Research Data Management Planning in Mathematics

Whitepaper by MaRDI, the Mathematical Research Data Initiative

October 6, 2023

Abstract Research data are crucial in mathematics and all scientific disciplines, as they form the foundation for empirical evidence, by enabling the validation and reproducibility of scientific findings. Mathematical research data (MathRD) have become vast and complex, and their interdisciplinary potential and abstract nature make them ubiquitous in various scientific fields. The volume of data and the velocity of its creation are rapidly increasing due to advancements in data science and computing power. This complexity extends to other disciplines, resulting in diverse research data and computational models. Thus, proper handling of research data is crucial both within mathematics and for its manifold connections and exchange with other disciplines.

The *National Research Data Infrastructure (NFDI)*, funded by the federal and state governments of Germany, consists of discipline-oriented consortia, including the *Mathematical Research Data Initiative (MaRDI)*. MaRDI has been established to develop services, guidelines and outreach measures for all aspects of MathRD, and thus support the mathematical research community. Research data management (RDM) should be an integral component of every scientific project, and is becoming a mandatory component of grants with funding bodies such as the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation). At the core of RDM are the FAIR (Findable, Accessible, Interoperable, and Reusable) principles.

This document aims to guide mathematicians and researchers from related disciplines who create RDM plans. It highlights the benefits and opportunities of RDM in mathematics and interdisciplinary studies, showcases examples of diverse MathRD, and suggests technical solutions that meet the requirements of funding agencies with specific examples. The document is regularly updated to reflect the latest developments within the mathematical community represented by MaRDI.

1 Introduction

Digitization has enabled an enormous growth of data availability. As a result, the scientific community, including funding agencies, are increasingly recognizing the value of research data and the necessity of RDM. This emphasizes the importance of making research data readily accessible and preserving them for future use rather than confining them to publications or local storage, ultimately leading to unavailable 'dark data' [24]. This drive stems from the reproducibility crisis [2, 3, 5, 16], and a broader call for open science and open data from different research communities [22]. Ultimately, appropriate research data handling should be considered as part of good scientific practice, see, e.g., [10]. However, questions regarding the types of data to be made available – how and where should it be stored and alongside which metadata – are being raised.

1.1 Open and FAIR data

In a bid to address the issues outlined above, Tim Berners-Lee's 5-star principles[1] and the more recent FAIR (Findable, Accessible, Interoperable, Reusable) principles [28] have been developed to establish guidelines and categorize levels of compliance for research output and research data. Similarly, the complementary FAIR4RS principles [6] have been developed to meet and match the unique demands of research software.

While open data and FAIR principles share the aim of achieving high levels of compliance while allowing for intermediate stages of implementation, they also have distinct differences. Open data and open source software should be freely available to everyone under an appropriate open license. On the other hand, FAIR data does not necessarily need to be open data. The FAIR principles focus more on clarifying how and under what conditions data can be used by people and machines. Furthermore, they introduce specific requirements for metadata descriptions and formats to ensure findability and potential interoperability and reusability. The use of such criteria aids in maintaining an overview of data through all stages of a research project, promoting a high level of FAIRness and a comprehensive project documentation as part of good scientific practice.

1.2 Scope of this white paper

The landscape of research data, including its challenges, opportunities, and benefits is broad and rapidly evolving. This is unsuitable for the adoption of high compliance levels of RDM practices by non-experts. Based on this observation, the NFDI has been instantiated as a community-driven nation-wide effort. Within the NFDI, MaRDI (the Mathematical Research Data Initiative) is responsible for mathematics and its manifold connections to neighboring disciplines [4, 26]. Additionally, funding agencies such as the DFG have introduced both the adoption of FAIR principles and inclusion of RDM concepts and plans as integral components of research projects and funding proposals. This has led to specific requirements by funding bodies at proposal phase that are specifically discussed in Section 4.1.

[1] `https://opendatahandbook.org/glossary/en/terms/five-stars-of-open-data/`

This white paper focuses on the FAIR principles within mathematics and provides researchers with guidance on achieving their goals, by reviewing and suggesting available services, standards, and concrete examples to facilitate effective implementation. It emphasizes the ubiquity of research data and the relevance of RDM across all areas of mathematics, and to a lesser extent, related fields. This white paper also examines the inherent data handling challenges within mathematical subdisciplines and offers suggestions for incorporating them as research data in RDM plans. Additionally, it explores various data handling approaches that arise in interdisciplinary projects, necessitating a unified treatment across mathematical and neighboring scientific disciplines. As supplementary material, the white paper offers illustrative RDM plans applicable to various mathematical subareas, serving as valuable references for project proposals. As such, this white paper complements the DFG's general RDM guidelines [12] and the mathematics specific guidelines [11]. Due to its general context, it is not limited to DFG proposals, but equally applicable for other funding bodies.

1.3 Towards FAIR research data

One of the key goals of good scientific practice is to enable reproducibility of published results without the need to contact the authors of the article. This ideal scenario ensures research quality assurance and facilitates the seamless reuse of existing knowledge, allowing for further advancements that build upon foundations laid by previous researchers.

