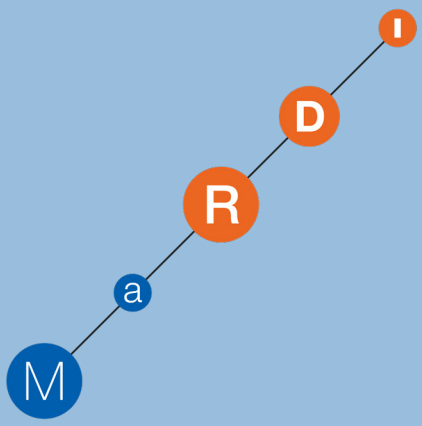
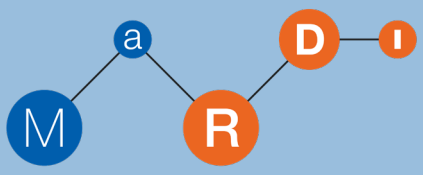
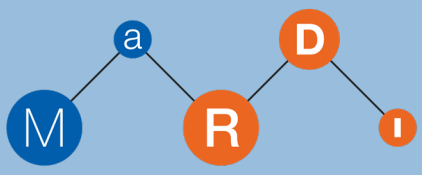
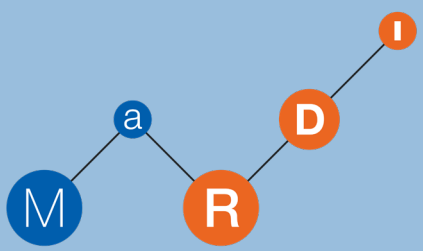
MaRDI
Mathematical Research Data Initiative









Author: The MaRDI Consortium –
Mathematical Research Data Initiative of NFDI
DOI: 10.5281/zenodo.6552436

Contents

Author: The MaRDI consortium –
Mathematical Research Data Initiative for NFDI

DOI:10.5281/zenodo.6552436

1 General Information

Name of the consortium in English and German

Mathematical Research Data Initiative
Mathematische Forschungsdateninitiative

Summary of the proposal in English

Mathematical research data (MRD) has become vast, it is complex, multifaceted, and, by the interdisciplinary potential of mathematics with its power of abstraction, wide spread in the sciences. It emerges within mathematical sciences but also in other scientific areas, and ranges from highly complex data from scientific computing to information bases like the standard reference data for special functions, tables etc. of, e.g., the US National Institute for Standards and Technology. The data volume and its creation velocity increase dynamically with the rapid unfolding of mathematics in data science and the ever-increasing compute power. In various scientific fields (such as, e.g., physics, chemistry, engineering sciences, humanities and life sciences), this leads to increasingly complex mathematical and computational models, and extremely diverse MRD. We hence aim at developing a MRD infrastructure useful in mathematics *and* with significant impact in other fields.

Motivated by demands of the mathematical community and other disciplines utilizing quantitative methods, MaRDI (**Ma**thematical **R**esearch **D**ata **I**nitiative), *the* consortial initiative of mathematical sciences, aims to set standards for certified MRD, the design of confirmable workflows, and it provides community services. The designated goal is to realize the FAIR principles across all of mathematics and its applications, interoperability of data handling, and to propel new research.

Targeting certified data, software developments, confirmable workflows, and the provision of services, the four research motivated pillars of MaRDI are computer algebra, scientific computing, statistics and machine learning, and interdisciplinary mathematical research. The latter leads to use cases and cooperations with the consortia NFDI4Ing, NFDI4Chem, NFDI4Culture, NFDI4Cat, and initiatives including PUNCH4NFDI, NFDI-MatWerk, NFDIxCS, NFDI-Neuro, BERD@NFDI, NFDI4DataScience, and NFDI4MobilTech.

For MRD from the aforementioned fields standardized formats, data interoperability and application programming interfaces will be developed. Services on pilot level will be expanded into full services providing added value to current research. MaRDI will also significantly enhance information retrieval services, including mathematical models as research data, a cross-disciplinary mathematical digital semantic atlas, ontologies, and metadata such as, e.g., an algorithm metadata library. Moreover, MaRDI will develop a digital service portal as a one-stop unique contact point for the scientific community to retrieve and consult MaRDI services. This portal will be developed in an agile fashion and installed permanently.

The sustainable realization of MaRDI findings requires a community adhering to a FAIR data culture and FAIR research workflows. MaRDI will thus build collaborative platforms which are pivotal for knowledge dissemination, the scientific discourse and quality control.

Summary of the proposal in German

Mathematische Forschungsdaten (MFD) sind umfangreich, komplex und vielfältig. Durch die Interdisziplinarität und Abstraktionskraft der Mathematik sind sie in den Wissenschaften weit verbreitet. So tauchen sie in der Mathematik, aber auch in anderen Wissenschaftsbereichen auf und reichen von hochkomplexen Daten aus dem wissenschaftlichen Rechnen bis hin zu Informationsdatenbanken wie den Standard-Referenzdaten für spezielle Funktionen, Tabellen usw., bereitgestellt z.B. vom US National Institute for Standards and Technology. Das Datenvolumen und dessen Entstehungsgeschwindigkeit nehmen mit der raschen Entfaltung der Mathematik in den Datenwissenschaften und der ständig steigenden Rechenleistung dynamisch zu. In verschiedenen Disziplinen (wie Physik, Chemie, Ingenieur-, Geistes- und Biowissenschaften) führt dies zu immer komplexeren mathematischen Modellen und MFD. Unser Ziel ist es daher, eine Forschungsdateninfrastruktur zu entwickeln, die nicht nur für die Mathematik, sondern auch in anderen Bereichen von großem Nutzen sein wird.

MaRDI (**Ma**thematical **R**esearch **D**ata **I**nitiative), *die* Konsortialinitiative der Mathematik, zielt auf Standards für zertifizierte MFD, reproduzierbare Arbeitsabläufe und Dienstleistungen für die Wissenschaftsgemeinschaft ab. Dabei ist es das erklärte Ziel, die FAIR-Prinzipien in der Mathematik und ihren Anwendungen sowie die Interoperabilität der Datenverarbeitung umzusetzen und neue Forschung voranzutreiben.

Die vier forschungsmotivierten Säulen von MaRDI sind Computeralgebra, wissenschaftliches Rechnen, Statistik und maschinelles Lernen sowie interdisziplinäre Mathematik. Sie zielen jeweils auf zertifizierte Daten- und Softwareentwicklungen sowie auf bestätigbare Arbeitsabläufe und die Entwicklung von Diensten ab. Die interdisziplinäre Stärke der Mathematik führt zu Kooperationen mit den Konsortien NFDI4Ing, NFDI4Chem, NFDI4Culture, NFDI4Cat und Initiativen wie PUNCH4NFDI, NFDI-MatWerk, NFDIxCS, NFDI-Neuro, BERD@NFDI, NFDI4DataScience und NFDI4MobilTech.

Für die dabei anfallenden MFD werden standardisierte Formate bzw. Dateninteroperabilität und Anwendungsprogrammierschnittstellen entwickelt sowie prototypische Dienste zu Volldiensten mit Forschungsmehrwert ausgebaut. MaRDI wird verbesserte Informationsdienste entwickeln, die mathematische Modelle als Forschungsdaten, einen interdisziplinären mathematisch-digitalen Semantikatlas, Ontologien und Metadaten, wie z.B. eine Algorithmen-Metadatenbibliothek, umfassen. MaRDI wird auch ein digitales Serviceportal als zentrale Anlaufstelle für die Wissenschaft aufbauen. Dieses Portal wird in agiler Weise entwickelt und dauerhaft installiert.

Die nachhaltige Umsetzung der Ergebnisse von MaRDI erfordert eine Gemeinschaft, die sich auf eine FAIR-Datenkultur und FAIR-Forschungsabläufe stützt. Dazu wird MaRDI Kooperationsplattformen zur Wissensverbreitung, für den wissenschaftlichen Diskurs und die Qualitätssicherung aufbauen.

Applicant institution	Location
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS) Leibniz-Institut im Forschungsverbund Berlin e. V. Director: Michael Hintermüller	Mohrenstraße 39 10117 Berlin

Name of the consortium spokesperson	Institution, location
Michael Hintermüller email: michael.hintermueller@wias-berlin.de	WIAS Berlin Mohrenstraße 39, 10117 Berlin

Co-applicant institutions	Location
Deutsche Mathematiker-Vereinigung e.V. (DMV) c/o WIAS President: Friedrich Götze, from January 2021 on: Ilka Agricola	Mohrenstraße 39 10117 Berlin
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH (FIZ) CEO: Sabine Brünger-Weilandt	Hermann-von-Helmholtz-Platz 1 76344 Eggenstein-Leopoldshafen
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V. für ihr Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM) Director: Anita Schöbel	Fraunhofer-Platz 1 67663 Kaiserslautern
Freie Universität Berlin (FUB) President: Günter M. Ziegler	Kaiserswerther Str. 16/18 14195 Berlin
Ludwig-Maximilians-Universität München (LMU) President: Bernd Huber	Geschwister-Scholl-Platz 1 80539 München
Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO) Director: Gerhard Huisken	Schwarzwaldstraße 9-11 77709 Oberwolfach-Walke
Max-Planck-Gesellschaft zur Förderung der Wissenschaften e. V. für ihr Max-Planck-Institut für Dynamik komplexer technischer Systeme (MPI DCTS) Managing Director: Udo Reichl	Sandtorstraße 1 39106 Magdeburg
Max-Planck-Gesellschaft zur Förderung der Wissenschaften e. V. für ihr Max-Planck-Institut für Mathematik in den Naturwissenschaften (MPI MIS) Managing Director: Bernd Sturmfels	Inselstraße 22 04103 Leipzig
Technische Universität Berlin (TUB) President: Christian Thomsen	Straße des 17. Juni 135 10623 Berlin
Technische Universität Kaiserslautern (TUK) President: Arnd Poetzsch-Heffter	Gottlieb-Daimler-Straße 47 67653 Kaiserslautern
Technische Universität München (TUM) President: Thomas Hofmann	Arcisstraße 21 80333 München
Universität Stuttgart (USTUTT) Rector: Wolfram Ressel	Keplerstraße 7 70174 Stuttgart
Westfälische Wilhelms-Universität Münster (WWU) Rector: Johannes Wessels	Schloßplatz 2 48149 Münster
Zuse Institut Berlin (ZIB) President: Christof Schütte	Takustraße 7 14195 Berlin

Names of co-spokesperson	Institution, location	Task area(s)
Michael Hintermüller	WIAS, Mohrenstraße 39, 10117 Berlin	T7
President of DMV (Jan 21 ++: Ilka Agricola)	DMV, Mohrenstraße 39, 10117 Berlin	T6
Peter Benner	MPI DCTS, Sandtorstraße 1, 39106 Magdeburg	T2
Bernd Bischl	LMU, Ludwigstraße 33, 80539 München	T3
Wolfram Decker	TUK, Gottlieb-Daimler-Str. 47, 67653 Kaiserslautern	T1
Mathias Drton	TUM, Boltzmannstr. 3, 85748 Garching	T3
Claus Fieker	TUK, Gottlieb-Daimler-Str. 47, 67653 Kaiserslautern	T1
Dominik Göddeke	USTUTT, Allmandring 5b, 70569 Stuttgart	T4
Michael Joswig	TUB, Straße des 17. Juni 135, 10623 Berlin	T1
Stephan Klaus	MFO, Schwarzwaldstraße 9-11, 77709 Oberwolfach-Walke	T6
Mario Ohlberger	WWU, Schloßplatz 2, 48149 Münster	T2
Harald Sack	FIZ, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen	T5
Anita Schöbel	ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern	T4
Christof Schütte	ZIB, Takustraße 7, 14195 Berlin	T5
Rainer Sinn	FUB, Kaiserswerther Str. 16/18, 14195 Berlin	T6
Bernd Sturmfels	MPI MIS, Inselstraße 22, 04103 Leipzig	T6

Participating institutions	Location
European Mathematical Society Department of Mathematics and Statistics	P.O.Box 68 00014 University of Helsinki, Finland
Gesellschaft für Angewandte Mathematik und Mechanik e. V. (GAMM)	Institut für Statik und Dynamik der Tragwerke Fakultät Bauingenieurwesen, 01062 Dresden
Gesellschaft für Operations Research e. V. (GOR)	Kackertstraße 7, 52072 Aachen
IMAGINARY gGmbH	Mittenwalder Str. 48, 10961 Berlin

Contribution of EMS: The European Mathematical Society (as *the* roof organization of all mathematical societies in Europe) pursues, as one of its central goals, the open and fair access to research data and publications. Its publishing house EMSPress, which publishes 20 top journals in mathematics, has its headquarters in Berlin. The EMS has just recently transformed all its journals to the subscribe-to-open model and is also partner in publishing the open Encyclopedia of Mathematics (formerly Springer), the European Digital Math Library (DML) and the very important information service zbMATH, which will be open access starting from January 2021. Within MaRDI, the EMS plans to make the Encyclopedia of Mathematics compatible with MaRDI. One of the many goals of EMS to join MaRDI is to take the developed concepts and ideas to the European level and to support other

European countries and the European Union in their transformation to the FAIR principles. Moreover, EMS plans to become a member of EOSC, the European Open Science Cloud[1].

Contribution of GAMM: The GAMM will support the dissemination of MaRDI results through its own publications like the “GAMM Rundbrief” or the “GAMM Mitteilungen”, via its mailing lists, as well as by providing time slots for dedicated sessions within its annual meeting (“GAMM Jahrestagung”) with an attendance of more than 1,200 researchers in applied mathematics and mechanics.

Contribution of GOR: GOR is mainly interested in standardizing mathematical models and developing cross-disciplinary metadata standards. For MaRDI, GOR will contribute use cases which arise in many discussions and applications in its working groups.

Contribution of IMAGINARY gGmbH: IMAGINARY will design and conduct training workshops and develop dissemination material to promote MaRDI within the mathematical community and to a broader scientific and public audience. Among this material is a highly innovative interactive open-source exhibition module, *the MaRDI station*.

Participating individuals	Institution, location
Peter Bastian for the Exzellenzcluster “STRUCTURES: A unifying approach to emergent phenomena in the physical world, mathematics, and complex data” (EXC 2181)	Ruprecht-Karls-Universität Heidelberg Institut für Theoretische Physik Philosophenweg 16, 69120 Heidelberg
Bettina Eick Technische Universität Braunschweig	Universitätsplatz 2 38106 Braunschweig
Thomas Ertl for the Exzellenzcluster “Data-integrated Simulation Science” (EXC 2075)	Universität Stuttgart, SC SimTech Pfaffenwaldring 5a, 70569 Stuttgart
Dieter Fellner for the Fraunhofer-Verbund IUK-Technologie	Anna-Louisa-Karsch-Str. 2, 10178 Berlin
Michael Kohlhase Friedrich-Alexander Universität Erlangen-Nürnberg (FAU)	Schloßplatz 4, 91054 Erlangen
Martin Skutella for the Exzellenzcluster MATH$^+$ (EXC 2046)	Technische Universität Berlin, Sekr. MA 2-2 Strasse des 17. Juni 136, 10623 Berlin

Contribution of Bettina Eick (Professor at the Institut for Analysis and Algebra of the TU Braunschweig): She has been a project leader in establishing the so-called Small Groups Library, a database of algebraic groups of certain orders. It has been an extensive project to collect resp. generate and curate the data in this library. Even today, it is still difficult to check the correctness of the data with state-of-the-art technology. The Small Groups Library is widely used in many areas of algebra, mathematics in general, and beyond. Hence the library has a very significant value for the research community. The research group of Bettina Eick will support MaRDI with knowledge and test cases arising from this database.

Contribution of Martin Skutella for the EXC 2046: MATH$^+$ offers to support MaRDI ’s activities by integrating MaRDI results into the MATH$^+$ research agenda. It will establish a direct cooperation through the Chief Research Data Officer, e.g., as a guest in MaRDI Boards in order to interlink specific agenda points of both institutions. MATH$^+$ will provide relevant use cases, also in cooperation with

[1] https://ec.europa.eu/research/openscience/index.cfm

other disciplines, as well as a training platform for early career researchers.

Contribution of Thomas Ertl for the EXC 2075: SimTech will contribute to a number of MaRDI goals. This is emphasized by the fact that SimTech contributes its resources equally to MaRDI, NFDI4Ing and NFDI4Chem. In addition to the commitment of the University of Stuttgart, SimTech commits itself to providing access to its RDM platform openDASH. All researchers receiving funding from EXC 2075 commit themselves to contribute data and software to openDASH, and the SimTech data stewards will be closely involved in curating ontologies and metadata schemes, and in providing interfaces to other data hubs within MaRDI.

Contribution of Peter Bastian for the EXC 2181: STRUCTURES will contribute to the standardization of input to and output of computer simulations, thus contributing to the interoperability of different codes, and provide use cases and reference results in numerical simulation and data analysis. The close connection to physicists and computer scientists provides also links to other RDIs, in particular in astrophysics and particle physics. Dissemination of MaRDI results will be part of the training in STRUCTURES.

Contribution of Dieter Fellner for the Fraunhofer-Verbund IUK: The Fraunhofer ICT Group will contribute to MaRDI by informing its members regularly about the progress of the work of the MaRDI consortium and will thus provide a platform of user exchange. Further, the Group will spread the information into the application areas of its respective institutes. This will foster the transfer of the results of MaRDI and eventually the wide use of mathematical research data.

Contribution of Michael Kohlhase (Professor and head of RD management services at FAU Erlangen-Nürnberg): He conducts research in automated theorem proving, representation and management of mathematical knowledge, and the linguistics of mathematical language. These have been applied for Math-on-the-Web standards, such as, e.g., MathML, OpenMath, and OMDoc/MMT and various end-user systems for the management of mathematical research data: MathHub.info hosts the major theorem prover libraries in the OMDoc/MMT format, arXMLiv hosts all 1.7M preprints from arXiv.org translated to semantic XHTML+MathML, MathDataHub.info hosts math object collections FAIRly and MathWebSearch makes them semantically searchable. He will contribute expertise in all of these areas to the MaRDI consortium.

Names and numbers of the DFG review boards (DFG Fachkollegien) that reflect the subject orientation of the proposed consortium

Mathematics 312-01

2 Scope and Objectives

The MaRDI consortium has a strong core in and a wide embedding into mathematics and its applications. Through applications it reaches out from mathematics to many different scientific disciplines as well as research and development in industry, which helps to identify relevant use cases. The interaction with users is indeed present on all levels of MaRDI. This includes in particular the interaction and joint activities with a large number of companion NFDI initiatives or consortia established in the first NFDI call as well as the alliance with major mathematical communities or societies such as the

European Mathematical Society (EMS), the “Gesellschaft für Angewandte Mathematik und Mechanik” (GAMM), “Gesellschaft für Operations Research” (GOR), and several more. These links specifically help to shape and evolve MaRDI services in a user-centric fashion. Moreover, the consortium will advance its agenda in an open, interactive and community oriented way.

2.1 Research domains or research methods addressed by the consortium, specific aim(s)

Mathematical research data has become vast, it is complex, diverse, and multifaceted, and, through the successful application of mathematics in interdisciplinary research, it is wide spread in the scientific landscape. It emerges within mathematics as a discipline but also in other scientific areas, and ranges from highly complex data in scientific computing to information bases such as the standard reference data for special functions, tables etc. as provided by the US National Institute for Standards and Technology and others, which are routinely consulted by experts from various disciplines. Motivated by the needs and requests from the mathematical community, but also from other scientific disciplines that utilize quantitative methods, MaRDI aims to set standards for certified mathematical research data, the design of confirmable workflows, and to develop further services for the scientific community. Even more, mathematical research data is extremely diverse and stems from many different disciplines which use mathematical methods and solution approaches. Moreover, mathematics is a science which can build abstract structures which can be applied not only within mathematics but also in other disciplines. We hence aim at developing a research data infrastructure not only needed in mathematics but which will be applicable also in other fields.

