with an increasing focus on research-software engineering and research-data management.

saak@mpi-magdeburg.mpg.de, orcid.org/0000-0001-5567-9637

Björn Schembera holds a diploma degree in computer science and a PhD in engineering. His research interests include dark data, semantic technology and research-data management and he currently works as a knowledge engineer in MaRDI at the University of Stuttgart.

bjoern.schembera@mathematik.uni-stuttgart.de, orcid.org/0000-0003-2860-6621

Karsten Tabelow holds a PhD in physics. He works in the field of medical image analysis with a focus on neuroimaging, quantitative imaging and statistical methods at WIAS Berlin. Since 2016 he has been also working on mathematical research data and the concepts behind MaRDI, the Mathematical Research Data Initiative.

karsten.tabelow@wias-berlin.de, orcid.org/0000-0003-1274-9951

Marcus Weber holds a PhD in mathematics and did a habilitation at FU Berlin. He is head of the research group 'Computational Molecular Design' at Zuse Institute Berlin. Marcus is mainly interested in molecular simulation and in the theory of Markov processes.

weber@zib.de, orcid.org/0000-0003-3939-410X