Achieving **FAIR**ness in research data is a crucial step towards this goal. FAIR data implies that the data are findable, accessible, interoperable, and reusable. **Findability** involves assigning a unique identifier, providing comprehensive metadata descriptions, and indexing the data in a human- and machine-searchable resource to facilitate access by researchers. **Accessibility** ensures that interested researchers can access the data through standardized, open, and free communication protocols, with metadata permanently accessible even if the data themselves are not, or become inaccessible in the course of time. **Interoperability** enables the use of data in different contexts comparable to their original purpose. **Reusability** demands proper documentation using community standards, allowing for reproducibility and enabling other researchers to rely on a dataset for their own applications. Reusability also requires releasing the data under a suitable license. Moreover, to foster compliance with the FAIR principles, the usage of semantic technology (i.e. ontologies and knowledge graphs) is strongly recommended, as it gives meaning to research assets and enables the interlinking of data [13, 19].

The following brief list summarizes important considerations for RDM in mathematics that aim at a high degree of FAIRness. Detailed recommendations are provided in Sections 2 and 3.

1. Data and metadata should be human- and machine-readable, irrespective of the software used to create them, and adhere to community standards when selecting the data format.

2. Software implementations are research data themselves. Open source research software is often preferable. If this is not possible, e.g., for legal reasons, measures should be implemented to address the impact of this ”black box”. In any case, the accompanying documentation should define input and output data formats, allowing for the development of format conversions if needed.

3. Publishing results, papers, software, input data, output data, and references together is crucial, by interlinking them with persistent identifiers.
4. Long-term preservation of research data should rely on dedicated storage services that comply with standards for long-term availability and trustworthiness.[2]

1.4 Benefits of RDM in mathematics

Proper RDM offers numerous benefits for a project beyond adhering to good scientific practice and achieving FAIRness. Prior to project commencement, RDM assists in identifying and establishing a comprehensive strategy, including infrastructure requirements and planned data flow. For an initial review of RDM in mathematics, see [5]. Here, we list some of the advantages of maintaining a living data management plan throughout a project, again referring to later sections for details on the design and implementation.

- Internally documenting any changes in approach and realization during a project is essential, prior to publishing the final state. For instance, in software-based projects, switching data formats, implementations or even entire packages may be necessary and beneficial. Versioning tools are essential here. Another example, not limited to software development, is the maintenance of (project-internal) clear documentation regarding the techniques used, for example, to prove that a certain assertion was tested, but failed for specific reasons. This helps to avoid re-encountering problems that have already been addressed, ensuring smoother and more productive progress. Effective RDM significantly minimizes the effort involved in such transitions and helps to proactively identify potential pitfalls.
- An RDM plan provides additional structure to a project, allowing for independent scrutiny or repetition of specific parts if needed. Examples in software development include test-driven development and unit tests.
- An RDM plan minimizes the impact of on- and off-boarding project members, considering the transient nature of academic careers and the potential span of research agendas across multiple (PhD) researchers. Ensuring seamless continuation of work from departing to incoming project or group members relies heavily on proper data management and corresponding documentation of software, articles, preprints, classification tables, etc. throughout a research project.
- Optionally, publishing the RDM plan serves as a valuable resource and an example for other researchers.

These examples showcase the importance and usefulness of properly structured RDM plans, outweighing the effort required in developing them. We expect this to become more evident in the future, given the increasing relevance of data in science.

[2] see, e.g., `https://www.langzeitarchivierung.de/Webs/nestor/DE/Zertifizierung/nestor_Siegel/siegel.html` and `http://www.oais.info/`

1.5 Example projects

To illustrate the concepts of mathematical research data and its management discussed in this paper, we use two running example projects that we will refer to throughout the document:

Example 1. A project aims to develop a new statistical method for image enhancement and analysis in the application area of neuroimaging and evaluates it using simulations and real data examples.

Example 2. A project wants to compute the Schläfli fan to analyze all possible patterns of lines on tropical cubics.

2 Mathematical Research Data

Research data comprise *any* data that arise in research. All researchers produce data, if only in the form of a LaTeX manuscript. All research data must be equipped with appropriate metadata. In the example of a LaTeX manuscript, the publication is in fact a machine-readable research artifact ready for further processing, e.g., to automatically extract metadata like author information or references. During a project, different types of research data emerge: data being imported, raw and intermediate data, and data that are being ”exported”, meaning either published or kept for further internal use. Each of these data types must be equipped with appropriate metadata.

The aim of this section is to illustrate various kinds of research data in mathematics, see also [4, 26] for a more detailed discussion.

2.1 Data within mathematics

Just like all disciplines, mathematics creates and uses multifaceted research data. Here, we provide guiding examples in different areas of mathematics. In general, MaRDI roughly differentiates among *exact and symbolic*, *numeric*, and *uncertain* data, in addition to rigorous proofs and metadata related to publications. The website `https://www.mardi4nfdi.de` provides more information which we briefly summarize here for reference.

Exact data are data with arbitrary precision that do not have any tolerance for errors. Examples for objects of this data type include vanishing sets of polynomials, lattice points of polytopes, Gröbner bases, exact solutions to (mixed integer) linear programs, etc. These data formats are developed and surveyed as part of the task area “Computer Algebra” of MaRDI [9]. Beyond this, exact data also include symbolic data, or mathematical objects like special integer sequences, mathematical functions or even abstract mathematical models.

Numerical data have finite precision and occur in many applied mathematical problems and related simulations. This data type is often the result of (or input to) a simulation code and ranges from limited size to very large data sets depending on the problem at hand. A consequence of the finite precision is that simulation results often depend on the implementation of the algorithm and the hardware used to run it. Very often, input parameters are also known only with finite precision and thus influence the results. These data types are subject of the task area “Scientific Computing” of MaRDI.