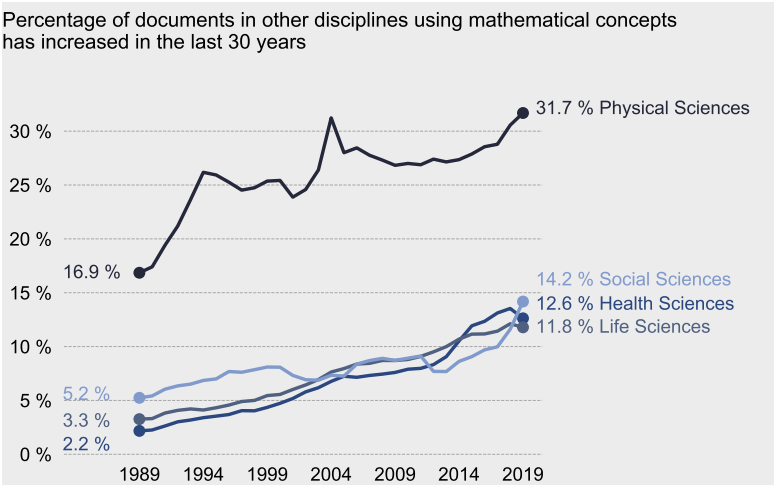


Figure 1: Percentage of peer-reviewed publications using mathematical concepts compared to the total number in each subject area excluding mathematics itself based on a Scopus query using mathematical keywords. For details see [SG].

This ambitious goal can be reached, since through the versatility and wide applicability of mathematics MaRDI is well connected to a considerable number of co-applying or already established NFDI consortia. This has lead to the embedding of several interdisciplinary use cases into this proposal, see task area **T4** for details, which address certified data requirements and confirmable work flows at the interface between mathematics and other disciplines. MaRDI services will propel further mathematical or quantitative research (leading to added value through research sophistication). They will also raise information retrieval to an entirely new level, e.g., by establishing an expanding data base on numerical algorithms along with their application scopes, flexible cross-disciplinary semantic libraries or ontologies

along mathematical objects and models.

For example, in case a user has identified a system of differential equations modeling a real-world phenomenon of interest, MaRDI's mathematical model library can be used to retrieve mathematical information on that model such as, e.g., references to results confirming existence and properties of solutions, a semantics library will help to connect mathematical descriptors with the ones relevant in the target application, and plugging in MaRDI's algorithm database will connect the aforementioned system of equations with appropriate solution algorithms and associated information. Moreover, through the integration of **swMath**[2] into MaRDI the user will get links to software packages with implementations of the target algorithm along with pointers to the relevant literature. In summary, MaRDI aims at providing comprehensive, versatile and agile access to a wide landscape of *interoperable* mathematical research data that will advance research within mathematics, but also in many other scientific fields. Achieving this ambitious goal is facilitated by the architecture of MaRDI in form of cross layers focusing on interoperability of data handling (core), findability and accessibility through repositories and suitable interfaces (data), innovative data services to enable new research and scientific added value (exchange), and, finally, the expansion of semantic knowledge through ontologies and knowledge graphs (knowledge). This architecture is designed to secure the adhesion to the FAIR principles and to advance research.

In this context, it is a designated goal of MaRDI to provide the vast majority of these services through an open digital MaRDI-Portal, which will be developed and implemented by the consortium. This agile portal concept serves as a one-stop contact point for mathematical research data for the scientific community. It is our goal to provide a sustainable service, hence, this service portal solution as well as the strategic conception of MaRDI are designed to guarantee long term availability and development. Through design and realization, all of these developments associated with MaRDI, the specific NFDI consortial application by mathematicians, will strictly adhere to the FAIR principles [Wil+16]. This will not only lead to sustainable and re-usable research data, but it will also open up mathematical concepts and strengthen their visibility to a broad scientific community. Indeed, findable and accessible mathematical objects (including models, equations, formulas, etc.) and mathematical results are of paramount importance for unleashing the full creative power of mathematics; compare Figure 1. Moreover, interoperability of MaRDI's versatile research data and services is another key focus within the consortium, but, of course, also in cooperation with all NFDI consortia with interlinkage to mathematical research data.

While we will be more specific on the need for realizing the FAIR principles in mathematical research data, which clearly goes beyond mathematics as a discipline, and on addressing the objectives of MaRDI below, we next anticipate the structural components of the consortium. The four research motivated pillars of this proposal (each of them and among others connected to specific data types) are computer algebra **T1** (exact data), scientific computing **T2** (floating point data), statistics and machine learning **T3** (uncertain data), and interdisciplinary mathematical research **T4** (mixed data along use cases). All these areas aim to identify certified data and software development as well as confirmable workflows in their respective realm. Concerning the resulting (portal) services, the development of each of these areas will lead to a prototype which will then be turned into a MaRDI-service or product

[2] https://swmath.org

in a designated domain. Finally, we emphasize in view of these areas related to computer-aided mathematics that MaRDI's reach is clearly beyond. In fact, it addresses mathematical research in its entire width through, e.g., standardization, services, and community interaction to guarantee overall FAIRness.

These four pillars circle around computational methods where new workflows and practices are of dire need to guarantee reproducible and trustworthy computational results, see [Her], [Feh+16], [BBS16], [Her19], [Per19], e.g. by making software, input data, and execution environment used to produce published results (openly) available. Unfortunately, such factual materials are often not available for editors and reviewers, e.g., of scientific journals as well as the reader in today's document-oriented workflows. This is, of course, also true for experts who wish to exploit such findings in their own work environment and may suffer from limitations in reproducibility. Indeed, this lack of availability of data and their FAIRness results in difficulties in understanding and penetrating the entire research environment of a result [DKK18]. This can been seen as a part of the "reproducibility crisis" in science as discussed, e.g., in [SBB13]. In a survey [Bak16], more than 80% of the scientists blamed the unavailability of methods and code as a factor for irreproducible research. One definition of research data is "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" [MCG]. In this sense, confirmability and reproducibility, FAIRness in general, of scientific results require a novel, comprehensive and agile research data management.

As indicated above, MaRDI's focus on its representative mathematical areas is very natural based on the rough division of mathematical research data in *exact and symbolic data* **T1**, *floating point data* **T2**, and *data with uncertainty* **T3**.

Starting from a merely mathematical internal perspective, computer algebra systems are prototypical for *exact* computations. They play a prominent role in formal proofing or proof verification and, thus, branch out to other quantitative scientific disciplines that use, given some hypotheses and a system of axioms (or validated, certified basic facts), logical reasoning to prove (or disprove) a scientific statement. The latter is highly important in obtaining rigorous certificates for complex proofs. Indeed, on the one hand, rigorous proofs based on an axiomatic system are one of the paramount strengths of mathematical science as they guarantee the indefinite correctness of results (typically expressed in form of theorems), on the other hand, due to a tremendous increase in complexity, proof validation by human expert reviewers comes under pressure. Formal proofs, which inevitably require computer-based methods, and their validation by automated proof assistants like Isabelle may thus become an increasingly important tool within the rigorous quality standard in mathematics. In a sense it provides a way out of a potential confirmability crises in theoretical mathematics (also applied in other disciplines) and will shape next generation peer review. Obviously, such formal tools must be based on certified data and confirmable workflows and software environments in order to secure objective, indefinite correctness. **T1** of this proposal is primarily concerned with this complex.

The field of scientific computing is confronted with a vast amount of data. These involve heterogeneous input / output data, mathematical software and data produced during runtime of algorithms.

For instance, simulations in fluid dynamics require inflow profiles, flow domains, mathematical model data in form of partial differential equations in addition to physical parameters characterizing the fluid. When considering, e.g., optimal design problems in fluid dynamics, sensitivity based

minimization schemes need in addition to the above an objective to be optimized and the adjoint equation which generates associated data in its own right and in such a way that the entire data stack requires smart processing and storage techniques to be handleable at all. Moreover, the attained numerical results will always crucially depend on the employed solution algorithm as well as discretization or model order reduction techniques, which are further relevant data components. In the MaRDI structure this realm is subsumed in Task Area **T2**.

Many problems in mathematics and its applications are related to uncertain data. Research in the related fields of Statistics and Machine Learning focuses on the development of broadly applicable methods for data analysis that solve prediction problems, support decision-making, and infer structure underlying a scientific phenomenon. While its methods draw on a wide variety of computational techniques that include many of the numerical methods considered in scientific computing (**T2**) and also symbolic computation as considered in computer algebra (**T1**), data processed in the area also have the distinguishing feature of being inherently *uncertain*, i.e., subject to *stochastic noise*. Separating this noise from the signal of interest is a key challenge that arises in virtually all branches of Science and Engineering. Addressing this challenge and designing methods that ensure that conclusions drawn in scientific studies are likely to persist in independent replication studies is a chief goal of the corresponding Task Area **T3**. Moreover, deep learning methods for specific components of a physics model, for example, require input/output measurements (of the underlying physical process) possibly along with statistical information on uncertainty, the topological structure of the deep network, the shape of transition and activation functions along with the underlying numerical optimization resp. solution algorithms for computing the network parameters. In nowadays applications this again leads to a vast amount of heterogeneous data. The area of mathematics of machine learning emanates from **T3**, but very naturally connects to numerical algorithms in **T2** and use cases in **T4**.

The fourth research motivated pillar of MaRDI building on (model) data, software, and workflows comes from the interdisciplinarity of mathematical research. 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 whole NFDI it is also expected that this applicability is even more strengthened. In **T4** we primarily use a variety of highly relevant use cases which will be developed together with interdisciplinary partners, but in particular also with many other NFDI consortia such as NFDI4Ing, NFDI4Cat, NFDI4Chem, NFDI4Culture and consortial initiatives like NFDI-MatWerk, PUNCH4NFDI, NFDIxCS, FAIRmat, NFDI-Neuro to name a few. In each context, a real world problem is simplified in order to be represented by equations or inequalities together with initial / boundary conditions (= model). The model and specific input data form an instance of the problem. Next, an algorithm exactly or approximately transforms input data (the instance) into output data (the solution). The algorithm must then be validated and, finally, the solution interpreted with respect to its original context. Models and algorithms are classes of research data that need to be linked in order to make them accessible. Furthermore, the connection with a large number of other NFDI consortia in particular through **T4** is a solid underpinning for achieving MaRDI's goal of (data) interoperability across the entire NFDI.

These four areas described above will work from recent representative research outcomes towards

a respective service prototype with respect to certified data, confirmable workflows and software. Towards the end of their respective development stack, these services will be integrated into the MaRDI-Portal in **T5**. The goal of this development is the provision of services and products for the scientific community through a permanent digital portal. **T1**, **T2**, **T3**, **T4** also have the advantage of naturally providing another dimension of "dynamization" of MaRDI's strategic development: By organizing MaRDI's tasks into data-type related sub-communities each of them can benefit from the developments in the other areas with respect to data standardization, services and community integration. Then, **T4** commences its work by interdisciplinary exchange to establish a workflow for reducing a real world problem to a mathematical model. It will also rely and expand on certified data, confirmable workflows and software principles, and enriches this by FAIR aspects from projects at the interface between disciplines.

2.2 Objectives and measuring success

The approach of MaRDI to the challenges mentioned above is to introduce mathematical research data together with workflows and services supporting all work phases of a mathematician, guaranteeing findable, accessible, interoperable and re-usable results and objects in mathematics as well as in other disciplines. Hence, mathematics has to deal with all aspects of the FAIR principles.

In order to reach its goals, the consortium will address the following objectives:

O1: Interoperable mathematical research data: The goal is to develop interoperability of mathematical research data (from symbolic, numerical or runtime time, uncertain data, equations, functions, to mathematical models and model hierarchies). This data conception should not only be useful within mathematics, but also suitable for scientific disciplines that process mathematical data.

O2: Secure confirmable and reproducible results: Mathematics has the unique property that its results can not only be made plausible by theory and experiment, but rigorously proven. However, complexity has increased vastly, and reproducibility of results has become an issue. In particular, MaRDI will focus on the interlinkage of mathematical results, software, and mathematical data with the aim to guarantee confirmability and reproducibility.

O3: Establish confirmable workflows: In mathematics, the confirmability of scientific results is primarily associated with proofs, however, since the advent of computer-aided mathematics and scientific computing, the entire workflow needs to be documented, including types and versions of software used, program code, intermediate results, and many more. MaRDI aims to establish tools and standards for making this workflow transparent and reliable, also for other disciplines.

O4: Development of mathematical services: Services need to be developed that make the creation of FAIR research data easy and attractive for the user. A central element in this service orientation is the creation of an easy to access agile digital portal that gathers the majority of MaRDI services.

O5: Establish next-generation peer review: Correctness of computational methods, program code, and software cannot be ensured by traditional means of peer review. MaRDI aims to establish standards and requirements for ensuring the correctness of these research results.

O6: Standardize semantic relations: Mathematical research data is used in many disciplines and is the foundation of cross-disciplinary methodologies, mathematical modeling and simulation (MMS) as well as statistics and data science. MaRDI establishes a common ground between the disciplines concerning standards for models, software and other data. These standards link semantically related research data and thereby guarantee findability.

O7: Culture development and community embedding: The long-term goal is the recognition of FAIR mathematical research data as an accomplishment in its own right. This includes establishing standards, services and workflows for mathematical research data by embedding into the mathematical and related communities, training of mathematicians in good practices of data handling and by broad dissemination through conferences and publications.

In the longterm MaRDI aims at securing and advancing the FAIR principles in the mathematical sciences and in disciplines relying on mathematically quantified research. For this it will contain agile task areas reflecting the latest challenges in mathematical research data. These areas will advance their agenda towards introducing prototypes which can then be transferred into services in the digital MaRDI portal. MaRDI's culture and community development as well as the standardization of data is aimed to foster community intrinsic appropriate standardization oriented data, workflow, and software development whenever novel data types arise. Finally, the digital portal is designed to become a permanent one-stop contact point for the scientific community for retrieving or using services concerning mathematical research data. Through the objectives above MaRDI will contribute to the general aims of the NFDI to establish data handling standards and guidelines, to develop an interoperable data management, to foster reusability of existing data within mathematics as well as beyond, and to disseminate its experience to the scientific community in Germany as well as worldwide. In this way MaRDI will be a reliable, robust and well-connected consortial node within the NFDI.

In order to monitor the success of achieving MaRDI's aims and objectives, MaRDI will observe the following key performance indicators (KPIs):

- Number of users of the MaRDI portal and specific services
- Number and volume of data sets integrated into MaRDI's infrastructure
- Number of data sets certified by the MaRDI badge
- Number of edits in the MaRDI knowledge graph (WikiBase)
- Completed deliverables and milestones from the task areas in the work program
- Number of participants in workshop, events, tutorials, and carpentries
- Number of events and workshops co-organized by MaRDI
- References to MaRDI data sets
- Number of publications adopting MaRDIs implementation of FAIR principles for mathematical research data
- Advancement in gender equality

We do not use specific indicators tailored to a individual task area, but rather evaluate the success of MaRDI as a whole.

3 Consortium

3.1 Composition of the consortium and its embedding in the community of interest

The subject of MaRDI is research data in mathematics and its application in other disciplines. Traditionally, the notion of mathematical research data has been focused on *information retrieval* or *catalogue services* for specific topics or general services like zbMATH and swMATH. However, recent discussions have broadened the perspective on research data in a wide interdisciplinary area. For example, mathematical models have been introduced as an important mathematical research data category [KTK16; KT16] as they are utilized in a wide array of disciplines. This idea initiated research on adequate representations of mathematical models which are both, human-understandable and machine-actionable, [Koh+17; Kop+18] and was sharpened by feedback from the mathematical and applied communities; see, e.g., [Kop+17].

At the end of 2018, an in-depth meeting of WIAS representatives at FIZ Karlsruhe including also an EMS representative identified the need for strategic developments of a mathematical research data management infrastructure on a national as well as international level. This led to the formation of the MaRDI group in January 2019 at WIAS spanning a wide range of institutions, representatives of mathematics communities and societies, like DMV, GAMM, and GOR, and research groups. The meeting (and regular follow-up MaRDI meetings) explore several dimensions: (i) computational mathematics including computer algebra, scientific computing, and statistics; (ii) mathematical research and other scientific disciplines; (iii) the scope of mathematical research data; (iv) current as well as future-oriented information retrieval services, based on in-depth indexing as well scientific quality assurement; (v) research driven services. In particular, the areas in (i) traditionally deal with mathematical research data and have a history of developing corresponding management strategies and services within their respective mathematical user communities. Thus, they are excellent prototypical development areas in MaRDI. The discussions on (ii) highlight the wide spread applicability and penetration of mathematical results and techniques across a wide array of scientific disciplines. (iii) shows the dire need to expand the current scope of mathematical research data. Finally, mathematics builds on strongholds such as zbMATH in connection with information indexing and retrieval in (iv), but faces tremendous challenges through the vast amount of nowadays and future data, open access and open research demands, and the need for establishing continued research enabling data driven services in (v).

The MaRDI group regularly presents its state of understanding of mathematical research data management in mini-symposia and panel discussions at the Annual Meeting of the DMV reaching out its target user groups. The connection to the Germany wide DMV community has already led to a targeted exchange including community feedback and a plan for future activities. The orientation of MaRDI was also presented at a Round Table Meeting for Applied Mathematics at the *Deutsche Forschungsgemeinschaft* in October 2019. There is an understanding with GAMM that GAMM-related consortia (in particular, MaRDI and MatWerk) will be invited to present their activities during special tracks at the GAMM Annual Meetings (usually, those are attended by about 1200 people). Together with the involvement of GOR on its topic MaRDI will reach large parts of the mathematical community

(in Germany and beyond). By incorporating the current mathematically oriented Clusters of Excellence MaRDI maintains strong connections to cutting-edge research. Furthermore, from the Leibniz network "Mathematical modeling and simulation" (MMS)[3], coordinated by WIAS, several institutions committed strong interested on MaRDI's development.

MaRDI submitted a proposal in the first round of funding calls within the NFDI. The proposal was not successful, however, the very encouraging feedback has led to a new structure for the MaRDI consortium, which now explicitly covers a new Task Area T3: Statistics and Machine Learning, a generally highly interdisciplinary topic which otherwise we think would not yet be well represented within the NFDI, especially its mathematical aspects. MaRDI furthermore decided to leave out highly risky, yet innovative developments with respect to a deepened view on FAIRness of research data (deepFAIR), which still require research progress in their own right. These latter projects will now be considered as part of MaRDI's perspective and are connected with MaRDI through Prof. Michael Kohlhase (FAU). Results from this ongoing research will be closely monitored and integrated at later stages where appropriate.

A further milestone of development towards Open Science with very positive implications for MaRDI was, that at the beginning of 2020 the mathematical information retrieval service zbMATH at our co-applicant FIZ Karlsruhe started to turn Open Access. Furthermore, it now provides an Open API for machine-integration enabling new services based on the platform and broadening of MaRDI services in general.