Uncertain data are central to MaRDI’s task area “Statistics and Machine Learning”. In this area of mathematics, randomness is key to numerical experiments. Thus, specific outcomes of statistical methods depend on the utilized random number generator or the seed value for initialization. Different levels of reproducibility of scientific results then depend on their documentation. Furthermore, for machine learning methods, the availability of training data sets is crucial.

Mathematics and MathRD can also be found in many other disciplines. The notion of MathRD in interdisciplinary research requires special considerations and is considered in MaRDI’s task area “Cooperation with other disciplines”. This includes data exchange as well as more general concepts of **mathematical models or workflows**.

MathRD from research projects are a valuable source of knowledge and can help generate new scientific results. For this, the application of the FAIR principles is central to the NFDI in general and within MaRDI in particular. Adding new research data to existing and emerging infrastructures, e.g., ontologies, metadata specifications, repositories or the knowledge graph within MaRDI and the NFDI [26], contributes to the quality of the scientific research cycle. The following non-exhaustive list of mathematical research data provides examples of different data types that can be considered and encountered in mathematical research.

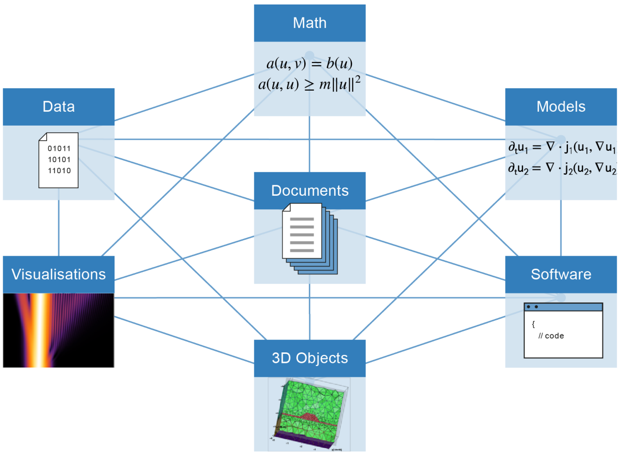

Figure 1 – *The multiverse of mathematical research data [26].*

- **Mathematical documents** in PDF, LaTeX, XML, MathML, etc. represent the most basic data that appear in every project. They encapsulate preprints, books, online documents, and other forms of mathematical literature. Consequently, collections of literature, e.g., in reference lists, constitute a valuable source of knowledge for the topic at hand.

- **Notebooks** e.g., in Jupyter, Pluto, knitr, or Mathematica, live at the intersection between narrative and code, and are an excellent resource to document research in a reproducible way. They usually consist of interchanging code blocks and text blocks that explain the code. Figures can be included. Such notebooks might even contain examples that are too large or complicated to fit within a publication, implementations of algorithms that do not justify a publication by themselves, or even computations that are essential for proofs. However, they usually are not sufficient as-is for FAIRly documenting a computer experiment, see for example [23].

- **Domain-specific research software packages and libraries** comprise for instance R for statistics, Octave, NumPy/SciPy or Julia for scientific computations, CPLEX, Gurobi, Xpress and SCIP for integer programming, or DUNE, deal.II and Trilinos for numerical simulation, and computer algebra systems like SageMath, SINGULAR, Macaulay2, GAP, polymake, Magma, Pari/GP, Linbox, OSCAR, and their embedded data collections. The same holds for **self-written codes**. **Programs and scripts** written in these packages and systems are research data themselves that must be documented, and appropriate metadata must be added, e.g., the version of the package used, and in case of numeric or uncertain data, the specification of the machine the software has been run on.

- **Simulation data** include series of states of representative snapshots of the system, discretized fields, and more generally large but structured data sets as simulation output or experimental output (equals simulation input and validation). They are typically stored in established data formats (e.g. HDF5) or in domain-specific formats, like DICOM or NIfTI for MRI scans in neuroscience or LAZ for LIDAR point cloud data in earth science. Again, documentation and metadata are obligatory.

- **Formalized mathematics** like Coq, HOL, Isabelle, Lean, Mizar etc. constitute the machine-readable counterpart to actual papers: They contain proofs that can actually be verified by software. Proof assistants and computer verifiable proofs have long existed as a niche subject, but with mathematics growing more complicated alongside the acceleration of community output, they play an increasingly important role. A recent example is the liquid tensor experiment project by Johan Commelin and Peter Scholze together with the Lean community [8]. Relevant metadata are very similar to the metadata for domain specific software, and include details regarding the program and version used.

- **Collections of mathematical objects**, e.g. 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, MORwiki are another category. Metadata must link to documentation of the file format and an explanation of how the data was produced.

- **References to other publications** embed a publication in the wider mathematical context. Using tools like bibtex, they are machine readable. Alongside other metadata, they play a crucial role in findability of the publication on portals like arXiv, zbmath, swmath, and mathscinet.

- **Data from other disciplines used in mathematical research** play an important role in applied mathematics. They include a broad range of types like image or auditory data, measurements of material properties or natural phenomena, medical, health, or social data and a lot more. Aspects of these data are discussed in the next section.

2.2 Interdisciplinary aspects

In addition to the identification of research data within mathematics, interdisciplinary aspects play a major role. Other scientific disciplines contribute motivation and test cases for new mathematical methods, ideas, models, algorithms and techniques, and simultaneously inspire novel development within the mathematics community. Vice versa, the applicability in a wide array of disciplines ranging from natural, engineering, and life sciences to humanities due to the tremendous progress in digitization has become one of the strongholds of mathematics. The digitization and standardization of research data within the entire NFDI is also expected to further strengthen this applicability. It is a dedicated goal of MaRDI to harmonize the interfaces between different disciplinary standards for data and metadata, with respect to FAIRness in mathematics.

One use case of interdisciplinary research is the application of mathematical methodology to research data that have been created in some other discipline. Examples include imaging data, survey data, or material data. Another use case is mathematics providing data to other disciplines, e.g., benchmark data for new simulation methods developed in the engineering sciences or training data sets for machine learning algorithms.

More abstract notions of research data come into play, e.g., when the solution of real-world problems require the application of multiple mathematical methods in analysis and simulation pipelines. Then, workflow descriptions and appropriate mathematical models should be considered as mathematical research data themselves.

The following recommendation is then obvious: *When dealing with interdisciplinary research data, follow disciplinary standards of the respective application domain if they exist, and consult MaRDI's publication on the status of their 'mathematicization'.* Prominent examples are the Systems Biology Markup Language (SBML) for computational models of biological processes, SPICE for modeling of electronic circuits and devices, MPS or LP as a modeling formats for integer programming and their modeling counterparts AIMMS and LINGO, CELLML for subcellular processes in biomechanics, ENCYMEML for protein folding, to just name a few.

In cases where established solutions are less obvious, more effort is required. The NFDI is working bottom-up, implying that in the case of an unsupported/unconsidered combination of disciplines, checking (upcoming) white papers by the disciplines is a good first step.

Example 1. Within the project, several types of research data are important: Neuroimage data in NIfTI format. Simulation data with generating R-code including seed for random number generator. Developed algorithm as contribution to the MaRDI knowledge graph. Implemented R-code for the developed method including documentation. Written preprints and papers published as scientific papers.

Example 2. The input data of this project consists of all the regular unimodular triangulations of $3 \cdot \Delta^3$. These come in a text format that is also used by `TOPCOM`[a]. This input data is then processed by `polymake` code, this code is made available as a `polymake` extension. Finally the output, the Schläfli fan, is made available in `polymake`'s own data format whose technical side is JSON. The last datatype is the LaTeX preprint connecting all these parts to the mathematics.

[a] `https://www.wm.uni-bayreuth.de/de/team/rambau_joerg/TOPCOM/index.html`

3 Tools and services for mathematical research data

Tools and services for storing mathematical research data have to meet certain criteria in order for the data produced and accommodated to be considered FAIR. In particular, this concerns findability and accessibility. We briefly elaborate on these aspects and give several hints for selecting a suitable repository for different types of data. The requirements differ by project phase, namely development or archival. Licensing and standardization aspects are also discussed. When selecting repositories, the requirements of the involved institutions and funding agencies should always be taken into account. The following survey is not exhaustive by any means.

3.1 Requirements for FAIR repositories

A repository combines raw storage with metadata information and (persistent) identifiers. Selecting the right repository is crucial in guaranteeing the FAIRness of data. We list some requirements derived from the FAIRness requirements for the data.

- **Findability**: The repository must offer discoverability features and provide meaningful metadata of the datasets it contains through an accessible API and, e.g., web service, both in a machine- and human-readable way. Additionally, it is crucial to assign a uniform resource identifier (URI), such as a fixed URL or a DOI, to each dataset.
- **Accessibility:** The primary purpose of a repository is to make data accessible to interested researchers and their tools. Repositories with dedicated hardware and maintenance staff tend to be able to run more stably and react faster in case of errors, leading to constant uptime and hence being always accessible.
- **Interoperability:** An open API allows for adapting other software to interact with the repository instead of delegating the task to the user. Furthermore, aspects of controlled vocabularies and metadata schemas are crucial here, see below.
- **Reusability:** Since data may contain errors and code may be upgraded for newer versions of underlying software, any repository should have a versioning mechanism in place. This is necessary for reproducibility.

Generic FAIR assessment tools exist [15], but none specific for MathRD. Nevertheless, these tools are helpful to get a first impression of a repository service. To this end, one can use these to evaluate a few datasets. Exhaustive recommendations are provided by the OpenAIRE[3] program funded by the European Commission, the FairSharing program by the University of Oxford[4] and the Database of Research Repositories.[5]. Certification programs for the quality, reliability, trustworthiness, long-term availability and other aspects of a repository include CoreTrustSeal, NESTOR and OAIS.[6]

[3] https://www.openaire.eu/find-trustworthy-data-repository

[4] https://fairsharing.org

[5] https://www.re3data.org/

[6] see https://www.coretrustseal, https://www.langzeitarchivierung.de/Webs/nestor/DE/Zertifizierung/nestor_Siegel/siegel.html and https://www.oais.info/

3.2 Repository examples

Prominent examples of repositories relevant for mathematicians are listed below:

arXiv: The arXiv[7] is a preprint server, funded by Cornell University. It was launched in 1991 and can be reached via the domain arxiv.org since 2001. Upon upload, metadata such as author details, title and abstract, Mathematics Subject Classification (MSC), and keywords are collected. A number of licenses may be chosen from and there is a default minimal license for publishing on the arXiv. Dedicated staff at Cornell University maintains the arXiv servers and provides basic curation of metadata. However, to enable a FAIR reuse, at least a CC-BY, such as CC-BY-SA 4.0[8] license should be chosen. A major advantage of arXiv in contrast to other archives is that the LaTeX source code will be stored. If properly licensed the information is more accessible than PDF documents. For example, the ar5iv [9] project presents arXiv articles optimized for web display.