The research motivated pillars of the MaRDI-proposal are based on the rough division of mathematical research data in *exact and symbolic data* (T1: Computer Algebra), *floating point data* (T2: Scientific Computing), and *data with uncertainty* (T3: Statistics and Machine Learning). Furthermore, we take into account the interdisciplinary role of applied mathematics by a specific task area T4: Cooperation with Other Disciplines. Doing so, MaRDI pursues the representation of its user communities by design. The tasks for this proposal have been identified and formed according to the respective *readiness* of preliminary work and services. Such a readiness indicator very well reflects the consolidated needs of the mathematical community. We expect that these flagship projects (task areas) will develop a pull effect on further parts of the mathematical community and also at an international level with regard to discussing their needs with respect to mathematical research data. Through the direct integration of the national mathematical society, DMV, see task area T6: Data Culture and Community Integration, we aim for the full feedback and dissemination cycle within the mathematical community and beyond.

Within MaRDI, users and providers are represented through institutions and professional societies representing the entire mathematical landscape. Specific services will be hosted and provided as well as used by institutions like WIAS, ITWM, LMU, MPI DCTS, MPI MIS, TUB, TUK, TUM, USTUTT, WWU, ZIB (see Section 1 for a detailed explanation of the abbreviations). Furthermore, FIZ Karlsruhe and ZIB will provide the one-stop T5: The MaRDI Portal for these services for the mathematical community as a whole and potentially other applied discipline relying on mathematics. DMV, EMS, GAMM, GOR, and MFO with its state-of-the-art meeting facilities and scientific agenda is of great importance for the contact to the national and international mathematical top-level research and for

[3] https://www.wias-berlin.de/research/Leibniz-MMS/

the dissemination of services and best-practices. Typically, users will develop pilot solutions that can be transferred into sustainable services integrated into the MaRDI Portal.

Composition of consortium

The **Weierstrass Institute for Applied Analysis and Stochastics (WIAS)** conducts project oriented research in applied mathematics and coordinates the "Leibniz-Network on Mathematical Modeling and Simulation (MMS)", which involves Germany-wide more than thirty institutions from a wide range of scientific disciplines. It also has a research focus on establishing mathematical models as research data and advances open access as well as open science policies. Further, WIAS is one of five co-operation partners of the Berlin Cluster of Excellence **MATH$^+$** (EXC 2046) in Mathematics, and it hosts the permanent Secretariat of the International Mathematical Union (IMU) as well as the Secretariat of the German Mathematical Association (DMV). Through these activities and connections to a large and versatile user community, it is the ideal institution for coordinating MaRDI. Moreover, its research agenda, application orientation and drive for scientific software innovation qualifies the institute to support the consortium's agenda in task areas concerning use cases with other disciplines, standardization of mathematical models, and user services.

Deutsche Mathematiker-Vereinigung e.V. (DMV), founded in 1890, is a professional organization representing mathematicians in Germany. With its currently more than 4500 members, it takes active part in all aspects of research and teaching and fosters international cooperation. It is a founding member of the European Mathematical Society (EMS). The DMV will support MaRDI in developing and establishing standards, services and workflows for mathematical research data by activity groups which help to address the needs of the mathematical community, training of mathematicians in good practices of data handling and broad dissemination through conferences and publications like the "DMV Mitteilungen". This will help to establish data culture in the mathematical community.

FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH (FIZ) makes significant contributions to sustainable information infrastructures, thereby supporting researchers in science, humanities and industry worldwide. Its mission is to research, develop and operate methods, processes and services. FIZ offers data, information and knowledge, software and services via open and legally compliant platforms by making them searchable, accessible, interoperable and reusable. Through its goal to support the value creation process in science and innovation at all levels, it fits well into the consortium at its information infrastructure side. In the area of mathematics, FIZ Karlsruhe develops and maintains various information services acknowledged and used by the community worldwide. Their core features are focused on facilitating indexing and retrieval of bibliographic information, reviews, free full texts, and software, notably zbMATH, EuDML, eLibM, and swMATH. FIZ Karlsruhe brings in expertise built up for processing, standardizing, storing and indexing mathematical (meta-)data which can be transferred to research data. Its Research Data Repository RADAR is a customized, cost-efficient and user-friendly service for the archival and publication of research data.

The **Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM)** in Kaiserslautern is one of the largest research institutes for industrial mathematics worldwide. Its mission is the development, implementation and application of mathematical methods for modeling, simulation and optimization of products, processes and services for business and society. In the High Performance

Center Simulation- and Software-based Innovation or the Science and Innovation Alliance Kaiserslautern, science, application and industrial partners, in particular small and medium enterprises, are brought together. The Fraunhofer ITWM does not only build a bridge between the real and the virtual world, but also between the academic mathematical research at universities and its practical application in other disciplines and industry.

Freie Universität Berlin (FUB) is a leading research institution. It is one of the 13 German universities being funded through the German government's Excellence Strategy and is part of the only University Consortium of Excellence, the Berlin University Alliance. FUB has 12 departments and almost 40.000 students. The Dahlem Research School promotes the training and development of junior researchers. It cohosts the Berlin Mathematics Research Center **MATH$^+$** (EXC 2046) that has recently launched the research area "Mathematics of data science". it is the home of the DMV Medienbüro.

Ludwig-Maximilians-Universität München (LMU) is the leading teaching and research university in Germany, ranking 1st in Germany in the latest Times Higher Education World University Ranking. Attesting to the excellence of its research, LMU Munich has been one of the most successful German universities in the Excellence Initiative, having been funded in all three funding lines since the beginning of the Initiative in 2006. It is currently involved in 43 Collaborative Research Centers funded by the German Research Foundation (DFG), and coordinating institution in 15 of them. LMU Munich offers 43 structured doctoral programs in a broad range of disciplines. As one of the leading statistical research centers in Germany, the LMU Department of Statistics is renowned for its development and application of statistical methods based on the interplay between experimental and theoretical work. In particular, the Chair of Statistical Learning and Data Science offers expertise at the cutting edge between machine learning and statistics and is part of the OpenML project, which focuses on the development of a collaborative open research platform for machine learning that can be easily integrated into the objectives of the MARDI initiative.

The **Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO)** is an international research institute with a main focus on mathematical research, scientific collaboration and training in mathematics and its related areas. Thereby, the promotion of young scientists plays an important role. Every year, approximately 2500 guest researchers who are leading experts in their field participate in the weekly changing workshop program as well as in the "Research in Pairs" program for smaller research groups. It has strong connections to and is a minor shareholder of the non-profit company IMAGINARY, an open source platform for interactive and participative math communication.

The mission of the **Max Planck Institut for Dynamics of Complex Technical Systems (MPI DCTS)** in Magdeburg is to develop mathematical models capable of describing complex chemical, biotechnological and energy-related processes and to analyze the system properties and dynamic behavior of these processes using the models. After validation, these models should be used to design and control efficient and sustainable production processes. The objectives also cover the establishment of new computational methods and advances in systems and control theory.

The **Max Planck Institute for Mathematics in the Sciences (MPI MiS)** in Leipzig is a world-class research center that also runs an extensive visitor program. It has the mission to foster the careers of young scientists and has a proven track record of leading PhD students and Postdocs to the next

career level. Due to this perpetual stream of incoming young scientists and international guests it harbors a constant opportunity of direct interaction with many mathematicians in all career stages, but especially the next generation of mathematicians. Additionally, the MPI MiS has a longstanding expertise of hosting many workshops and conferences every year. The institute's 'Eberhard Zeidler Library' will also support MaRDI by contributing its professional network, its services, resources and the expertise of an outstanding mathematical library.

With almost 34,000 students, about 100 course offerings and 40 institutes, the **Technische Universität Berlin (TUB)** is one of Germany's largest technical universities. Freie Universität Berlin, Humboldt-Universität zu Berlin, and Technische Universität Berlin, along with Charité-Universitätsmedizin Berlin applied successfully in the German government's Excellence Strategy competition under the name Berlin University Alliance. The Excellence Cluster **MATH$^+$** (EXC 2046) and its Graduate Program "Berlin Mathematical School" are directly relevant to the proposal. Further, TUB hosts DepositOnce, a repository for research data and publications.

The **Technische Universität Kaiserslautern (TUK)** is a research university with a focus on engineering and natural sciences and an international profile. TUK's application-oriented research is moreover pursued in close cooperation with the ten research institutes on and close to the TUK campus, like the Fraunhofer Institute for Industrial Mathematics (ITWM). With respect to its research strongholds TUK has identified six core research areas, two of which "Mathematical Modeling, Algorithms and Simulation" and "Digital transformation in Economies and Societies" immediately relate to the MaRDI objectives.

The **Technical University of Munich (TUM)** is one of Germany's leading universities and was selected as a University of Excellence in each one of Germany's excellence initiative competitions. TUM educates over 42,000 students across 15 departments and 6 integrative research centers. MaRDI will be actively supported by the TUM Department of Mathematics, which represents a wide range of research areas and is internationally recognized for its excellence in applied mathematics.

The **Universität Stuttgart (USTUTT)** is a leading German technical university. Basic research that is both insight-oriented and application relevant is one of its keys of success. Endorsed by their data policy, the University of Stuttgart expressly promotes and supports the free access to research data and established the Competence Center for Research Data Management FoKUS, which supports all phases of the data life cycle by a technical infrastructure as well as consulting and training services. The University's role within MaRDI is the interface between applied mathematics and engineering, biochemistry and material science and its expertise and infrastructure for interdisciplinary research data management. This will be coordinated within the Stuttgart Center for Simulation Technology **SimTech** (EXC 2075).

The **Westfälische Wilhelms-Universität Münster (WWU)** is one of the largest comprehensive universities in Germany. It is internationally recognized for its top-level research, and it hosts the cluster of excellence **Mathematics Münster: Dynamics, Geometry, Structure**. It is a stronghold for mathematical research covering the whole range from theoretical to applied mathematics. The WWU is shaping digitalization strategically and in a coordinated manner, especially with regard to digital infrastructures. A central element here is the management of research information as the basis of many digital infrastructures and processes. Through theses expertises and focus areas, concerning

MaRDI it contributes to the area of scientific computing with an emphasis on verified research data in mathematics and its applications, FAIR-principles for computer based experiments entailing also data, confirmable workflows for trust worth computations and dissemination.

The **Zuse Institut Berlin (ZIB)** is an interdisciplinary research institute for applied mathematics and data-intensive high-performance computing. Its research focuses on modeling, simulation and optimization with scientific cooperation partners from academia and industry. The institute is located at the interface between mathematical method development and the analysis and management of large data sets. It is a link and a communication amplifier between the applied sciences and mathematics. The ZIB is active within the framework of the Berlin based Research and Competence Center for Digitalization (digiS), an institution for cross-disciplinary digitalization projects. With FIZ, it develops and maintains swMATH, a freely accessible information platform for mathematical software, which contributes also to the FORCE11 Software Citation Implementation Group.

The **European Mathematical Society** is the umbrella organization of almost all European mathematical societies. Currently, it has about 5,800 individual members, and employs several committees which address all diverse relevant aspects of mathematical research and teaching. The quadrennial European Congress of Mathematicians is the largest meeting of mathematicians in Europe, and numerous joint events with partners worldwide foster international cooperation.

The **Excellence Cluster EXC 2046 "MATH$^+$ - Berlin Mathematics Research Center"** has a research focus on application-oriented data-driven modeling, simulation and optimization and fosters interdisciplinary projects. It integrates the excellence graduate school "Berlin Mathematical School" (BMS). Math$^+$ runs thematic Einstein semesters which bring international top-level researchers to Berlin.

The **Exzellenzcluster EXC 2075 "Data-integrated Simulation Science" (SimTech)** constitutes a long-standing prime example on establishing and structurally supporting interdisciplinary research. It is dedicated to the goal of developing and evolving simulation technology from isolated numerical approaches of individual disciplines to an integrated systems science. This is achieved, for example, by simultaneously increasing the performance, accuracy, precision, and validity of computer simulations of mathematical and data-driven models. Since 2007 SimTech has a pioneering research data management.

The **Exzellenzcluster EXC 2181 "STRUCTURES: A unifying approach to emergent phenomena in the physical world, mathematics, and complex data"** explores new concepts and methods for understanding how structure, collective phenomena, and complexity emerge from the fundamental laws of physics. These concepts are also central for finding structures in large datasets, and for realizing new forms of analogue computing.

The **Fraunhofer-Verbund IUK-Technologie** is the largest provider of applied research in the field of information and communications technologies in Europe. It marshals key expertise for business and society to utilize in exploiting opportunities and meeting the challenges that result from the comprehensive digitalization of virtually all aspects of today's new world.

The **Gesellschaft für Angewandte Mathematik und Mechanik e.V. (GAMM)** fosters the scientific development in all areas of applied mathematics and mechanics. As far as applied mathematics is concerned, GAMM is traditionally embracing areas like numerical analysis and scientific computing,

computational science and engineering (CSE), optimization, uncertainty quantification (UQ), computational materials science, which are data and software intensive and nowadays very much involved in producing FAIR research data.

The **Gesellschaft für Operations Research (GOR)** represents researcher working in operations research as well as companies which apply methods of operations research for solving real-world problems. The discipline is on the boundary between mathematics, computer science and business administration.

The **IMAGINARY gGmbH** creates and distributes interactive exhibits that communicate current research of mathematics. As a non profit organization it promotes mathematics education and knowledge worldwide by offering interactive software, hands-on exhibits and trainings under open licenses. IMAGINARY has been working in setting up digital infrastructure to promote mathematical research through open source projects as Hilbert (dynamic museum infrastructure) or its open platform for mathematical exhibits https://imaginary.org. IMAGINARY has strong expertise in communicating scientific information and relies on its well-established communication channels to inform about the newly developed standards, services and workflows.

The consortium at its current state has the expertise to fulfill the proposed work program. For new topics emerging by the strategic development within the Consortium Council (CC) MaRDI will identify the required expertise and potential new partners.

3.2 The consortium within the NFDI

Mathematics as a common scientific language is a basic prerequisite for the theory development in many other disciplines. Furthermore, cross-disciplinary methodologies like mathematical modeling and simulation (MMS), statistics and data science are used everywhere in science. Within the NFDI, MaRDI aims to establish common standards for research data as well as meta data schemas related to mathematical models, numerical algorithms and software, benchmarks and workflows and all involved mathematical objects. This includes model parameters and model data from other disciplines used as input data for simulation studies or data analysis.

On the other hand it is important to link the mathematical concepts to their domain-specific context and semantics. The same mathematical object or model can occur in completely different contexts, having different semantic meanings. By storing the discipline-specific semantics we can contribute to a terminology service which helps to translate between the disciplines making results, methods and the involved data better findable, accessible and re-usable. This unique contribution to the NFDI will enable researchers from other disciplines to link their research data to mathematical concepts, and vice versa, to allow mathematicians to find application domains, test cases, and benchmark problems for their own research activities. Within NFDI, MaRDI will play the role of an integrating factor with respect to modeling and simulation in all disciplines. Thus, MaRDI will contribute to one of the main goals of the NFDI: to enable posing entirely new research questions and to foster innovation.

There are a number of cross-cutting topics of the NFDI which are relevant for MaRDI, among them one finds:

- Authentication and Authorisation Infrastructure (AAI)
- Cloud-based Data Processing
- Knowledge Graph and Semantic Technologies
- Federated Repositories
- Metadata for Scientific Software, Workflows, Computer-Experiments
- Journal Integration / Next-Generation Peer Review
- Author Identification (Data, Software, Publications)
- Data Culture and Training

MaRDI will provide the MaRDI-Portal **T5** as a one-stop unique and easy to access contact point for the scientific community to retrieve and consult specific MaRDI services, which will be distributed among MaRDI partners. The heterogeneous nature of the involved mathematical data requires viable Federated Repository infrastructure involving sophisticated data descriptions. On a technical level MaRDI will share its expertise in managing high volumes of heterogeneous data as well as in developing web-services with high user frequency as well as their secure and sustained 24/7 operation, like zbMATH and swMATH, to other NFDI consortia, who are also aiming for the creation of central portals.

The reproducibility crisis and the ever-growing amount of data and use of computational methods in science as well as mathematics has lead to the urgent need of establishing Next-Generation Peer Review and to provide guidelines for Journal Integration. MaRDI will contribute to this topic in collaboration with a number of mathematical scientific journals especially in **T1** and **T3** as prototype before working on distribution widely within the mathematical publication community.

Metadata for Scientific Software, Workflows, Computer-Experiments is key for the reproducibility of scientific results especially for computer-assisted mathematics. Therefore, MaRDI will dedicate substantial effort to the definition of suitable metadata for the related research data objects. This will not only allow for the reproducibility of scientific workflows but also for the re-usability of established models and processing pipelines in **T1**, **T2**, **T3** and **T4**.

The Knowledge Graph and Semantic Technologies are not only important for the distribution of mathematical knowledge within the mathematical community but also within the other disciplines that rely on mathematical objects, models, workflows, and software. MaRDI will therefore build up an agile knowledge graph to provide reliable knowledge sources and to prevent ”re-invention of the wheel” in a scientific project. In this direction, the MaRDI partners have already established several services that will be integrated into the knowledge graph and that will be built upon. Last, but not least, a cross-cutting concern is the training of the research community in the NFDI standards, data sets, services, and workflows to establish a mathematical research Data Culture in **T6**.

The cooperation with other disciplines and other NFDI consortia is established via a dedicated Task Area T4: Cooperation with Other Disciplines; see page 72ff. MaRDI co-applicant institutions are active in NFDI4Culture (funded), NFDI4Chemistry (funded), NFDI4MatWerk, NFDI4Memory, NFDI4Objects, NFDI4DataScience, NFDI4Agri (2020 proposals), NFDI4MobilTech, NFDI4Phys (2021 proposal) via it co-applicant institution FIZ Karlsruhe. NFDI4Objects, and Text+ (2020 proposals) via ZIB, NFDI4Cat (funded) via MPI DCTS, and BERDNFDI via LMU.

3.3 International networking

The MaRDI consortium is strongly embedded into other German and international initiatives in the context of the FAIR treatment of research data. In its capacity as MaRDI participant, the European Mathematical Society (EMS), *the* roof organization of all mathematical societies in Europe (not only the EU), will act as a coordinator of similar activities in Europe and also worldwide. This helps to avoid duplication or conflicting work and to create internationally accepted standards. The mathematical community is widely connected internationally, with the EMS being a key player in all of Europe and also acting on the world scale within the International Mathematical Union (IMU), with its permanent Secretariat hosted at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS), and the International Consortium for Industrial and Applied Mathematics (ICIAM). In the forefront of this proposal, already collaborations with other mathematical organizations in Europe and worldwide have been discussed and shaped. EMS will also take the lead in the dissemination of the results of the MaRDI project to the international partner. When it comes to open research and open access policies, MaRDI aspires, through its connection to the EMS Publishing House, the realization of FAIR information retrieval (such as, e.g., mathematical formulas, potential connection to discipline specific semantics, etc.) from initially selected publications, but with the clear longterm goal of a comprehensive realization of mathematical information retrieval from publications. In connection with IMU, MaRDI will have an outreach to a large number of national mathematical societies (like the DMV for Germany). This generates an enormous multiplication potential for dissemination of MaRDI outcomes and user feedback in all fields of mathematics.