Zenodo: Zenodo[10] is an open science repository hosted by CERN and funded by, e.g., the European Commission's OpenAIRE program. Upon upload, Zenodo collects all metadata like author, description, version and keywords to make datasets findable, if filled out diligently. It also generates a DOI for every dataset. As Zenodo is aimed at the researcher community, important information is collected that might be overlooked in other contexts, like funding and license information.

GitHub: GitHub[11] is a hosting service for git repositories with source code, provided by a subsidiary of Microsoft. It is the de facto standard for the development of open-source software. In addition to hosting git repositories, GitHub offers various continuous integration tools. Authors can specify metadata by including a .bib or .cff file in the main folder of the git repository[12]. GitHub provides long-term preservation mechanisms[13]. In addition, workflows exist for synchronizing it with platforms that do, e.g., Zenodo.

Institutional repositories Many institutions already offer collaborative platforms suitable for hosting research data, like nextcloud or gitlab instances. Some institutions even have their own preprint servers. Examples for institutional repositories are mathREPO at MPI MiS Leipzig for augmenting publications with code, and DaRUS at Stuttgart University. Such repositories are often hosted by libraries or central university administration. They do not always offer persistent identifiers and have varying metadata standards. It is worth checking with your current institution what the specific requirements for research data publication are.

[7] https://arxiv.org/
[8] https://creativecommons.org/licenses/by-sa/4.0/deed.en
[9] https://prodg.org/talks/welcome_to_ar5iv
[10] https://zenodo.org/
[11] https://github.com/
[12] https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-citation-files
[13] https://archiveprogram.github.com/

Domain-specific repositories arXiv and Zenodo already collect meaningful metadata, although only in a minimal fashion. Other repositories aimed at particular subcommunities within mathematics collect metadata better suited to their needs and provide decidated metadata schemas, increasing findability and accessibility within that subcommunity.

3.3 Data ownership and licensing

Using different repositories leads to data fragmentation, making it difficult to maintain a holistic view of the research data. This can impede collaboration and hence, slow down project progress. However, interlinking various data that comprise, e.g., a published paper and its supplementary material, mitigates this issue when persistent identifiers are used.

The transfer of ownership of the data to the providers of the repositories raises questions about control and access. In particular, the handling of commercial and academic repositories often differ. Academic providers are usually more focused on open access and tend to also think about long term availability. Commercial services can be profit-driven. However there are some commercial services that are considered to be de facto standards, like GitHub for source code. In that case it makes sense to enable tracking by software heritage [7] for mirroring the repository by an academic service.

The choice of appropriate licenses is a fundamental and crucial aspect in publishing research results. Often, the choice of licenses is limited by the repository provider. It is recommended to use much weaker licenses for metadata than for the actual research data. Similarly, different data sets comprising, e.g., a journal publication can be licensed differently, e.g., using an open source license[14] for software, a CreativeCommons[15] license or the publisher's OpenAccess license for the actual research article, and yet another license for data sets stemming from, e.g., a publication, or some data set used in the research project from a curated collection, like, e.g., polydb.

3.4 Metadata schemas

Metadata are a mandatory prerequisite for achieving compliance with the FAIR principles, in particular for reaching FAIR subgoals F2 and R1.[16] The main task and goal is to provide meaningful information in the form of attributes. Descriptive, technical, process, and domain-specific metadata are needed for a holistic documentation. Descriptive metadata include citation metadata, and describe the data from a higher level, typically including title, authorship, licenses or date. The typical metadata standard for this is DataCite[17]. Technical metadata (sometimes called administrative or preservation metadata) contains information about the data formats, encodings, or checksums. These first two types of metadata are generic across all subject areas. Process metadata hold information about the research process, namely, for example, what steps, equipment, software, or people went into generating the data. The Metadata4Ing standard developed by NFDI4Ing serves

[14] https://opensource.org/licenses/
[15] https://creativecommons.org/
[16] https://force11.org/info/the-fair-data-principles/
[17] https://schema.datacite.org/

as an exemplary process documentation model for engineering science and beyond [1]. Subject-specific metadata, together with process metadata, ultimately provide the information needed to achieve reproducibility, i.e., all relevant information needed to acquire the research results: the scientific statement based on the data, comprehensible, verifiable and reproducible for third parties. For example, in mathematics, this refers to the presentation of models, quantities, mathematical formulations or the algorithms used.

MaRDI is working on the standardization of such descriptions and metadata for models and algorithms in the form of knowledge graphs. A knowledge graph for algorithms is already in operation[18] and another for models is under development. It is recommended to first check if existing metadata standards already exist to avoid duplication of effort.[19] Often, existing standards can also be easily adapted to own use case.

Example 1. The neuroimage data is obtained from an open database and accompanied with rich metadata on data acquisition parameters and data meaning within the NIfTI format itself and related json-files. The simulation data with the generating R-code is uploaded to Zenodo with rich metadata describing it for findability and availability. It is referenced from the related scientific papers. The developed algorithm is contributed to the MaRDI knowledge graph. The R-code with the implementation of the developed method is compiled as an R package including required and extensive documentation. The package is uploaded to CRAN; the development of the code is done within a git repository at the institutional server and cloned to GitHub. Data and code is given a suitable license for re-use by the scientific community. The preprints of the scientific papers are equipped with meaningful keywords and MSC codes and uploaded to the institutional preprint server or `arXiv`. The scientific papers are preferably published in diamond open access journals. The PI of the project takes responsibility together with the research data steward of the institution.

Example 2. The triangulation data is obtained from an open database. The code is uploaded to `GitHub` and the output data is uploaded to `polyDB` and `Zenodo`. The preprint is uploaded to the `arXiv` and contains links to all other datasets. All datasets are equipped with rich metadata and link back to the preprint. The PI of the project takes responsibility together with the research data steward of the the institution.

4 Recommendations for RDM in Maths

In the previous sections, various options for handling research data in mathematics have been presented. Based on this, recommendations for RDM along the different phases of a project are given below. A distinction must be made between the proposal phase, in which intentions for action are defined, usually depending on the requirements of the funding agencies; the project phase, in which these intentions are implemented, reviewed and adapted, and the archiving/archival phase, in which the research data are finally made available to the community. Therefore, the RDM plan must be viewed as a living document.

[18] `https://algodata.mardi4nfdi.de/`

[19] Index of metadata standards: `https://rdamsc.bath.ac.uk/`

Like for any other publication, be it article, preprint, book or software, the RDM plan should be peer-reviewed, and ideally it should be tried to reproduce the results with the information provided. Annotated and curated RDM plans from other scientists, e.g., published on DMPonline[20], may also provide guidance. In addition, there are already several publications that address how to create a *good* RDM plan, e.g. [17, 20]. If specific problems or questions arise that cannot be answered satisfactorily, it is recommended to contact the appropriate program officer(s) at the funding agency or the institution's RDM counseling service.

Currently, there is a choice between institutional repositories like local Git installations and central repositories like GitHub. It is foreseeable that in the future, a federated approach will reduce the advantages of central commercial services like GitHub. Ultimately, one can imagine that mathematical research data will be stored provider-independently using a decentralized service stack [27, 14], and interlinking will be done based on intrinsic identifiers such as hashes or DOIs. This will eliminate the problem of choosing where (but not how!) to store the data. As a rule of thumb, at least for large repositories, there will be a migration path to future research data infrastructure, preserving the most commonly used metadata and keeping the original deposits in an immutable format. Currently, this means that the focus should be on finding a pragmatic solution for the time being. Central solutions offer significant advantages, despite the single point of failure risk and ideological concerns, that can be mitigated via local backups.

4.1 Proposal phase

Existing Guidelines Funding organizations and research institutions may have specific RDM plan guidelines or rules to follow. These guidelines should be referenced in the RDM plan. For instance, to harmonize RDM plan policies, Science Europe has defined six core requirements that every funding organization should request in an RDM plan[21]:

- **Data description.** Description of data creation, reuse and processing and their characteristics (formats, volumes). Examples of mathematical research data are discussed in Section 2.
- **Data documentation.** Description of the standards and measures used to describe the data in an understandable way and to ensure high quality. This includes appropriate and rich metadata, see Section 3.4.
- **Data storage.** Description of data storage and backup during the project, including the protection of sensitive data. For requirements and examples of data repositories, see Sections 3.1 and 3.2.
- **Legal obligations.** Description of the legal obligations, professional standards and scientific codes applicable to the data. In particular, contextualized with regard to possible restrictions on publication and accessibility. Data licenses are discussed in Section 3.3.

[20] https://dmponline.dcc.ac.uk/

[21] https://www.scienceeurope.org/media/urspcz0a/se-rdm-template-1-core-requirements-for-data-management-plans.docx

- **Data Exchange.** Description of data suitable for re-use in other contexts and where and how it is made available to third parties. Re-use requires suitable archiving after the project. For choices of repositories see, e.g., 3.2.

- **Responsibilities.** Description of responsibilities for data management and curation. In particular, the time and financial effort involved.

To accommodate the specific needs and focus of each funding agency, the order of the core requirements and their detailed design is variable. The DFG follows these core requirements within their checklist for the appropriate handling of research data in connection with DFG projects [12] for every funding proposal. In alignment with the creation of the German NFDI from within the different scientific disciplines, the checklist contains rather general hints on how to write an RDM plan at proposal phase, while discipline specific recommendations are provided in additional documents. The corresponding specifications for mathematics are provided in [11].

Regular Updates Indicate the intervals at which the RDM plan is reviewed and updated. Short-term adjustments due to significant changes remain unaffected.

4.2 Project phase

- **Data storage and backup.** Host data on suitable platforms offered by the research institution, e.g. nextcloud, Overleaf, gitlab and institutional repositories, see also Section 3.2, to enable collaborative work. In addition, backed up network drives could be used to prevent data loss. At this point of the project, findability and accessibility has only to be guaranteed for project members. Nevertheless, using external repositories like `GitHub` or `Zenodo` should be considered as soon as possible. One must also consider how much and which intermediate results of the data should be stored. As a general guideline, computationally expensive data and data on which scientific results are based hold significant value compared to data that can be easily reproduced by software.