MaRDI is also tightly connected to the Society for Industrial and Applied Mathematics (SIAM), which is based in Philadelphia, USA. Through this connection, MaRDI expects feedback from a worldwide community of applied mathematicians, computer scientists, engineers and researchers in quantitative life sciences, and other quantitative scientific fields. This will occur through direct connection to SIAM governance bodies as well as MaRDI's presence at SIAM Annual Meetings as well as meetings of various fields covered by SIAM with the SIAM Conference on Computational Sciences and Engineering being perhaps the largest instance. SIAM, as a worldwide renowned non-profit organization, has been actively pushing the FAIR principles. On the computational science side it has in particular guided its community to reproducibility [SBB13; Her19]. This also includes building a community culture by developing teaching modules for securing reproducible results, another area connecting to MaRDI.

Through its co-applicant FIZ Karlsruhe, MaRDI benefits from the cooperation with zbMATH and swMATH. zbMATH is *the* primary resource for mathematics when it comes to information indexing, retrieval and accessibility, e.g., on published work. It will be open access starting from 2021. zbMATH maintains a very high quality standard by involving international review experts for assessing all scientific articles in zbMATH's data base (complete since 1886). This large worldwide network of experts will be accessible to MaRDI and provides a multiplier effect in both dissemination as well as user feedback. MathSciNet, run by the American Mathematical Society, is similar to zbMATH (though it covers a shorter period of time), and both services cooperate in developing and maintaining the Mathematical Subject Classification. Through zbMATH, MaRDI also seeks to connect to MathSciNet.

MaRDI will establish a close collaboration with OpenAIRE which is the Open Science pillar of the EOSC. By exposing metadata about software and publications and their links (via standard APIs and formats), the MaRDI knowledge graph will become an OpenAIRE data source. This will include the integration of swMATH data to OpenAIRE, see specifically **T5**. Through this MaRDI is giving more visibility to its services to all OpenAIRE customers and benefits from the functionality and additional information available in OpenAIRE. EMS as one of MaRDI's participants plans to become a member of EOSC and will thereby strengthen the connection between the consortium and EOSC. Furthermore, the metadata standards in MaRDI will be aligned to other source, like WikiData or Zenodo, to enable data donation to other services.

Through its Research Group on Numerical Mathematics and Scientific Computing, WIAS, the MaRDI coordinating node, is connected to FORCE11 (https://www.force11.org), an international community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. In particular, it pursues the goal of making research data "Findable, Accessible, Interoperable and Reusable" [Wil+16; DKK18]. This embeds MaRDI into an international forum for fostering the realization of the FAIR-principles in science, and - under the guidance of MaRDI - it is foreseen to establish a FORCE11 acknowledged working group on mathematical models.

While several networking strategies have already been mentioned above, we highlight two further directions. One objective of MaRDI is the development of metadata standards and schemes for mathematical objects and correspondingly persistent identifiers (PIDs) allowing their citation. On this topic, we plan a close collaboration with DataCite, the global leading provider of persistent identifiers (PIDs) in research. Mathematical research data is closely connected to international initiatives aiming at the creation of a comprehensive mathematical knowledge base, like the International Mathematical Knowledge Trust (IMKT), or the Digital Library of Mathematical Functions (DLMF) developed by National Institute of Standards and Technology (NIST), USA. Here, FIZ cooperates with IMKT on standardization and with NIST on the integration of DLMF into zbMATH services, which will be also available in the MaRDI-portal.

MaRDI will incorporate international knowledge by including experts within the advisory board and users via the User Forum. There, the European Open Science Cloud with ESCAPE and Ex-PaNDS as well as connections to the Exascale Computing Project are of particular interest. Through its co-applicants, MaRDI can rely on experience from previous projects, and it developed infrastructure for collaborative work, the inclusion of mathematical software into notebooks, or the distribution of software containers for cloud computing. MaRDI will work on this with respect to reproducible computer experiments.

3.4 Organisational structure and viability

Within the task areas the strategic work packages of the consortium are completed. Each task area has an elected spokesperson who represents the task area in the MarDI Consortial Board and the Council. In this context, **T7** holds governance responsibilities for the entire consortium. In this endeavor it will be supported by several internal decision boards; see Figure 10 on p. 106. The various

roles of the boards are described in detail in T7 below. Major board tasks comprise strategic topics oriented decisions, supervision of task area work and work across the task areas as well as the use and distribution of funds. The Consortial Board also runs a Technical Committee whose task is to certify services which shall be launched in MaRDI's digital portal, or to provide guidelines for MaRDI service developers on how to provide certified products and services. This represents another level of quality management in MaRDI. It should also be emphasized that standardization and interoperability of data and services are taken care of by a dedicated board, which, through the expertise of FIZ, will also team up with other NFDI consortia in order to secure a high level of interoperability for the entire NFDI. Further, MaRDI seeks advice from experts with various backgrounds (including science, decision and policy making), subsumed in an Advisory Board.

The integration of users takes place on several levels. First of all, the task areas T1-T4 not only reflect different data types, but they also directly connect to broad user groups within mathematics, but also in other scientific disciplines. The latter is in particular related to the many NFDI interactions that are at the heart of T4. The tight relation of these task areas to the state-of-the-art in their respective research area is of paramount importance for the viability of MaRDI. Further, through the FAIR-data orientation and interoperability agenda, the development benefits the entire NFDI-user groups in quantitative fields. Further, MaRDI foresees a dynamical integration of users on an operational level through manifold activities ranging from developments in T6, over forming alliances with scientific societies to conference participation. On the more organizational level, MaRDI runs a User Forum for community feedback concerning service orientation and to shape the strategic development. On all levels, members of other cooperating NFDI consortia are invited to participate with a guest status. In this way, the NFDI network concept is strengthened and joint use cases as well as cooperation models and services will be explored via direct integration of partners.

This board structure implements the MaRDI quality management and it will enable discussions and communication of strategies on a governance as well as organizational level. MaRDI will also hold regular general assemblies which are open to all MaRDI participants and team members. This fosters a direct communication of information MaRDI wide and it enables community building within MaRDI.

The budget will be allocated at and distributed by the applicant institution WIAS. The Council, together with the Advisory Board, reviews and sets the strategic goals for MaRDI and its task areas over typically an at most two to three years (internal) project period during MaRDI's run time. Then the task areas need to formulate corresponding work plans including timelines, work packages and milestones in project format. Based on these project applications and their critical review, budget allocations are decided by the Council. Travel and internal cooperation funds are also distributed correspondingly to the project hosting institutions.

Concerning membership status of individuals and organizations, MaRDI's governance will also foster consortial viability by integrating (through the MaRDI portal) further research data programs such as, e.g., an envisaged corresponding FAIR data agenda of the Journal of Statistical Software or DFG-project driven activities, to mention only a few. In particular, MaRDI envisages associated memberships for researchers who run third-party funded research projects which have an interlinked agenda with MaRDI. Ideally, this will lead to agreed and quality secured, autonomous integration of services into the MaRDI portal.

Further, for the viability of the consortium T7 will hold some strategic funds for an agile development, MaRDI-wide workshops, conferences and symposia. These agile funds will be allocated according to short-term or strategic project needs, to follow new developments as they appear during MaRDI's runtime and to reinforce, if needed, critical work packages. Specific examples for the use of the agility funds include: (-) a project for expanding math-web-search along with its portal integration which can only be started once the MaRDI portal has reached a sufficiently mature state; (-) programming and data integration support for the agenda on next generation peer review and journal integration in conjunction with the EMS Publishing House; (-) the extension of the algorithm data base from objects maintained by T1-T4 to a connecting middle layer between zbMath and swMath and its integration into the portal; (-) the support of increasingly autonomous, quality secured service integration into the MaRDI portal through associated members and active user groups–this project can only be started once the portal runs in a stable and robust mode which is anticipated between year 2 or 3 of the MaRDI project period. These projects are only four examples which are already now identified as projects which need to be implemented once MaRDI has reached the respectively necessary level of maturity.

This envisaged overall administration of strategic topics, budget distribution based on a rigorous quality management, and community building (internally and externally with users) guarantees the viability and agility of MaRDI.

MaRDI's viability will also be secured through monitoring and properly reacting to KPIs as listed above in Section 2.2.

Finally, MaRDI pays attention to EOSC compatibility of its agenda and work results. This will be partly helped by the EMS as a communicator in both directions, i.e., from MaRDI to EOSC and vice versa. This will couple MaRDI with the single forum envisaged for EOSC which combines EOSC stakeholders such as research funders, policymakers, research performing organizations and operators of research infrastructures in order to actively contribute to and monitor the future EOSC developments.

3.5 Operating model

MaRDI will set out with an operating model that will rely on a cooperation agreement between the co-applicants' as well as participants' institutions. This agreement will regulate the regular proceedings of MaRDI, the ways of transfer of funds, and it will fix the in-kind contributions by the respective institution. Moreover, it will secure the institutional inputs into MaRDI in the sense that a robust continuation of MaRDI's strategic work is guaranteed even if one of the partners would unilaterally discontinue the cooperation agreement. This is aimed to underline the spirit of the overall NFDI to establish a robust, agile and permanent structure. In particular, the MaRDI consortium will not only adhere to the FAIR principles, but it will work as a non-profit / public-benefit operation. Through its proceedings it is entirely devoted to fostering Open Access, Open Research and Open Science policies, services and products, with the typically indicated legal and financial approach towards interests from potential commercial customers. This approach ensures an integration into a common NFDI governance structure, such as a public-benefit association, at any point in time as outlined in the Articles of Association of the

”NFDI-Verein” currently being in founding.

The participating institutions may provide services to the MaRDI portal, which will be run as a one-stop contact point for MaRDI services and products. Many of the basic services (e.g., on information retrieval) will be easy access services; selected other ones may require registration to receive login data. All services and products will be run on a non-profit / public-benefit basis.

MaRDI pursues an open and dynamic, yet quality assuring, membership policy. In particular, researchers or institutions running (third-party funded) quality assured projects with a mathematical research data agenda adhering to the FAIR principles and interlinked with MaRDI may become non-MaRDI-funded members. These members may want to insert services and links to data bases into the MaRDI portal, and they participate in MaRDI's general assembly. This fosters viability, keeps pace with current research, and dynamically expands the scope of MaRDI.

Concerning users and providers of services and products, MaRDI benefits from an “integrated” consortium. This means in particular that co-applicants provide services throughout the task areas, with **T5** (involving scientific information providers such as FIZ-Karlsruhe and ZIB) providing the digital portal of MaRDI. Among the MaRDI participants one finds a nationally as well as internationally widespread user community (DMV, EMS, GAMM, GOR) which is even further enhanced through strategic partnerships of MaRDI with Clusters of Excellence, the interdisciplinary Leibniz-Network “Mathematical Modeling and Simulation” as well as close links to a number of other NFDI consortial initiatives. Also, the co-applying institution Mathematical Research Institute in Oberwolfach (MFO) is worldwide *the* place where mathematicians meet and exchange. This is a unique way and opportunity of exchange with the worldwide mathematical user community. This will not only help to shape MaRDI services in a user-centric way, but it will foster the wide-spread development of a FAIR community culture and a rapid dissemination of MaRDI findings. Strategically, users provide proof-of-concept mathematical data/objects/workflows or use cases which will be converted into service pilots in a joint venture between users and providers. Finally, providers expand the pilot projects into fully fletched services which, after passing a certification check, will be offered through the MaRDI digital portal. MaRDI's information retrieval services and software database goals benefit from zbMATH and swMATH, respectively, two services of FIZ-Karlsruhe of worldwide renown in the scientific community with connections to mathematics.

The consortial co-applicants provide a wide span of in-kind contributions see the summary in Section **??**. These range from (long term) positions for personnel, technical equipment facilities to digital solutions for basic IT services and components, as well as the provision of conference facilities and services.

4 Research Data Management Strategy

As we already detailed in Section 2, mathematical research data are multifaceted and heterogeneous. Within mathematics as a discipline it emerges in form of specific entities and objects required for the advance of mathematical research, but it is also highly relevant in other scientific disciplines that rely on mathematics as language of abstraction. In order to set the data specific stage for this proposal,

in the following section we describe typical examples of mathematical research data, and we explain the state of the art in data handling by using the relationship between the mathematical research process and the data life cycle. The pertinent available infrastructures and services that constitute the starting point for MaRDI's research data management strategy are summarized. In order to realize the FAIR principles for data in mathematics and to propel new research in mathematics, but also in other disciplines, MaRDI is designed to follow a layered architecture which is relevant across all task areas. This architecture consists of the four layers **X1: Core**, **X2: Data**, **X3: Exchange** and **X4: Knowledge**, whose interplay along with the added value generated for developers, researchers and users through this architecture is explained by a simple guiding example, namely a benchmark framework for linear system solvers. We identify the synergy topics **S1: Algorithms**, **S2: Mathematical Models**, **S3: Workflows** spanning over multiple fields of mathematics and which will be considered in MaRDI from a strictly mathematical viewpoint. Based on use cases from experimental and computational mathematics, mathematical modeling and simulation, and statistical data analysis and machine-learning we discuss the typical problems and limitations of todays research data management and how MaRDI can address these problems by its work program. The section is completed by a discussion of metadata standards, the implementation of the FAIR principles and quality assurance as well as services provided by MaRDI.

4.1 State of the art and needs analysis

Current practices for mathematical research data are very heterogeneous (domain-specific) and have differing degrees of formality, accessibility and interoperability, depending on the various data sources and data usage. For MaRDI the following typical sources are most relevant:

- **Mathematical documents** in PDF, LaTeX, XML, MathML, etc.
- **Notebooks** e.g., in Jupyter or Mathematica.
- **Domain-specific research software packages and libraries** like 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** like SageMath, SINGULAR, Macaulay2, GAP, polymake, Pari/GP, Linbox, OSCAR, and their embedded data collections.
- **Programs, and scripts** written in the packages and systems above, and in systems not developed within the mathematical community; but also input data for these systems like algorithmic parameters, meshes, mathematical objects stored in some collection, the definition of a deep neural network as a graph in machine learning etc.
- **Simulation data**, usually series of states of representative snapshots of the system, discretized fields, more generally very large but structured data sets 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, Mizar, NASA PVS library.
- **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

- Descriptions of **mathematical models** in mathematical **modeling languages**, e.g. 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.

These artifacts of research data are created, transformed and re-used within the mathematical research process. This constitutes the data life cycle characterized by the computation of individual facts, observation of common structure in this "experimental data", derivation of a conjecture and formulation of mathematical statements, identification of appropriate arguments by exploration of the existing theory as well as by introducing novel techniques, exploration of the validity of the conjectures or their limits by further experiments, organization of the argumentation in a coherent statement or proof following the current mathematical standards, and publication of the results to the community for application. With MaRDI we aim to improve the current practice in handling these inter-linked, complex, diverse and multi-faceted mathematical research data along the whole research workflow as depicted in Figure 2.

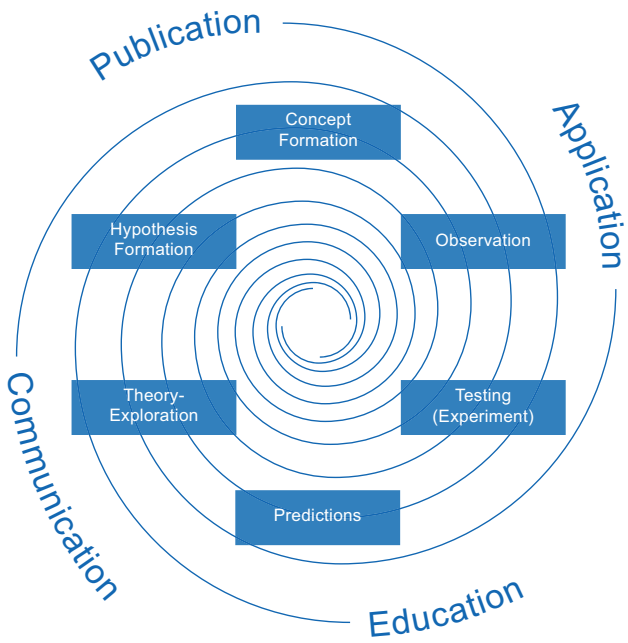


Figure 2: The data life cycle in mathematics. Mathematical research data is generated and re-used at all stages of the mathematical research workflow.

In mathematical modeling and simulation (see **T2** and **T4**), this workflow spans from a real-world problem, its formulation in terms of a mathematical model, the development of a solution method for that model by exploring the mathematical knowledge and the validation of the model predictions against experimental data. In practice, mathematical research data are complex, diverse and multi-faceted. It can be abstract like mathematical models and also very concrete such as visualizations, and it may consist of interlinked objects of different typesThis is especially true for the emerging field of scientific machine learning in which complex multi-scale multi-physics dynamics are combined with machine learning techniques, see **T3**.

Currently, different infrastructures or services exist for mathematical research data. We illustrate this through selected examples. A more comprehensive list of sources related to the partners of the consortium can be found in the appendix. With respect to mathematical research software the *swMATH* database[4], maintained by FIZ and ZIB, establishes a connection between scientific publications and mathematical software. Dedicated to the exploration of mathematical knowledge in

[4] https://www.swmath.org

publications *zbMATH*[5], edited by FIZ, the EMS, and the Heidelberg Academy, is the world's most comprehensive and longest-running abstracting and reviewing service in mathematics, linked to the preprint portal *arXiv.org*.

The Encyclopedia of Mathematics (EoM)[6] is an Open Access resource designed specifically for the mathematical community. The *Model Order Reduction Wiki (MORwiki[7])*, initiated by MPI DCTS, provides interactive wiki article pages on benchmarks for the comparison of different methods, algorithms and software. The *Open Machine Learning*[8] project is an open science project to build an open, organized, online ecosystem for machine learning. The *Archive of Formal Proofs*[9] is a collection of proof libraries, examples, and larger scientific developments, mechanically checked in the theorem prover *Isabelle*. Furthermore, there exist large databases for collections of specific mathematical objects like the *L-functions and modular forms database (LMFDB[10])*, the *On-Line Encyclopedia of Integer Sequences (OEIS)*[11] or *FindStat*, a database and search engine for combinatorial statistics. Finally, we mention specific repositories like *polyDB*[12] which is a database for objects in discrete geometry and related areas or GAP data libraries which can be downloaded from a central portal, e.g. the *GAPs Small Groups library*[13] (450 million finite groups).

Currently, the infrastructure environments for mathematical research data are primarily concerned with **information retrieval** or **catalogue services** for specific topics such as, e.g., the OEIS, polyDB, LMFDB, MORwiki, OpenML or general services like zbMATH, and swMATH. Generally, **methodic requirements from research** initiated these collection efforts of special mathematical objects and their description. Complementary, research infrastructures are services like SageMath, JupyterHub with mathematical notebooks in R, Julia, or Python. These services are mostly individual solutions to specific mathematical research data management problems, and are based on a wide range of RDM software frameworks. This hinders a FAIR use of data repositories over larger branches of the mathematical community, and the interlinking of similar approaches in disciplines that build upon mathematics.

A current trend in science is to combine the infrastructure for data processing with the infrastructure for research data storage, e.g., in grid computing in particle physics or Big Data analysis centers for machine learning. These efforts provide a stable and sustainable environment to researchers which are not able to host such resources at their respective institution. This is also one of the goals of the European Open Science Cloud (EOSC), and we discuss this issue further in Section 4.2.

MaRDI layer architecture Our needs analysis within the mathematical community represented by its sub-communities in MaRDI, see **T1**, **T2**, **T3**, and **T4**, has resulted in the following layer architecture:

X1: Core: Ensuring the **interoperability** in data handling by means of data formats and data types

[5] https://www.zbmath.org
[6] https://encyclopediaofmath.org
[7] https://morwiki.mpi-magdeburg.mpg.de
[8] https://www.openml.org
[9] https://www.isa-afp.org
[10] https://www.lmfdb.org
[11] https://oeis.org/
[12] https://db.polymake.org
[13] https://www.gap-system.org/Packages/smallgrp.html

for storage in the computer and programming languages, on external storages (input/output), metadata formats, and software application interfaces (APIs).