- **Data protection.** Publish hashes of individual contributions in shared documents automatically, e.g. in Overleaf [25], to diminish the risk of stolen intellectual property.

- **Data formats.** Whenever possible, rely on widely adopted, non-proprietary file formats and apply open standards. If closed formats are unavoidable, consider open alternatives and format conversion.

- **Metadata annotation.** Add descriptive, technical and process metadata as required, see Section 3.4. Therefore, identify the information that need to be captured to enable others to find, access, interpret and use the data and determine if there are community-based metadata standards that can be adopted. Suitable standards would include, for example, OntaMathPro[22] or the MaRDI developments on algorithms and mathematical models. Software tools managing and creating metadata may also be helpful.

[22] https://ontomathpro.org/

- **Existing, well-established and open source software.** Keep maintenance and documentation work low by using existing, well-established and open source software. If closed source solutions have to be used, minimize the impact on the FAIRness of the results. Choose software languages that seamlessly integrate within the corresponding community instead of opting for exotic ones that have a low probability of re-use or integration into other projects. For documents, LaTeX is preferable over proprietary formats.

- **Continuous integration best practices.** When writing software, use a versioning system like git, set up a test framework to avoid errors and regression, use a linter for code quality and use code coverage to gauge the quality of your tests. According to the SIRS recommendations [21], all research software sources should be deposited in the universal software archive. From there, stable citations are possible, for example, using the biblatex-software package. [23]

- **Living RDM plan.** Review and update your RDM plan continuously to keep track of the project. More and more research institutions, projects, and NFDI consortia provide their members with their own instance of the *Research Data Management Organiser* (RDMO[24]) to support the RDM along the whole project life cycle using standardized templates. Use MaRDMO[25], an RDMO plugin, to document and publish interdisciplinary workflows on the MaRDI portal.

- **Machine-actionable RDM plan.** Add persistent identifiers to the RDM plan and use a controlled vocabulary to create a human-readable and machine-actionable RDM plan [18, 20].

4.3 Archival phase

- **Preprint publication.** Upload preprints to `arXiv`, the standard repository for mathematical publications, using appropriate MSC classifiers. In interdisciplinary projects, other preprint servers may be used, like bioRxiv, medRxiv, ChemRxiv, etc.

- **Source code publication.** Upload your source code to `GitHub` and activate preservation with software heritage. Software packages for e.g. `Julia` or `python` should also be made available in the respective package manager.

- **Data set publication.** Upload your data sets to a public repository. When choosing a repository, consider the implications of the FAIRness criteria. While `Zenodo` is considered as the general purpose choice for a repository, many specialized repositories from various mathematical communities exist. To maximize findability and accessibility, it always makes sense to upload data to a general purpose repository and to a community repository. When using RDMO, a direct export of data sets and corresponding metadata to `Zenodo` is possible. Define a decision process to ascertain the data that are to be stored – definitely archive the data needed to reproduce published results, but raw and intermediate data of failed experiments/simulations might also be useful!

[23] https://ctan.org/pkg/biblatex-software

[24] https://rdmorganiser.github.io/

[25] https://github.com/MarcoReidelbach/MaRDMO

- **Persistent identifiers.** When referencing papers, software and data sets, use persistent identifiers like DOIs, ORCiD IDs, swmath IDs, Wikidata IDs and more. MaRDI will provide persistent identifiers for algorithms, mathematical models, and interdisciplinary workflows soon.
- **Intrinsic identifiers.** Use intrinsic identifiers like git commit hashes and software heritage IDs to increase reproducibility.
- **RDM plan publication.** Make the RDM plan accessible to the research community to link all of the project's published data through a single resource. The plan itself can in turn serve as a template for other researchers [18, 20].

5 Conclusion

Appropriate research data management has become an important part of research in several fields, including mathematics. Given that the topic is relatively new to the field, this White Paper intends to be a first resource and guiding document to researchers in mathematics. Further development with respect to research data management is to be expected in the near future – initiated by NFDI across disciplines and by MaRDI for mathematics. Therefore, this is a living document and will be updated regularly and eventually accompanied by online resources.

Contributing authors

The writing team comprised:

Peter Benner (Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg/ Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg), Renita Danabalan (WIAS Berlin), Dominik Göddeke (Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Stuttgart Centre for Simulation Science), Lars Kastner (TU Berlin), Tabea Krause (Universität Leipzig), Daniel Mietchen (FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Berlin), Marco Reidelbach (Zuse Institut Berlin), Björn Schembera (Institute of Applied Analysis and Numerical Simulation, University of Stuttgart), Moritz Schubotz (FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Berlin), Rainer Sinn (Universität Leipzig), Karsten Tabelow (WIAS Berlin)

The current version of this document was approved by the MaRDI Board on 19 September 2023.

References

[1] Susanne Arndt et al. *Metadata4Ing: An ontology for describing the generation of research data within a scientific activity.* Version 1.1.0. Feb. 2022. DOI: 10.5281/zenodo.7706017.

[2] Monya Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533 (2016), pp. 452–454. DOI: 10.1038/533452a.

[3] Wolfgang Bangerth and Timo Heister. “Quo vadis, scientific software?” In: *SIAM News* (2014). URL: https://sinews.siam.org/Details-Page/quo-vadis-scientific-software-1 (visited on 08/30/2023).