X2: Data: Securing **findability** and **accessibility** by data repositories and interfaces to find and access these data. Interface to or integration with the repositories of other NFDI consortia. Machine-actionable web interface and web portal for interaction by humans.

X3: Exchange: Innovative data services to enable new research and added scientific value through **re-use** of quality data. This area also includes workflows for automated data handling, e.g., data analysis workflows, Jupyter notebooks etc.

X4: Knowledge: Semantic knowledge in the form of ontologies and **knowledge graphs** to open up mathematical research data for new research. Examples are the linking of algorithms with references to implementations, benchmarks and publications or linking the model database with references to application problems, equations, simulation software, test problems and numerical or experimental data. This layer enables **Findability** and **Accessibility** also in documents.

This layered architecture was inspired by the EOSC architecture. We consider this structure compatible with the needs of the mathematical community and the requirements of mathematics as a cross-cutting discipline for many other scientific areas. This is reflected in the need for a semantic level for mathematical data in other NFDI consortia, see Section 4.2 and **T4**, and in particular **M4.3**. We partnered with OpenAIRE and will integrate the MaRDI **X4: Knowledge** layer as closely as possible into the OpenAIRE knowledge graph.

In the context of our work program described below in section 5 we will present the measures associated with the respective task area along with the pertinent cross layer. In order to make the notions behind these layers transparent we next demonstrate the architectural interplay of this MaRDI structure by means of a fundamental problem in mathematics.

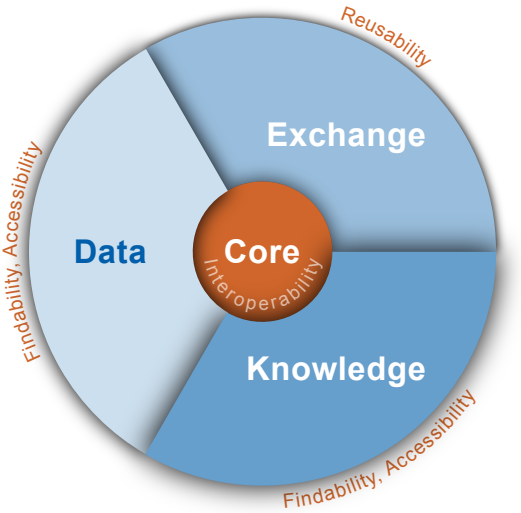


Figure 3: MaRDI Layer Model.

Guiding example: a benchmark framework for linear solvers We explain the MaRDI layer concept with a (simple) guiding example from the task area T2: Scientific Computing: We consider the solution of linear systems of equations

$$Ax = b,$$

over $\mathbb{R}^n$ where A is a $n \times n$ matrix, b is the vector of the right side and x is the unknown solution vector. The MaRDI layers can be interpreted in this example as follows:

X1: Core: Providing data types for matrix (A), vectors (b, x), and solution parameters. Different representation formats (CRS, CSC, Coordinate List, ...) and exchange or file formats (Matrix Market, Harwell-Boeing, ...). Software interface to the solver itself.

X2: Data: Repository with relevant test cases, i.e., concrete matrices, solution vectors, assigned with persistent identifiers (PIDs), web portal and web interface. Metadata standards and formats for the description of the properties of the matrix (e.g., symmetric positive definite, sparsity, M-matrix) and the solver (e.g., direct, iterative). Query function for test problems with special properties, e.g. symmetric, positive definite, size. Support of curated collections of test problems.

X3: Exchange: Benchmark framework for testing and comparing solvers. For this purpose the implementation of a solver code is adapted to the software interfaces related to the **X1: Core** and the **X2: Data** layer. Execution of the benchmark either locally or using a predefined software environment provided by MaRDI via containerized cloud computing allows to determine performance data (runtime) and memory requirements.

X4: Knowledge: In the semantic layer, knowledge about solution algorithms is linked to descriptions in the algorithm database, to publications via zbMATH, to software via swMATH and the data itself (test data, performance data, also from other NFDI consortia). To create semantically rich metadata, vocabularies and ontologies for matrices and linear solvers are required.

This results in the following added values: (a) **method developers** will benefit because they can run new algorithms on many test problems and also on special collections. (b) **users** will benefit, because they can generate performance data for selected algorithms or their implementations regarding various test problems from their application context. (c) All results of solver runs using the benchmark framework can be logged and recorded in the data layer again, e.g. performance data. The collection of these results contributes to **systematic evaluation** and analysis of the limits of solvers or their implementations.

The knowledge graph enables search and retrieval of publications related to a specific solver or its implementations. Conversely, one can refer to the test data, algorithm database, benchmarks in publications and make them traceable in terms of next-generation peer review. Such a research data infrastructure enables scaleable, transparent and re-usable research in the field of linear solvers.

A prototypical example for the realization of this concept in the area of machine learning is the portal https://www.openml.org, which is co-developed by MaRDI co-applicant Bernd Bischl (LMU Munich).

Synergies The specific requirements of the individual deliverables and milestones in the sub-projects of the individual task areas lead to a constant maturation of the MaRDI technologies (see T5: The MaRDI Portal) during the development process in the funding period. Synergies in the realization of new deliverables are expected along the following cross-cutting topics, cf. the requirements identified in use cases above:

S1: Algorithms are the core component in the mathematical research data life cycle for data processing. Among such algorithms are, e.g., the Buchberger-algorithm to determine a Gröbner basis for exact computations, the Gauß-elimination for systems of linear equations, iterative methods such as Newton's method, multi-grid methods, or Monte-Carlo methods. Metadata of these algorithms then describe mathematical properties like the convergence order of an iterative or approximation methods, the regularity of solutions, or the properties of admissible input (e.g. being a symmetric positive definite matrix).

S2: Mathematical Models are key buildings blocks for the application of mathematical methods to real-world problems.MaRDI will describe their mathematical properties, like, e.g., the existence or uniqueness of solutions, discretization methods for numerical simulation and perturbative methods to derive approximate models (simplified problems, boundary layers) with exact solutions.

S3: Workflows are processing steps of mathematical research data to achieve scientific results, see Figure 2. In mathematics the role of a proof is to document the methods which were used to obtain a result. In computer-aided mathematics workflows take this role: They collect auxiliary information, including types and versions of software used, program code, intermediate results, formal proofs, and more to achieve confirmability, a key feature of mathematics.

These synergy topics that emerge in all task areas **T1**, **T2**, **T3** and **T4** are the basis for an organic, vital convergence of MaRDI's bottom up approach addressing the needs of the mathematical community.

Use cases MaRDI aims to support mathematical research data management for pure as well as applied mathematics and at the interface to other disciplines. The methods, standards, and services will be developed and deployed by MaRDI in a bottom-up approach starting with specific developments in sub-communities by use cases and case studies representing different users and types of mathematical research data. In the following we describe three important example categories of such use cases and how the MaRDI's layered infrastructure and synergy topics contribute to better mathematical research.

Use case A: Experimental and Computational Mathematics While it is a common task in applications of computer algebra to browse a database searching for mathematical objects with a specific property, additional challenges arise in experimental mathematics. Often, we need to analyze one given object and certify its properties. An example is a system of polynomial equations, which we can solve by computing an elimination Gröbner basis. Here, it may be desirable to store that certificate, which is hard to compute but easy to verify independently, as an invariant entity. Another example is the secondary fan of a fixed finite point configuration, which encodes all regular triangulations; this can be used, e.g., to derive information about moduli spaces of tropical curves of fixed genus. This asks for an infrastructure where individual researchers can upload their certificates, typically complementing a classical publication. Standardized and open data formats, as much as possible independent from existing software systems are a task for MaRDI. This is realized by MaRDI's **X1: Core** and **X2: Data** layer and involves the synergies **S1: Algorithms** and **S3: Workflows**.

Use case B: Mathematical Modeling & Simulation Applying mathematics to other disciplines leads to a variety of different and often strongly inhomogeneous research data. Publications at the interface to other disciplines and applications are mostly interdisciplinary and should therefore be found and accessible in different portals. Mathematical models describe the (real-world) problems. However, there is no cross-disciplinary standard to systematically describe and categorize the models for concrete applications. Problem instances that contain the parameters and input data for the concrete questions are diverse and their description depends on the field of application. Software developed for numerically solving the problems has to be linked to the existing models and also to the problem instances. Experiments and evaluations are often huge, unsystematic, and incompletely docu-

mented. An assignment to the corresponding models, problem instances, and simulation tools as well as a searchable description of the achieved results and enhancement with appropriate metadata is missing. MaRDI will provide a database of algorithms (**S1: Algorithms**) and of mathematical models (**S2: Mathematical Models**) and connect the related objects in its knowledge graph (**X4: Knowledge**). Numerical methods for simulation software will be made interoperable within **X1: Core** allowing for benchmarks and validation of methods (**X3: Exchange**).

Use case C: Statistical Data Analysis and Machine-Learning Experiments In the context of machine learning (ML), benchmark experiments provide information about the performance of algorithms on different datasets. They contribute to a better understanding of the behavior of ML algorithms in different scenarios, which is indispensable in the development of ML algorithms. For example, new insights can be gained by analyzing benchmark results together with meta-information such as data characteristics and the used hyperparameter settings of the algorithm. However, benchmarks of more complex ML pipelines are rather rare compared to conventional ML algorithms. Ongoing efforts to improve ML pipelines lead to an ever-increasing need for more large-scale benchmarks. These complex benchmarks require the use of more efficient methods to define a proper experimental design and to analyze their results. The high-dimensional, mixed, and nested input space of ML pipelines complicate the design and analysis of benchmarks. Concrete guidelines and tools to simplify and ultimately unify this task are still missing. This will be addressed by MaRDI's **S3: Workflows**. Furthermore, the sole use of static descriptive visualizations to analyze such complex benchmarks is impractical and often not target-oriented. In this case, providing tools that simplify the statistical analysis of complex benchmarks will support researchers (**X3: Exchange**). Making the experimental results easily accessible in a machine-interpretable form (**X1: Core** and **X2: Data**) will facilitate collaborative research and foster collaboration.

4.2 Metadata standards

The initial purpose of metadata is to yield findability of data. Standardized metadata have the potential to facilitate data interoperability. There is no established standard yet for metadata in computer algebra, scientific computing, statistics and machine learning or applied and computational mathematics in general; let alone across scientific disciplines. The task areas **T1**, **T2** and **T3**, in cooperation with **T4** and **T5**, will work towards metadata standards by implementing controlled vocabularies and ontologies in dedicated measures.

In terms of the MaRDI layer architecture, such metadata (initially) fall into the layer **X2: Data**: MaRDI needs to go beyond purely descriptive, bibliographic metadata (e.g., author, title, date, version, …), where it will employ well established standards in close collaboration with other initiatives within the NFDI. For mathematics specific applications, such as the database *polyDB* for object in discrete geometry and related areas, metadata-formats already have been established. Within MaRDI, we will homogenize those formats iteratively to make use of synergies with similar initiatives. Additionally, we will closely collaborate with the EngMeta [IS19; SI19; SI20] effort towards standardizing experiments in engineering sciences as relevant for **T2**, **T4** and partially **T3**. This effort suggests to additionally include workflow metadata (e.g., software, input, output, the person who generated the actual data, …)

and content (subject-specific) metadata, i.e., a description of the observed/simulated system (components, variables, boundary conditions, parameters, …). MaRDI will work towards interlinking such schemes by ontology supported crosswalks and towards enhancing by developing micro standards for metadata to describe characteristics of mathematical and computational concepts like performance and applicability, to be used in the envisioned MaRDI *Database of Numerical Methods* (see **M2.1**) and the MaRDI *Model Database* (see Measures **M4.2** and **M4.3**). The recently started AIMS platform[14] for creating, combining and sharing metadata profiles can serve as an interlinking point to other disciplines and their standards.

The envisaged resulting metadata enrichment already constitutes a major improvement in terms of (disciplinary, both within and outside mathematics) FAIRness advances along the **X2: Data** layer, and towards constructing knowledge graphs (**X4: Knowledge** layer) for subareas in mathematics. The recommendations of the *European Open Science Cloud (EOSC)* interoperability framework[15] can be addressed by importing hierarchically linked micro-schemata into the MaRDI knowledge graph provided that the used concepts are mapped to corresponding ontologies.

The main conceptual challenge in the cooperation with other disciplines towards the vision of unified, interoperable knowledge graphs is to embed the outcome of MaRDI into the metadata frameworks developed by other NFDI consortia and vice versa, i.e., to attach the documented research data descriptions of input and output data as well as the iterative cycling through workflows. This means that already from the beginning, the metadata schemes, vocabularies and ontologies should be designed in such a way that future interoperability is promoted. This problem is faced by all NFDI consortia alike and should be coordinated in a common effort. MaRDI will actively take advantage of the fact that some of its partners (see **T4**) benefit from a head start: Based on the findings of, e.g., topical workshops organized by already funded NFDI consortia in 2020 and in which MaRDI participated, the conclusion arises that the *Wikibase* knowledge graph implementation will be the platform of choice to create such a hierarchy of interlinked micro-schemes in a way that allows direct interlinking to ontologies and thereby leveraging explicit semantic representations to enable sophisticated queries and fundamental added benefits in terms of the **X4: Knowledge** layer. In particular, the Wikibase implementations support the linking between different vocabularies via external identifiers and the inner-graph linking via properties to represent facts, which can be enhanced with qualifiers to represent additional semantics via meta-statements, as e.g. information about provenance, validity, or context of represented facts.

4.3 Implementation of the FAIR principles and data quality assurance

MaRDI implements the FAIR principles by design and by individual measures and deliverables. Specifically, MaRDI's layer architecture (cf. Figure 3) addresses each of the FAIR principles:

Findability and **Accessability** of mathematical research data are realized by the **X2: Data** layer through interfaces to repositories and services of other NFDI consortia and metadata schemes. By linking data sets to mathematical methods and models, to publications, to research software and to benchmarks and using semantic descriptions by vocabularies and ontologies the **X4: Knowledge**

[14] DFG project "Applying Interoperable Metadata Standards (AIMS)", read more https://tinyurl.com/y2gllpul
[15] https://op.europa.eu/en/publication-detail/-/publication/78ae5276-ae8e-11e9-9d01-01aa75ed71a1/language-en

layer enables semantic search and cross-walks in the sense of Web of Data or WikiData. MaRDI will set up a central platform, the MaRDI portal, through which relevant data and applicable services can be found (even if the user does not know of their existence or applicability).

Interoperability on the data level is established by MaRDI's **X1: Core** interfaces and APIs. On a knowledge level interoperability of semantic technologies will be realized by utilizing WikiBase and WikiData standards.

Re-useability is established by MaRDI's **X3: Exchange** layer at two different levels. On a low level, MaRDI will develop and provide shared components for data processing and analysis. On an application level MaRDI implements platforms for research like the benchmark framework for linear solvers or OpenML for machine learning, to name only two. These services lead to an increasing re-use of research data in mathematics and other disciplines. Moreover, through the **X4: Knowledge** layer MaRDI's knowledge graph can be integrated within other knowledge bases, e.g, from other NFDI consortia. This enables the re-use of semantic web information, and allows for a mutual search and access.

MaRDI will introduce a score describing the level of implementation of the FAIR principles for mathematical research data. This will be realized in **T5** as part of the MaRDI Portal with respect to data life cycle management services. For example, a software or dataset that is found to be incomplete, faulty, or no longer up to date will be marked as deprecated and potentially replaced by a new version.

A second level of data quality assurance is dedicated to the scientific aspects of mathematical research. MaRDI will address these by establishing techniques of next-generation peer review for data and software, as traditional methods of peer review are not well suited for computational experiments in mathematics; see Use Cases above. For this purpose, MaRDI will support certification of data, mathematical software, and results. For example, MaRDI will distinguish between raw data and one that has been endorsed, tested, or proven to be correct as in the case of algorithms. In the guiding example a solution for a linear problem or the property of the related matrix could be certified. The certification of mathematical research data is essential for the conformability of mathematical results, see objective O2.

4.4 Services provided by the consortium

The MaRDI consortium will provide a multilevel infrastructure for developing and maintaining a sustainable system of services. In particular, the MaRDI infrastructure will cover the full spectrum from data repository solutions to advanced math-oriented applications, including

- distributed repositories, long-term preservation, creation of stable unique identifiers, authentication and authorization tools;
- processing of data streams and distributed computations, generation of mathematical knowledge graphs and terminology services;
- standardized APIs allowing for computation, validation, and data submission, as well as for interlinking internal MaRDI services, other NFDI resources, and external platforms; and
- advanced mathematical retrieval functions like formula search, filtering by mathematical properties.

MaRDI will provide the following services categorized along its layers:

Data Services. These services are related to **X2: Data**. A selected list of planned services includes:

- Repositories for Computer Algebra (polyDB, Small Groups Library) (in **T1**)
- Library of curated benchmark data (in **T2** and **T3**)
- Library of mathematical models (in **T4**), cf. **S2: Mathematical Models**
- Persistent Identifier Registry (in **T5**)

Services for research. This layer is dedicated to **X3: Exchange** and provides services with added value for the scientific community through combination of different data sources. A selected list of planned services includes:

- Benchmark framework (in **T2** and **T3**)
- Workflows (in **T1**, **T2**, **T4**), cf. **S3: Workflows**
- Library of Statistical Analysis (in **T3**), cf. **S1: Algorithms**
- Notebooks for reproducible computer experiments (in **T1**, **T2**, and **T3**)
- Software environments, MaRDI containers for computing and experiments (in **T1** and **T2**)
- Workflow and data certification service for next-generation peer-review (in **T1** and **T3**)

Knowledge Services. These services are related to **X4: Knowledge** and the mathematical knowledge graph . A selected list of planned services includes:

- Knowledge graph of numerical algorithms (in **T2**), cf. **S1: Algorithms**
- Benchmark information knowledge base (in **T2** and **T3**)
- Knowledge graph for mathematical modeling and simulation (in **T4**), cf. **S1: Algorithms**, **S2: Mathematical Models**
- Terminology service (in **T2**, **T3**, **T4**, and **T5**)
- zbMATH open
- swMATH

In particular, the algorithm database, or the database of mathematical models will contribute to the collection and redistribution of mathematical knowledge within mathematics as well as within related disciplines. The integration with the *Encyclopedia of Mathematics* will complete this service layer of MaRDI. Additionally, the integration of swMATH in OpenAIRE will expand the knowledge graph of mathematical software on the level of EOSC. The basis for the creation of the mathematical knowledge graph are the semantic-rich metadata provided by the terminology service in all areas.

More services, especially at the level of a pilot or demonstrator, e.g., the MORwiki benchmark tool in **T2** or the software tool for the analysis of benchmark results in **T3**, can be found in the description of the respective task area.

Finally, the **MaRDI Portal** developed in **T5** will provide a one-stop entry point to these MaRDI services. Users access all data using a selection of unified interfaces that let them transparently search over the whole data collection. For MaRDI users and developers community workshops and outreach activities are provided within **T6**.

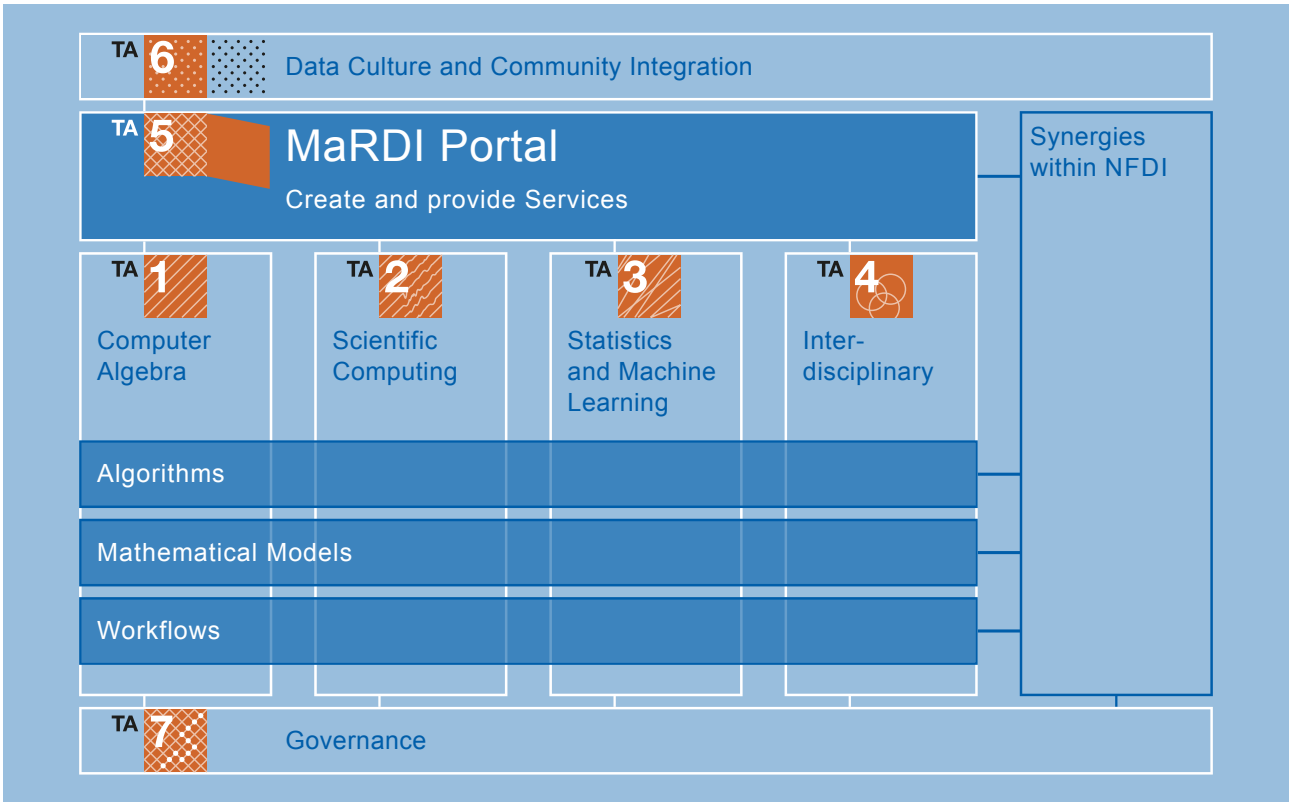


Figure 4: Organization of MaRDI's work program

5 Work program

MaRDI pursues the representation of its user communities by design. Thus, the working program of the MaRDI-proposal relies on research motivated pillars of the MaRDI-proposal corresponding to a division of mathematical research data in *exact and symbolic data* (T1: Computer Algebra), *floating point data* (T2: Scientific Computing), and *data with uncertainty* (T3: Statistics and Machine Learning). These areas cover an enormous extent of mathematical data types. Furthermore, we take into account the interdisciplinary role of applied mathematics by a specific task area T4: Cooperation with Other Disciplines. This structure is a consequence of MaRDI's bottom-up approach, and reflects the heterogeneity of mathematical research data and the requirements for the research data management, see Section 4. The specific requirements of the individual deliverables and milestones in the sub-projects of the individual task areas lead to a constant maturation of the MaRDI technologies within T5: The MaRDI Portal during the development process in the funding period. Synergies in the realization of new deliverables are expected along the following cross-cutting topics, **S1: Algorithms**, **S2: Mathematical Models**, **S3: Workflows**. These synergy topics that emerge in all task areas **T1**, **T2**, **T3** and **T4** are the basis for an organic, vital convergence of MaRDI's bottom up approach addressing the needs of the mathematical community. Through the direct integration of the national mathematical society, DMV, see task area T6: Data Culture and Community Integration, we aim for the full feedback and dissemination cycle within the mathematical community and beyond. Within task area **T4** we define a portfolio of case studies with explicit references for the collaboration with other

NFDI consortia. As such **T4** is the bridge from the mathematics- and computation-related research in other scientific disciplines into the mathematical communities represented by **T1**, **T2**, **T3**.

The tasks for this proposal have been identified and formed according to the respective *readiness* of preliminary work and services. Such a readiness indicator very well reflects the consolidated needs of the mathematical community. We expect that these flagship projects (task areas) will develop a pull effect on further parts of the mathematical community and also at an international level with regard to discussing their needs with respect to mathematical research data.

The aim of task area T5: The MaRDI Portal is to develop, implement and maintain a user-friendly way to make MaRDI's mathematical knowledge, research data and services accessible to the scientific community. The MaRDI portal is designed to achieve this goal and to become a one-stop contact point for mathematical research data in the mathematical community and beyond. It makes research data accessible through a low barrier unified user interface and machine accessible interfaces (APIs) as well as through the overarching NFDI knowledge graph. The task area will develop and host a storage service for mathematical research data, the portal infrastructure for the front- and backend, including the necessary interfaces for integration of (internal and external) data repositories and services.

With the work program we specifically address MaRDI's objectives

O1 Interoperable mathematical research data
O2 Secure confirmable and reproducible results
O3 Establish confirmable workflows
O4 Enable the development of mathematical services
O5 Establish next-generation peer review
O6 Standardize semantic relations
O7 Culture development and community embedding,

see the following table. MaRDI's **offer to other consortia in the NFDI** is realized via Case Studies in a dedicated task area **T4** on a scientific level and in **T5** on an infrastructural level. Moreover, representatives of other consortia as well as NFDI stakeholders are integrated into the board structure in MaRDI's governance; see **T7**. Summarizing, all task areas contribute to this portfolio; see Figure 4.

T1: Computer Algebra (Michael Joswig, Wolfram Decker, Claus Fieker)	Objectives
Measure 1.1: Confirmable workflows for computer algebra	O2 O3
Measure 1.2: Data formats and data bases	O1 O4
Measure 1.3: Technical support for publishers and journals	O5
Measure 1.4: Predefined software environments	O4
Measure 1.5: Training of researchers and technical staff	O7

T2: Scientific Computing (Peter Benner, Mario Ohlberger)	Objectives
Measure 2.1: Knowledge Graph of Numerical Algorithms	O6
Measure 2.2: Open Interfaces for Scientific Computing	O1 O4
Measure 2.3: Benchmark Framework	O1 O2 O3

Measure 2.4: Description and Design of FAIR CSE Workflows	O3 O6 O7
T3: Statistics and Machine Learning (Mathias Drton, Bernd Bischl)	**Objectives**
Measure 3.1: Library of Curated Benchmark Datasets	O1 O2
Measure 3.2: Library of Statistical Analyses	O3 O6
Measure 3.3: Empirical Analysis of Machine Learning Experiments	O4
Measure 3.4: Standards for Peer Review of Numerical Experimentation	O5
T4: Cooperation with Other Disciplines (Anita Schöbel, Dominik Göddeke)	**Objectives**
Measure 4.1: Documentation and Analysis of Interdisciplinary Workflows	O6 O3
Measure 4.2: Standardization of Mathematical Descriptions across Disciplines	O1 O6
Measure 4.3: MaRDI Platform for Interdisciplinary Exchange	O4
Measure 4.4: Transfer beyond Case Studies	O7
T5: The MaRDI Portal (Christof Schütte, Harald Sack)	**Objectives**
Measure 5.1: MaRDI Portal Technology	O4 O7
Measure 5.2: Data and Service Lifecycle Management	O4
Measure 5.3: Standardization & Interfaces	O1 O2
Measure 5.4: Service Infrastructure	O4
Measure 5.5: Distributed Computing and Storage Infrastructure	O4
Measure 5.6: Service Development, Integration and Maintenance	O4 O7
T6: Data Culture and Community Integration (Bernd Sturmfels, Andreas Matt)	**Objectives**
Measure 6.0: Internal Communication and Dissemination Coordination	
Measure 6.1: Engagement with Mathematical Community	O7 O5
Measure 6.2: Engagement with Information Specialists	O7
Measure 6.3: Engagement with General Public	O7
T7: Governance and Consortium Management (Michael Hintermüller)	
Measure 7.1: Establishing MaRDI Administrative Structure and Governance	
Measure 7.2: Strategic Embedding and Organization	

T1: Computer Algebra

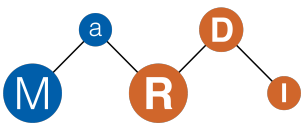
Today computers paired with sophisticated mathematical software systems allow for far reaching experiments which were previously unimaginable. In the realm of algebra and its applications, where exact calculations are inevitable, the software tools are provided by computer algebra systems. These systems are large, complex pieces of software, containing and relying on a vast amount of mathematical reasoning. It is an important aspect that through these systems a large treasure of mathematical knowledge becomes accessible to and can also be applied by non-experts.

Scientific results gain a lot of their value from documenting the methods by which they have been obtained. In mathematics, this role is traditionally played by proofs. Additionally, computer-aided mathematics necessitates to collect auxiliary information, including types and versions of software used, program code, intermediate results, formal proofs, and more. A major challenge in computer algebra is the technically well-coordinated modeling of heterogeneous data with complex semantics. Further, it is a characteristic of computer algebra that the generation of interesting data often takes a great deal of time, but the amount of data generated is often rather small, i.e., within the gigabyte regime (nonetheless, certain applications may produce terabytes of data). This is an important difference to areas such as Optimization, Numerical Analysis or Scientific Computing, where the amount of data is often significantly larger, but the procedures applied to it are relatively simple (in relation to the amount of data). However, there is an overlap with Scientific Computing in **T2** and Statistics and Machine Learning in **T3**, and we will collaborate closely to find the common ground.

From its very beginnings, computer algebra has been connected with algorithmic and experimental methods in commutative algebra, group and number theory, algebraic geometry and neighboring fields. Here we interpret the term computer algebra in a very wide sense to include computational logic, symbolic computation, computational methods in all areas of geometry, combinatorial optimization and more. In recent developments, the success story of computer algebra in the afore-mentioned fields makes its use in more and more mathematical subdisciplines attractive. Even further, applications to all areas of science came into focus. This is documented best through the formation of the SIAM Activity Group on Algebraic Geometry, with its flagship journal SIAM Journal on Applied Algebra and Geometry (SIAGA). The most recent installment of the SIAM Conference on Applied Algebraic Geometry, which took place in Bern in July 2019, saw more than 800 participants; which marks a steep increase. This additional focus, on applications outside mathematics, also requires to re-think the connections, e.g., between Computer Algebra and Scientific Computing, as research in, e.g., biology or chemistry may want to employ techniques from both fields. Investigating and developing such connections, in collaboration with **T2** and **T3**, is thus expected to have an impact beyond research data in the narrow sense. For such applications we will also cooperate with **T4**.

We propose the following list of measures. They will be implemented jointly between the groups at TUB and TUK, which have a long-standing cooperation, e.g., within the DFG SPP 1489 and TRR 195.

- Measure 1.1: Confirmable workflows for computer algebra
  - **Layers:** **X1: Core** (**X3: Exchange**)
- Measure 1.2: Data formats and data bases



  - **Layers:** X2: Data (X1: Core, X4: Knowledge)
- Measure 1.3: Technical support for publishers and journals
  - **Layers:** X3: Exchange
- Measure 1.4: Predefined software environments
  - **Layers:** X4: Knowledge (X1: Core, X3: Exchange)
- Measure 1.5: Training of researchers and technical staff
  - **Layers:** X1: Core (X4: Knowledge)

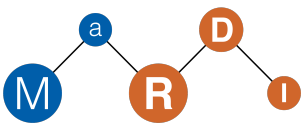
The target user group of this task area is the entire mathematical community, with a focus on developers and users of computer algebra software. The measures **M1.1**, **M1.2** and **M1.4** mainly target the software developers. The target user group of **M1.3** are publishers, editors and reviewers of software and data publications from computer algebra. Establishing a process for such publications will benefit the mathematical community as a whole. Measure **M1.5** aims at spreading the knowledge gathered by the other measures, as well as educating and recruiting new developers.

To set a standard for confirmable work flows requires a very flexible approach. On the lowest level this may involve just an informal documentation of the kind suggested by Bailey, Borwein and Stodden (2016), see **M1.1** below. On the highest level this may involve a formal proof of a theorem in mathematics (e.g., given in Isabelle or Coq). This flexibility is important for the acceptance of any standard among users, with different research goals and different levels of technical expertise. The goal for the entire MaRDI project is to encourage higher levels of formalization, without forcing this beyond some basic requirements, e.g., to assert searchability or cover legal aspects. **M1.1** will survey the available guidelines and adapt them to the specific requirements of computer algebra.

Computer algebra results constitute (parts of) mathematical proofs, and as such they hold indefinitely, if correct. To adhere to the FAIR principles of reusability and interoperability it is particularly important how data of this kind is encoded. **M1.2** addresses the development of data formats for objects from computer algebra. The User Forum (see **T7**) will have an important role for giving feedback and also raising the acceptance of standards in the mathematical software community.

Any file format considered here needs to allow for the complete migration of the data into any future file formats of the next century. Established standards such as XML and JSON, with a wide software infrastructure (e.g., for syntax validation), play a key role. Binary file formats are explicitly discouraged, and closed standards are entirely unacceptable. While data compression is relevant, this aspect will be largely ignored in **T1**. In practice, standard compression tools (e.g., gzip or xz) are employed. These are orthogonal to any file format specifications. **M1.2** will profit from decades of experience with serialization within the polymake project and liaise with developers from the OSCAR project, who are working on serialization of a variety of objects from algebraic geometry.

Interaction with the mathematical community is crucial for the success of MaRDI. While this is generally emphasized by the task areas **T4**, **T5** and **T6**, the specific developments from **T1** will be communicated in **M1.3**, **M1.4** and **M1.5**. For any development to be effective at all, it is essential to provide prototypes in the early stages to gather feedback. Open source mathematical software can be hard to install and even when it is provided by a package manager it is not guaranteed to interact



with other packages as expected. An example for this are the packages for polymake and Singular in the Linux distribution Debian. Both can be installed via Debian's package manager apt, but the library interfaces provided to each other are dormant. To overcome this obstacle, **M1.4** will provide predefined and easy to install software environments with up-to-date versions of many mathematical software packages. These can then be used for further development within the other measures.

Within **T1** we will not provide any capacity for storing user data. Instead we will provide software tools (e.g., docker containers) and guides (e.g., web based tutorials and interactive Jupyter notebooks) for establishing confirmable work flows in computer algebra. Users will receive digital object identifiers (DOI) and recommendations for storing their data via the MaRDI Portal (**T5**).

The currency within mathematics are publications. In recent years the idea that software should count as such has gained momentum. This requires software to undergo a peer-review process similar to other publications. To help shaping such a process is the key task of **M1.3**. It requires to interact with journal editors and publishers, as well as with scientific authors. This is an involved process strongly connected to the FAIR principles of findability and accessibility. We anticipate problems with, e.g., the code quality, proprietary software components, licenses, dependency on software which is no longer available, extremely long running times, and the (lack of) stability of web-based resources. Solutions will spread awareness of the value of software contributions, and improve software quality via feedback from the peer review process.

Training of researchers and collecting feedback is the purpose of **M1.5**. This includes organizing workshops, arranging meetings, and traveling to research institutions. **M1.5** also serves as a connection to **T4**, **T6** and **T5**. It is of utmost importance to make the community aware of new guidelines and best practices, as developed in **M1.1**. We need to know whether the designs actually meet the needs of the community and communicate this feedback to the developers in **M1.2** and **M1.4**.

Measure 1.1: Confirmable workflows for computer algebra

Synopsis To follow the FAIR principles, we need to provide rigorous guidelines for conducting computer experiments. From the guidelines it then becomes clear what metadata needs to be collected in order to verify compliance. This measure will survey existing guidelines for software experiments, adapt them to the setting of computer algebra, and specify the necessary metadata to collect alongside.

Tasks

1. Considering the significance of experimental mathematics and the key role played by computer algebra methods in this context, it is of utmost importance to establish guidelines with regard to the reproducibility of experiments using these methods and to their documentation in scientific papers. In [BBS16], Bailey, Borwein and Stodden addressed Principles and Practise of this for the Scientific Computing community: (1) a precise statement of assertions to be made in the paper, (2) the computational approach, and why it constitutes a rigorous test, (3) complete statements of, or references to, every algorithm employed, (4) auxiliary software (both research and commercial software), (5) test environment (hardware, software and number of processors), (6) data reduction and statistical analysis methods, (7) adequacy of precision level and grid resolution, (8) full statement

or summary of experimental results, (9) verification and validation tests performed, (10) availability of computer code, input data, and output data, (11) curation: where are code and data available?, (12) instructions for repeating computational experiments, (13) terms of use and licensing; ideally code and data "default to open," (14) avenues explored and negative findings, (15) proper citation of all code and data used. The first task is to generalize these guidelines to fit the setting of computer algebra. The workflows developed should suggest data formats from **M1.2** and software environments from **M1.4**. (**X1: Core**)

2. Collect feedback from publishers in cooperation with **M1.3**. (**X3: Exchange**)
3. Collaborate closely with **T2** and **T3**, aiming at uniform workflow recommendations. (**X1: Core**, **X3: Exchange**)

Added value By adhering to the newly designed guidelines researchers make it possible to better understand their experiments and allow for reproduction, enhancing the overall acceptability. This understanding is crucial for software and computer experiments to play a bigger role in the peer review process.

Deliverables

D-TA1-A1 Confirmable workflow guidelines adjusted to the needs of computer algebra The above will be adjusted and extended to the needs of computer algebra. Depending on the application, different encodings of the same mathematical object may be necessary. This poses specific challenges which will be illustrated by examples in **M1.2** below. (**X1: Core**)

D-TA1-A2 Protocol for error reporting With **T6** we will develop a protocol for reporting and correcting errors; cf. Bailey et al. (2013), Section 9.7. Despite all our efforts our data will be faulty. But eventually, errors will be detected, and thus should lead to corrections. We think it is very important to keep track of and give credit to researchers who found such errors. In this way users of mathematical software and the MaRDI services can be rewarded and thus encouraged to scrutinize the data. Ultimately this will lead to an increased reliability. (**X3: Exchange**)

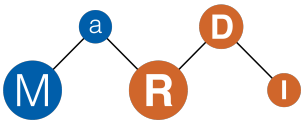
Milestones

M-TA1-A1 Presentation of guidelines to developers of the OSCAR project The OSCAR project, see https://oscar.computeralgebra.de/, is the common umbrella for a wide variety of software communities within computer algebra. We plan to closely cooperate with the OSCAR developers within this task area. After the draft stage, the guidelines will be presented to them, for collecting initial feedback and anticipating potential future issues.