[4] Peter Benner et al. “Die Mathematische Forschungsdateninitiative in der NFDI: MARDI (Mathematical Research Data Initiative)”. In: *GAMM-Rundbrief* 01/2022 (2022), pp. 40–43. URL: https://www.gamm-ev.de/wp-content/uploads/2022/05/GAMM_1-22web.pdf (visited on 10/05/2023).

[5] Tobias Boege et al. *Research-Data Management Planning in the German Mathematical Community*. 2022. arXiv: 2211.12071 [math.HO].

[6] Neil P. Chue Hong et al. *FAIR Principles for Research Software (FAIR4RS Principles).* Research Data Alliance. 2021. DOI: 10.15497/RDA00065.

[7] *Collect, organise, preserve and share the Software Heritage of mankind*. 2016. URL: https://www.softwareheritage.org/wp-content/uploads/2016/06/PressReleasePressKit-2016-06-30.en_.pdf (visited on 10/05/2023).

[8] Johan Commelin. “Liquid Tensor Experiment”. In: *Mitteilungen der Deutschen Mathematiker-Vereinigung* 30.3 (2022), pp. 166–170. DOI: 10.1515/dmvm-2022-0058.

[9] Antony Della Vecchia, Michael Joswig, and Benjamin Lorenz. *A FAIR File Format for Mathematical Software*. 2023. arXiv: 2309.00465.

[10] Deutsche Forschungsgemeinschaft. *Guidelines for Safeguarding Good Research Practice*. URL: https://wissenschaftliche-integritaet.de/en/code-of-conduct/ (visited on 10/05/2023).

[11] Deutsche Forschungsgemeinschaft. *Handreichung zu RDM in der Mathematik*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/handreichung_fachkollegium_mathematik_forschungsdaten.pdf. (Visited on 10/05/2023).

[12] Deutsche Forschungsgemeinschaft. *Umgang mit Forschungsdaten*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf. (Visited on 10/05/2023).

[13] Giancarlo Guizzardi. “Ontology, ontologies and the “I” of FAIR”. In: *Data Intelligence* 2.1-2 (2020), pp. 181–191. DOI: 10.1162/dint_a_00040.

[14] Cornelius Ihle et al. “Incentive Mechanisms in Peer-to-Peer Networks — A Systematic Literature Review”. In: *ACM Comput. Surv.* (2023). DOI: 10.1145/3578581.

[15] Kevin Lang et al. *FAIR Assessment Tools Overview*. Version 2.1. Feb. 2023. DOI: 10.5281/zenodo.7701941.

[16] Thomas Ludwig and Beate Geyer. “Reproduzierbarkeit”. In: *Informatik Spektrum* 42 (2019), pp. 48–52. DOI: 10.1007/s00287-019-01149-2.

[17] William K. Michener. “Ten Simple Rules for Creating a Good Data Management Plan”. In: *PLOS Computational Biology* 11.10 (2015), pp. 1–9. DOI: 10.1371/journal.pcbi.1004525.

[18] Tomasz Miksa et al. “Ten principles for machine-actionable data management plans”. In: *PLOS Computational Biology* 15.3 (2019), pp. 1–15. DOI: 10.1371/journal.pcbi.1006750.

[19] Mark Musen. “Making Data FAIR Requires More than Just Principles: We Need Knowledge Technologies”. In: *2019 15th International Conference on eScience (eScience)*. 2019, pp. 530–532. DOI: 10.1109/eScience.2019.00071.

[20] European Commission. Directorate General for Research and Innovation. *Turning FAIR into reality*. Publications Office, 2018. DOI: 10.2777/1524.

[21] European Commission. Directorate General for Research and Innovation. *Scholarly Infrastructures for Research Software: Report from the EOSC Executive Board Working Group (WG) Architecture Task Force (TF) SIRS.* Publications Office, 2020. DOI: 10.2777/28598.

[22] Christian Riedel et al. *Including Data Management in Research Culture Increases the Reproducibility of Scientific Results*. INFORMATIK 2022. 2022. DOI: 10.18420/inf2022_114.

[23] Sheeba Samuel and Daniel Mietchen. *Computational reproducibility of Jupyter notebooks from biomedical publications*. 2022. arXiv: 2209.04308 [cs.CE].

[24] Björn Schembera and Juan M Durán. “Dark data as the new challenge for big data science and the introduction of the scientific data officer”. In: *Philosophy & Technology* 33 (2020), pp. 93–115. DOI: 10.1007/s13347-019-00346-x.

[25] Moritz Schubotz et al. *Repurposing Open Source Tools for Open Science: a Practical Guide*. 2018. DOI: 10.5281/zenodo.2453415.

[26] The MaRDI consortium. *MaRDI: Mathematical Research Data Initiative Proposal*. 2022. DOI: 10.5281/zenodo.6552436.

[27] Dennis Trautwein et al. “Design and evaluation of IPFS: a storage layer for the decentralized web”. In: *SIGCOMM ’22: ACM SIGCOMM 2022 Conference, Amsterdam, The Netherlands, August 22 - 26, 2022*. Ed. by Fernando Kuipers and Ariel Orda. ACM, 2022, pp. 739–752. DOI: 10.1145/3544216.3544232.

[28] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3 (2016). DOI: 10.1038/sdata.2016.18.