M-TA1-A2 Setting up a web presence with a continuously updated guidelines For adhering to the guidelines it is crucial that an up to date version of them is easily accessible at all times. We plan to setup a website, containing an automatically updated version of the guidelines, as well as a collection of examples how to use these.

M-TA1-A3 Presentation of guidelines at computer algebra conferences A major milestone is the presentation of the guidelines to the general public.

M-TA1-A4 Journal publication of guidelines



Measure 1.2: Data formats and data bases

Synopsis This measure has a particularly wide range, which includes the following aspects: data serialization, making existing data available to a wide range of users, advising researchers in preparing and storing new data, establishing software interfaces and more. It establishes the fundaments for the main goals of securing confirmable and reproducible results, as well as defining and standardizing mathematical research data.

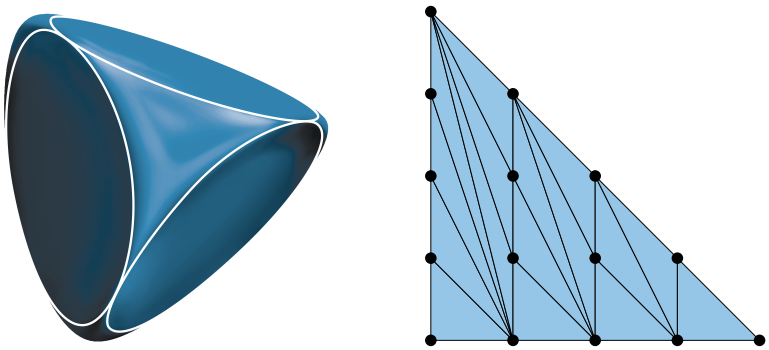


Figure 5: A spectrahedron (left) and a triangulation (right).

Depending on the usage of mathematical objects, several different encodings may be necessary, and this results in a multitude of challenges. For instance, a tremendous amount of research on the border between real algebraic geometry and optimization deals with spectrahedra, like the one in Figure 5 (left). A spectrahedron is convex body that is defined by polynomial inequalities. It is the set of feasible points in semidefinite programming. These objects represent a vast, and highly useful, generalization of convex polyhedra and linear programming. Yet to allow for visualization one needs to know an explicit representation as a set of points, for which an exact solution is often unavailable. So it may not always be feasible to switch between desirable representations of one object and our data formats, which we have to take into account in order to be practically relevant. The challenges can also come from an entire different direction. Some objects in tropical and toric geometry can be encoded as a finite (but very large) set of triangulations of a fixed point set (corresponding to monomials); see Figure 5 (right). For instance, the moduli space of smooth tropical cubic surfaces in 3-space is described by more than 14 million triangulations of the 20 lattice points in a dilated 3-simplex [JJK18]. Here the entire moduli space as one object is of interest as well as some individual examples among the 14 million triangulations. The data being encoded consists almost exclusively of vectors and matrices and only becomes valuable provided the context. For example a vector can be an element of a vector space, the coefficient vector of a polynomial or the exponent vector of a monomial. Hence it is necessary to develop complex semantics dictionaries. This asks for a representation as a complex hierarchical database, resulting in nontrivial requirements concerning storage and retrieval. Furthermore, the design needs to take existing data bases into account; a particularly relevant example in computer algebra is the GAP Small Groups Library [BEO01].

Tasks

1. **Standardize data formats for objects from computer algebra.** Currently there is no standardized data format for most mathematical objects from computer algebra. If data is stored and loaded this is mostly done as text files by the researcher experimenting with the software. This is a major

barrier for interoperability and reusability of said data. Some software frameworks use existing data standards, e.g. polymake uses XML ([GHJ16]) and JSON. There are ongoing efforts in the OSCAR project to extend serialization via Julia to bigger parts of computer algebra. Databases such as the ATLAS project of Finite Group Representations ([Wil+]) illustrate many of the underlying problems that need addressing: most of the involved objects (finite fields, number fields, groups, and their representations) do not admit a canonical presentation which is immediately useful from the mathematical point of view. As a result, the description of the presentation itself needs to be included in the data set. On the other hand, a representation might include, e.g., more than 10,000 algebraic numbers, so that the presentation of the field cannot be repeated each time. Solutions to this are known, and partly available in e.g. polymake, see [GHJ16; Paf17], but need to be expanded for many more data types. (**X2: Data**)

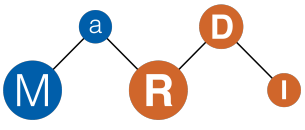
2. **Ensure that the data formats developed work seamlessly within the workflows developed in M1.1.** This requires to take into account existing file formats of existing software in order to convince developers and users of computer algebra systems. Clearly, this ties in with conceiving standardized software environments in **M1.4**. An important goal is to make the data formats as independent as possible from individual software systems. This will increase the accessibility, interoperability and reusability of data according to the FAIR principles. Due to involved and varied semantics in computer algebra we expect substantial differences to data formats relevant in **T2**. (**X1: Core**, **X2: Data**)

Added value Standardizing data formats will allow for easier transfer of data between different softwares, allowing for easier interfaces, leading to interoperability.

Deliverables

D-TA1-B1 Expansion of existing serialization methods to more data types To improve the access to already existing data, we will study and improve existing paradigms for data serialization in the field of computer algebra. One main challenge is that various subfields within computer algebra have competing traditions and standards, which need to be unified without losing their usefulness to the specialists. Other tasks include data migration paths and interfaces to existing software systems. For the highly relevant metadata aspects we will cooperate with **T5** to satisfy the FAIR principle of findability. (**X1: Core**)

D-TA1-B2 Import of datasets suitable as tutorials and test cases For identifying and solving problems, and for gathering feedback in the early stages of the development of data formats and protocols, rapid prototyping is necessary. The advantage is threefold. Firstly, new tools can be developed alongside existing datasets. Secondly the datasets will then also serve as test cases for any future changes, to avoid regression. Thirdly, the datasets can be turned into tutorials to guide new users. Repeatedly migrating the datasets throughout the development process will allow us to design robust and reliable versioning procedures. This also serves the main goal of securing confirmable and reproducible results. (**X2: Data**)



Measure 1.3: Technical support for publishers and journals

Synopsis One of the ultimate goals is to establish and formalize a refereeing process for software and datasets. Thus it is essential to engage the mathematical community and its institutions and based on this work with publishers and journals. Just like an article which has not been peer-reviewed does not count as a valid publication, the same is true for software. Peer-review improves the quality of the code and helps discover errors in both software and datasets. In the future code submissions will become far more frequent, as they already are in computer science. Hence any researcher wanting to publish software, will submit their code to journals with the appropriate procedures in place.

Some journals are already aware of these problems, and have implemented different kinds of peer-review for software. These include *Journal of Software for Algebra and Geometry*, *Mathematical Programming Computation* and the Section on Computational Algebra of the *Journal of Algebra.*

Added value This measure lays the groundwork for designing the next generation peer review by enabling reviewers to understand how output was produced. It enhances the findability and accessibility from the reviewers point of view, and in the long run, for the reader of a paper augmented by software and data.

Tasks

1. Liaise with selected journals publishing in the area of computer algebra. On submission of constructive papers, the journal editor should contact the MaRDI project. The PostDoc, as a technical editor, then contacts the authors of the paper and get access to the data and code base. The technical editor can also add a score to the data and code to indicate reliability. Thus the user can immediately get an idea of the quality of any programming code, documentation and data, independently of the paper. (**X3: Exchange**)
2. Convey the new guidelines and data formats to journals. (**X3: Exchange**)
3. Collect and relay feedback to **M1.1** and **M1.2**. (**X1: Core**, **X3: Exchange**)

Milestones

M-TA1-C1 Agreement with publisher finalized
M-TA1-C2 First publication with new procedures

Measure 1.4: Predefined software environments

Synopsis A different, complementary approach to increasing the reproducibility of the work is to provide predefined software environments which researchers can easily use, and which provide a complete set of interdependent software (different CAs, libraries and databases they use, etc.). By making this easy to set up and use, it becomes attractive for researchers to use, bringing the benefits of improved reproducibility as a side-effect, instead of an added burden. This measure is a key ingredient in enabling the development of mathematical services.

On a technical level, there are various ways to implement such environments, e.g. based on Docker images, virtual machines, curious containers or simply a software distribution such as Conda. Jupyter

notebooks are a nice way for documenting an experiment from the user perspective, although these are not useful for large cluster computations.

Added value This measure overlaps with and benefits **M1.3** by providing a basis to document software environments. Furthermore, it will greatly enhance the reusability of code by providing a reusable surrounding.

Tasks

1. Evaluate the various pros and cons of existing solutions. (**X1: Core**)
2. Make the environments accessible via a web interface. Possibly host these "in the cloud" as well, bringing further potential usability advantages to researchers. (**X1: Core**, **X3: Exchange**, **X4: Knowledge**)

Deliverables

D-TA1-D1 Repository of containerized software environments The most widely used container technology currently is Docker. Every Docker container is described by a text file, the Dockerfile. We plan to set up a repository of Dockerfiles with various combinations and versions of the software components involved in the OSCAR project. Basic work in that direction has already been done, since these containers are needed for the continuous integration frameworks of both polymake and OSCAR. Besides updating the containers for new software and OS versions, changes to Docker also require continuous maintenance of the repository. (**X1: Core**, **X3: Exchange**, **X4: Knowledge**)

D-TA1-D2 Increase longevity and maintainability of computer algebra software Given the appropriate environment, old software can run on new machines and can be maintained continuously. However, this requires access to the source code, which is why open source software is strongly preferred. It is crucial to document the data formats involved in interaction with **M1.2**. Predefined software environments are an important tool for documenting the workflow as in **M1.1**, and for refereeing the result, as in **M1.3**. The programming language Julia with its package manager and binary builder already provides useful tools to keep interfaces to other software usable. Thus, there will be close collaboration with the OSCAR project of the TRR 195 "Symbolic Tools in Mathematics and their Application". The environments developed can then be embedded in the infrastructure and services provided by **T5**.

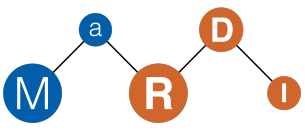
The software used in computer algebra is very different from the software in scientific computing. As in **M1.2** we expect substantial differences to **T2**. (**X1: Core**, **X3: Exchange**, **X4: Knowledge**)

Milestones

D-TA1-D1 Repository setup with basic containers
D-TA1-D2 Automatic build setup

Measure 1.5: Training of researchers and technical staff

Synopsis For establishing guidelines, best practices and data formats it is crucial to provide the necessary support to the target audience, to spread the knowledge and to encourage their use.



Added value By spreading knowledge and collecting feedback, this measure is crucial for **M1.1**, **M1.2**, and **M1.4** to succeed. Thus it is essential on the road towards FAIR data in computer algebra.

Tasks and Deliverables

D-TA1-E1 Data carpentry workshops for computer algebra To improve the availability and reliability of research data and mathematical software in the long run, authors and researchers need to be supported (**M1.1** and **M1.2**), but also informed and trained. Best practices for scientific programming are well known - but not to all scientists. For instance, current standards for continuous integration during the development of mathematical software systems were impossible in the past, due to being expensive in terms of CPU time and storage. It is necessary to raise the awareness of new tools through organizing workshops. (**X4: Knowledge**)

D-TA1-E2 Interdisciplinary communications Guidelines developed in **M1.1** and new data formats established in **M1.2** need to be communicated to other disciplines via **T4**.

D-TA1-E3 Feedback collection Staying in touch with the users of software systems will help to collect feedback to improve existing tutorials and tool chains. This will be done is close cooperation with **T6**. (**X1: Core**, **X4: Knowledge**)

Milestones

M-TA1-E1 First workshop

M-TA1-E2 Internal workshop with M1.1 incorporating collected feedback

Services

Confirmable workflows for computer algebra A major goal is to make computer experiments accessible to the entire mathematics community. The guidelines of **M1.1** will be developed such that anyone is able to understand how the experiment was conducted, to confirm that it produces the result it claims to produce, and to reproduce the experiment on their own computer if desired.

Data formats for computer algebra Establishing data formats across different software systems is crucial for interoperability. These will also be referenced in the workflows developed in **M1.1**.

Peer review process for software contributions Following the guidelines developed in **M1.1** will make software and computer experiments accessible to peer review. Together with **M1.3** this will culminate in a standardized way to publish software and the outcomes of computer experiments from computer algebra.

Predefined software environments Fixing the software environment, i.e. the versions and underlying container, enables one to reproduce computations on different machines. At the same time this is very important for debugging software, if a computation produces a different result depending on the environment.

Workshops and tutorials on good practice of computer experiments For the data formats and workflows developed to have any effect, they need to be communicated actively to developers and mathematicians outside of MaRDI.

Embedding into NFDI

Via standardizing data formats and guidelines for confirmable workflows, this task area mainly contributes to the objectives **O1**, **O2**, and **O3**. The main incentive for the community to adopt these guidelines and data formats is a new kind of publications based on large datasets and computer experiments. To this end we will work with publishers and journals in **M1.3**, and hence contribute to **O5**.

Requirements/own preliminary work (MaRDI Expertise)

The research group “Discrete Mathematics/Geometry” at TU Berlin has extended expertise in developing mathematical software systems and carrying out computational experiments [GHJ16; JJK18]. They will provide use cases and data sets from a wide range of subfields within computer algebra.

Work in this task area will particularly benefit from interacting with research projects of the DFG Collaborative Research Center TRR 195 “Symbolic Tools in Mathematics and their Application”. Its mission is to provide and use a comprehensive computational open source infrastructure to produce and manage vast amounts of data important to the mathematical community. Its core areas are: group and representation theory, algebraic geometry, commutative and non-commutative algebra, tropical and polyhedral geometry and number theory. The software development within the TRR 195, and its flagship project OSCAR, employs the new programming language Julia, which is becoming the first choice for developing new mathematical software systems. The TRR 195 projects will form a rich source of relevant use cases for MaRDI. Conversely, the entire TRR 195 will benefit from the new technical infrastructure for research data to be developed here. This gives a unique opportunity for setting up a new framework for research data in mathematics.

Both TU Berlin and TU Kaiserslautern have vast experience developing software and are aware of the risks of large software projects with many developers. Every possible measure to avoid errors and regression will be taken throughout development. This also includes usage of tools from professional software development, like git for version control, continuous integration frameworks like Jenkins and Travis, and a test driven approach to coding.

Risks and Mitigation

There are two main outside factors posing risks to this task area. One is the acceptance within the mathematics community, the other is finding a publisher willing to cooperate on finding a peer review process for software.

We plan to mitigate the risk of missing acceptance within the community by circulating prototypes of the guidelines, data formats, and predefined software environments as early as possible within the OSCAR community. Furthermore the different teams involved within MaRDI have vast experience even with publishing software, meaning that the ideas on confirmable workflow and data formats originate from a very strong fundament.

T2: Scientific Computing

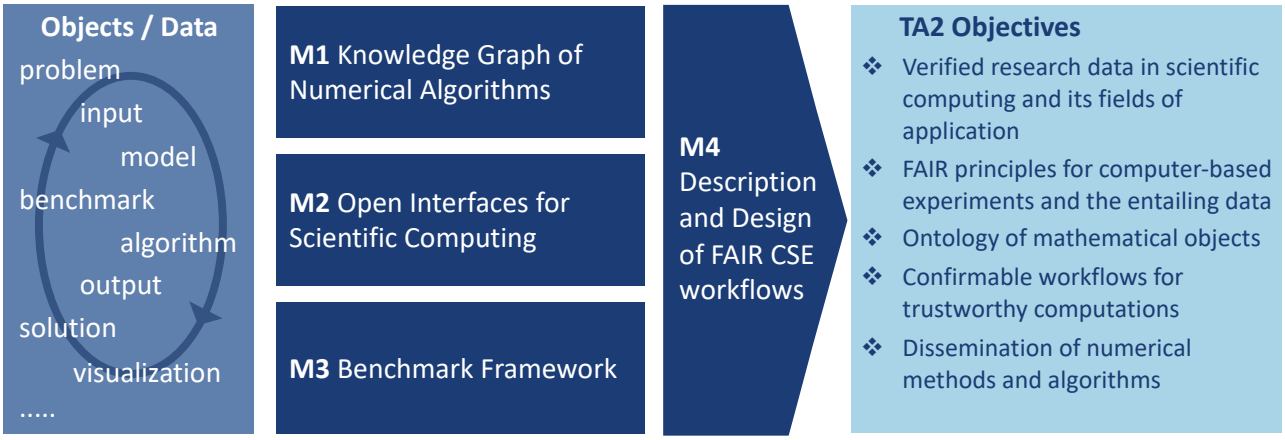


Figure 6: T2 measures and major objectives.

Scientific computing is a cross-disciplinary topic at the borders of applied mathematics, computational sciences and engineering (CSE), as well as other scientific areas involving numerical computations like digital humanities or computational medicine. Many of these interdisciplinary areas form the basis of use cases in T4. Here, we focus on certain aspects of scientific computing that focus on numerical computations, in contrast to the symbolic computations treated in T1. That is, the principal data type involved are fixed precision real numbers that are prone to roundoff errors in the scientific computing workflows. But scientific computing also handles various heterogeneous data structures, which depend on the chosen method and implementation details of the involved numerical methods. Beyond the data types also found in engineering, such as input/output data of numerical software, in computational mathematics, and specifically in scientific computing, also the algorithms, implementations, procedural data, and their metadata descriptions are research data. Moreover, standardized test problems, and prototypical demonstrator projects are in demand by the scientific computing community.

Via this task area, we will establish the FAIR principles for numerical computer-based experiments and the entailing data in scientific computing. The goal of the measures in this task area is facilitating discoverability, interoperability and comparability of (competing) algorithms and their implementations as well as discernibly tracking the subject's state-of-the-art, which is vital for scientific and efficient research conduct. Practically, pilot services and platforms to enable scientific computing data FAIRness will be implemented and experiences will be distilled into actionable guidelines.

Central objective of scientific computing is to develop mathematical algorithms which can be practically evaluated on a computer and yield approximations of mathematical models, while maintaining rigorous control over the approximation and roundoff errors. The developed algorithms, e.g., for solving large systems of equations, are foundational for all of scientific computing, and a vast body of methods has been proposed in the literature. There is a huge number of major numerical algorithms, which typically come with an even larger number of modifications and implementations. The "list of

algorithms" at Wikipedia[16] only gives a very limited overview. Traditionally, numerical algorithms are published in articles or books and implemented in a tremendous variety of software packages across all disciplines. We, e.g., refer to [EMNW11] for a documentation of classical numerical algorithms, which is, however, by far not comprehensive.

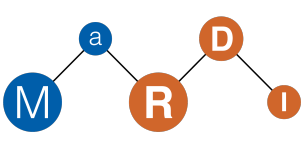
While mathematical algorithms are to be seen as one of the main research artifacts produced by scientific computing and despite their importance for a large part of science, there is currently no way of directly searching for data associated with a certain algorithm, such as journal articles discussing the algorithms, (benchmark) experiments and implementing software, or to discover algorithms that solve a certain mathematical problem. Databases for numerical algorithms are only available with limited scopes. In **M2.1** we establish a knowledge graph of numerical algorithms that will enable a directed search for algorithms applicable to given problem classes and to easily discover associated research data.

Often, a single algorithm is not sufficient to obtain a sought for result efficiently, but a cascade of methods is applied, yielding an often interdependent family of numerical models. For instance, given an inverse problem constrained by a PDE in space and time, an optimization problem may be solved, which depends on a forward problem discretized in space by the finite element method, and is then approximated in time by a Runge-Kutta scheme. This forward model, for example, may again have been replaced by a more efficient, surrogate model, which is updated from the original forward model during the optimization method. The entire inverse problem may again be only a part of a more involved scientific computing workflow. Usually, implementations of the methods producing the different models in such complex workflows are developed by independent research groups specialized in the respective mathematical methods, and often competing methods are realized in independent software packages. As a consequence, making the numerical models in such workflows interoperable requires significant implementation work, and individual modeling steps cannot easily be exchanged even though each building block may already be available.

Examples for established interfaces for numerical models that are implemented and used by multiple software packages are rare. At a very low level, the BLAS [Law+79] and LAPACK [And+99] standards define well-established interfaces for the manipulation of vectors and matrices. In the context of multi-physics modeling, several packages such as MpCCI [JK06] or PreCICE [Bun+16] exist that define and implement interfaces between individual simulation codes for the exchange of coupled solution fields. Originally developed for automotive applications, the Functional Mockup Interface [Blo+12] defines an open standard for the exchange and co-simulation of dynamic models. xSDK is a distribution of selected scientific computing libraries that adhere to joint packaging standards and which are regularly tested for their interoperability [SBD18]. However, interoperability is only ensured between individual libraries and no interface standards have been defined so far.

The goal of **M2.2** is to establish a widespread use of standardized interfaces throughout scientific computing by building a developer community around open scientific computing software interfaces and providing technical foundations for defining and implementing these interfaces. The measure is a core enabler for building flexible benchmark (**M2.3**) and simulation workflows (**M2.4**) within scientific

[16] https://en.wikipedia.org/w/index.php?title=List_of_algorithms&oldid=916759028



computing, but also for making these workflows available from other disciplines that are users of scientific computing (**T4**).

Typically, scientific computing publications are accompanied with *in silico* experiments to illustrate theoretical findings, validate proposed methods on real-life applications or compare competing methods in practice. A crucial weakness in this process is, that the test problem data set, the implementations of the existing methods and the execution environments are not standardized. In addition to being an inefficient scientific process, it is also not a fair comparison, as the test problem could happen to be a corner case on which a newly proposed method is working well, but not in general; or the competing methods could be implemented pessimistically. Thus, a benchmark framework, as addressed in **M2.3**, is of high relevance for scientific FAIRness, as it alleviates these problems, since researchers can adapt it to their community to test accepted benchmarks and use implementations from the respective experts to compare a new method or implementation.

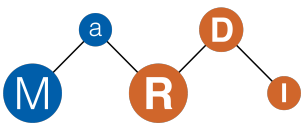
Technically, this involves implementing a generic benchmark framework, which can be specialized, for example, with community-specific comparison measures, which contributes to objective **O4**. A database is part of the benchmark framework maintaining benchmark data sets and benchmarking results as well as allowing visualizations or rankings. Furthermore, a common or at least well documented computer hardware platform then allows a fair comparison.

While there is a knowledge base for CSE workflows from a (software) engineering point of view [HW09; Bri+19] and while it has been acknowledged that for documentation, model descriptions and code can complement each other [Feh+16], a perspective on the underlying mathematical structures is not yet anchored. As for combining models, code, and data for the description of CSE simulations in a virtual lab notebook, Jupyter notebooks have gained popularity [Klu+16]. Also services like *code ocean*[17] target the combination of code and model descriptions. Still, little efforts is made to use abstraction for CSE workflow components in view of documentation tools that are generally applicable and that scale well with ever more demanding and sophisticated simulations. We note that many aspects of scientific workflows in CSE have commonalities with workflows involving exact and symbolic data (**T1**) as well as computational statistics (**T3**). We aim at a strong interaction between all measures in these task areas that deal with the respective workflows and plan to standardize these as much as possible while respecting domain-specific aspects as much as necessary.

Overall, four measures are planned, addressing these demands (cf. Fig. 6):

- Measure 2.1: Knowledge Graph of Numerical Algorithms
  - **Layers:** **X4: Knowledge**
- Measure 2.2: Open Interfaces for Scientific Computing
  - **Layers:** **X1: Core**
- Measure 2.3: Benchmark Framework
  - **Layers:** **X1: Core** and **X2: Data**

[17] https://codeocean.com



- Measure 2.4: Description and Design of FAIR CSE Workflows
  - **Layers**: **X1: Core** and **X3: Exchange**

Altogether, the measures in this task area address the full scientific process in scientific computing: In *Measure* **M2.1**, abstract numerical algorithms are made *findable*, while *Measure* **M2.2** makes their practical implementations *interoperable*. An implemented algorithm can then be compared to competing alternatives via *Measure* **M2.3**, rendering implementations *accessible*, while *Measure* **M2.4** facilitates their *reusability* in real-life applications. All four measures are deeply interlinked, since each measure facilitates, or is used by, the others, as delineated in the following.

Target user group Target users of this task area are mathematicians developing and validating numerical algorithms and their implementations, as well as all practitioners in scientific computing, such as research engineers and scientists that need to select a particular algorithm to solve a given numerical problem or understand its properties. While the focus in this task area will rest on models, algorithms and software used by the mathematical community, we will ensure in collaboration with **T4**, that the deliverables also meet the requirements of neighboring scientific fields. Potential users of this task area's work program are all actors in scientific computing, including industry; potential users will be addressed in GAMM Annual Meetings ($\approx$ 1,200 participants), SIAM Conferences on Computational Science and Engineering ($\approx$ 1,800 participants) and ICIAM Congresses ($\approx$ 4,000 participants).

Measure 2.1: Knowledge Graph of Numerical Algorithms

Synopsis and relevance In this measure we establish a knowledge graph of numerical algorithms, which interlinks those algorithms with the addressed mathematical problems and associated research data such as journal papers or implementing software packages. As often many variations of essentially the same numerical algorithm are discussed in the literature, an editorial board of domain experts will define the individual algorithm items in the database. The interlinking with the associated data, however, is carried out semi-automatically, using various data sources such as text-mining or metadata search in other databases and MaRDI services.

The measure improves the findability of research data associated with given numerical algorithms and allows non-experts, who need to solve numerical problems in their scientific work, to quickly gain an overview of available methods, their characteristics and implementations. Mathematicians in scientific computing are empowered to easily track the field's progress by data on competing methods.

Tasks

Prototype development. Based on Wikidata or other data standards like OAI-PMH or schema.org a graph ontology is defined and realized in a software prototype. The algorithms will be interlinked with mathematical objects/models to which they are applicable, publications and software implementations. Formal definitions of the algorithm will be given by linking to original articles introducing the algorithm or corresponding entries in existing knowledge bases such as such as Wikipedia [Wik04], the Encyclopedia of Mathematics [SE19] or domain-specific resources such as the MORwiki [MOR19].

A basic web frontend for querying the graph or browsing it by category is developed. For editors, a dashboard is realized that allows to easily modify the knowledge graph based on auto-generated data from integrated services or user suggestions.

Integration with the MaRDI portal and other services. As the knowledge graph will interlink various types of mathematical research data across MaRDI, a close integration with the MaRDI portal and other services is a central development goal. In particular, a mapping of ids between the algorithm knowledge graph and data from other MaRDI services is realized via the MaRDI portal in coordination with **T5** at an early stage of development. Further, a public API for querying and modifying the graph is developed for communicating with other services, as well as end-user clients such as Jabref. The integration with other MaRDI services, in particular with zbMATH [ZBM], swMATH [SWM] and the *Benchmark Framework* **M2.3** will be carried out in collaboration with **T5** (**M5.6**). Conversely the zbMATH review system is integrated to provide suggestions for extending the graph.

First algorithms. As soon as a first software prototype is available, an initial knowledge graph of model order reduction algorithms is set up based on the data already available in the MORwiki. This database will then be gradually extended by methods from numerical linear algebra and numerical PDEs.

Production deployment. Based on the experience with the software prototype, a production version is developed and deployed. In collaboration with **T5**, sustainable hosting and maintenance of the service for a large user base, as well as a stable integration with other MaRDI services is ensured.

Community building and support After the initial setup of the graph with algorithms from our own fields of expertise as a demonstrator, building and supporting a diverse editorial board will be a crucial task for long-term success of the knowledge graph. To this end, advertisement and feedback through MaRDI communication channels (**T6**) is an important part of this measure. An editorial office is set up as an official contact for editors and users of the service. As soon as an initial group of editors has been established, main editorial policies, e.g. objective inclusion criteria for new algorithms, will be laid out in collaboration with these editors. As part of this task, we will reach out to further data sources such as MathSciNet [Ame19] or journal publishers to interlink their metadata with the knowledge graph.

Deliverables (X4: Knowledge)

D-TA2-A1 Graph ontology and editorial guidelines. The ontology of the knowledge graph is defined. Editorial guidelines ensure the consistency of the graph over the various sub-disciplines.

D-TA2-A2 Initial knowledge graph. A knowledge graph of numerical algorithms is established for selected fields of scientific computing.

D-TA2-A3 Frontend, API and service integration. A web frontend allows querying and modifying the knowledge graph. A public API integrates the graph with the MaRDI portal and other services.

D-TA2-A4 Editorial board. An editorial board of experts from various fields of scientific computing further extends and maintains the knowledge graph.

D-TA2-A5 Editorial office. An editorial office organizes the interactions between the editorial board members and provides technical assistance.

Milestones

M-TA2-A1 After 12 months the software prototype is finalized and deployed on private test system. First entries from model order reduction have been created.

M-TA2-A2 After 24 months a beta-version of the service is publicly available. Several editors have extended the graph for further fields of scientific computing. The graph ontology is completed; First editorial guidelines have been set out.

M-TA2-A3 After 48 months the service is fully in production, integrated with other MARDI services. The editorial board is established, and the editorial office has been set up.

Measure 2.2: Open Interfaces for Scientific Computing

Synopsis and relevance The goal of this measure is to develop and establish open interface standards between numerical software packages that allow to seamlessly interconnect and exchange the models and algorithms realized by these packages in complex modeling or simulation workflows. Specifically, in this measure we realize such interfaces for discrete PDE models based on a reusable language-agnostic core API toolkit. Through a developer platform we assist other scientific computing communities with establishing open interface for their respective software stacks.

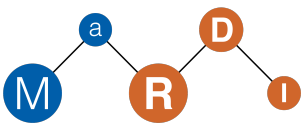
The measure contributes to **X1: Core** and serves as the technical basis for **M2.4** and the integration of MaRDI workflows with models provided by software packages from other disciplines (**T4**).

Open interfaces will improve the reusability of numerical models and facilitate their recombination in complex simulation workflows. By enabling researchers to reuse existing realizations of numerical models, significant development time can be saved, while collaboration between experts in different fields of scientific computing is fostered. Moreover, interface standards for numerical models improve the comparability of numerical methods by facilitating computations with competing algorithms for the same model. In particular, the developed interfaces will serve as a foundation for the benchmark framework in **M2.3**.

Tasks

Core specification and API toolkit development. A prototype C-language based API toolkit is developed that allows accessing interfaces implemented with the toolkit by loading the other software component as a shared library plugin. Language bindings for this toolkit are implemented for C++ [Str13] and Python [VR95].

After a first version of the PDE solver interfaces developed in this measure has been implemented and tested using the prototype toolkit, the toolkit is gradually extended to other technical means of coupling the components, in particular inter-process and network communication. Bindings for further relevant programming languages in scientific computing are developed, in particular for Julia [JP] and MATLAB [Mat] / Octave [OD]. During this process, common core specifications for interfaces developed within this measure are laid out and realized in the toolkit. This includes, e.g., error handling conventions, definition of atomic data types or data ownership policies. Interfaces implemented with this core toolkit will automatically be interoperable with any other software that uses the toolkit, independent of the implementation language. Different means of communication between the components can be chosen easily without additional programming effort.



Establish PDE solver interfaces. With DUNE [Bas+08] and pyMOR [MRS16], the WWU has extensive experience in the development of scientific computing software that interfaces with other numerical codes. pyMOR [MRS16] is a software library of model order reduction algorithms that are formulated in terms of abstract interfaces for seamless integration of full order models realized by external partial differential equation solver packages. So far, pyMOR has been successfully used with several PDE libraries such as DUNE, deal.II, FEniCS or NGSolve.

In this task, pyMOR's interfaces serve as basis for the development of PDE model interfaces that generalize the one-to-many coupling of pyMOR with these libraries to a many-to-many coupling, between the PDE solvers and various PDE user codes like model reduction, uncertainty quantification or optimization packages. To this end, pyMOR's interfaces and the existing PDE solver bindings in pyMOR are reimplemented using the open interfaces core API toolkit.

From the beginning, we keep close contact with the developers of major open source PDE toolkits and user codes and work with them on the final interface specification and the implementation of the interfaces in their packages.

In collaboration with **T4** we will further develop interfaces for coefficient functions and geometries defining analytical PDE models, as well as interfaces for the solution fields of the resulting discrete PDE models that are realized by the solver packages.

Community building and support. The objective of this measure is to establish open interface standards in many areas of scientific computing. This is only possible through a broad engagement of the relevant developer teams. To this end, outreach to various mathematical communities promoting open interfaces is an ongoing task. MaRDI will offer a web-based platform to support the specification process of new interfaces and their implementation driven by the respective communities and provide technical assistance to developers.

Deliverables (X1: Core)

D-TA2-B1 Core specifications and API toolkit. Common technical core specifications are laid out. A C-language based API toolkit with bindings for major programming languages in scientific computing allows dynamic run-time access via shared libraries, inter-process or network communication to any software implementing a given interface.

D-TA2-B2 Interfaces for PDE models. Interfaces for PDE models are defined and implemented in several relevant PDE solver packages. Multiple user codes that depend on at least one these packages are adapted to utilize the new interfaces, making these codes interoperable with all targeted packages.

D-TA2-B3 Community platform. A web-based community platform which hosts the interface specifications and allows their further development and discussion is established. MaRDI provides technical support for developers through this community platform.

Milestones

M-TA2-B1 After 24 months prototypes of core specifications and API toolkit have been developed. Interfaces for PDE models are integrated in at least two PDE solvers and user codes.

M-TA2-B2 After 48 months PDE solver interfaces are standardized and officially supported in multiple PDE solvers. Open interfaces community is established and has been extended to at least one other application field.

Measure 2.3: Benchmark Framework

Synopsis and relevance A common theme in scientific computing is the race for the most efficient, accurate, universal and elegant algorithm for a class of problems. Yet, this principally healthy competition is only beneficial to mathematics, science, industry and society if the research output is actually comparable to prior results. Even then, the comparison of algorithms can be a complex endeavor as the implementation, configuration, compute environment and test problems need to be well defined. For a single mathematician this is partly addressed by best practices for mathematical software, see for example [Feh+16; Feh+19] and references therein. But, due to the increase in computer-based experiments in mathematics, also communities, around specific problems need infrastructure for exchange, comparison and tracking the progress in the field.

To this end, we propose a benchmark framework, which is a generic toolkit to compare implementations of algorithms using benchmarks native to a community. Its value lies in the capability to fairly compare and validate existing methods for new applications, but also new methods to existing ones. Thus, it contributes to "establish confirmable workflows" (**O3**), even beyond scientific computing to all algorithmic branches of mathematics, see for example **T1** and **T3**. This measure contributes to RDI Layers **X1: Core** and **X2: Data**, as it provides important definitions and the benchmarking results for the MaRDI-portal (**T5**).

Tasks

Assembly of domain-independent specifications. The assessment of a generic benchmark run constitutes a function of the implementation of an algorithm, the specific realization of a benchmark problem and the precise execution platform (i.e. software and hardware). While the algorithm will be determined using the knowledge graph from **M2.1**, the abstract interaction with the benchmark problem requires the use of the open interfaces from **M2.2**. The identification of the benchmark problem will use the unique identifiers assigned in the MaRDI-Portal. This measure will derive specifications of the objects recorded there.Synergies are used from the exchange of experience with **M3.1**, **M3.2** and also **M4.2**.

The actual execution of the benchmark run overlaps with the workflows in **M2.4** and will be determined in close collaboration. In contrast to the platform documentation, there, for direct comparability of the assessments, here, it is crucial to develop clear specifications of distinguished and well defined execution platforms, covering both software and hardware properties.

Database of curated benchmarks for various model classes Data from and for the model order reduction (MOR) community is already collected in the MORwiki [MOR19], a collection of living documents, based on the MediaWiki software [Med19]. It features three main categories: Benchmarks [CVD02; KR05], methods and software. An editorial board curates submissions and edits. Data sets for linear and parametric-linear models are well represented in the existing collection. Data sets for non-linear or procedural models and models for which only evaluation data, rather

than equations are available need to be added or extended. Properties, and interesting characteristics used for the later benchmark selection and assessments are recorded in the model meta data. The current manual selection of data sets will be supplemented by an automatic selection API adapting the ideas from the SuiteSparse Matrix Collection [Sui].

Development of a Demonstrator. The MORwiki collection will be the data basis for the demonstrator, a model reduction benchmark tool. To this end, experiences from [Bau+17] will serve as a prototype and will get extended to the remaining model classes and methods. The MORwiki will serve as a proof of concept for a living document progress tracker of a field in mathematics in combination with fair comparability of new findings and methods. Its core information will be mirrored in the MaRDI-Portal (**T5**). The benchmark tool will be executed on defined hardware to fill the assessment database, but also made available to enable users to make assessments further problems on their own hardware.

Generalization to the generic benchmark framework. The (generic) benchmark framework aims to allow as many as possible branches of computational mathematics to compare their implementations. The above generalization to the other model classes in the MORwiki will be a crucial learning experience towards the full abstraction required here. Collaboration with **M2.1** is required for the determination of competing implementations. The integration of the results into the MaRDI-Portal (**T5**) is key for the findability of the benchmarking assessments by researchers outside the specific domain searching for the best implementation for their problem.

In a modular setup, the framework will provide an API for the interaction with the knowledge graph (**M2.1**), the benchmark problems and the MaRDI-portal (**T5**). Specialization of the interaction modules enable seamless integration of community collections into the MaRDI infrastructure. In the MaRDI consortium the model databases established in **T3** and **T4**, and beyond MaRDI community collections, such as the ones in optimization constrained by PDEs [Her+; Her+14] or discrete optimization [Mip], are expected pilot users for framework and an implementation for algorithms from computer algebra (**T1**) is projected.

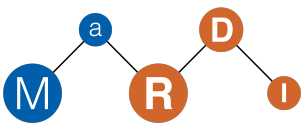
Presentations on domain-specific workshops (e.g. meetings of related activity groups of GAMM and SIAM), as well as at research software engineering (RSE) events (SORSE, deRSE Conference, RSEConUK) ensure that the modules are flexible enough and follow appropriate software engineering standards.

Deliverables

D-TA2-C1 Domain-independent workflows for benchmarking. (**X1: Core**)
D-TA2-C2 Universal definitions for benchmark collections enabling automated comparison. (**X1: Core**)
D-TA2-C3 General purpose guidelines for the specification of execution platforms. (**X1: Core**)
D-TA2-C4 Curated model reduction benchmark database. (**X2: Data**)
D-TA2-C5 MORwiki benchmark tool as a demonstrator.(Service, **X2: Data**)
D-TA2-C6 Generic modular benchmark toolkit. (Specification, **X1: Core**)

Milestones

M-TA2-C1 After 6 months workflow and benchmarking specifications have been stabilized.



M-TA2-C2 After 18 months first reference execution platforms have been defined and guidelines for future platform specifications have been drafted.

M-TA2-C3 After 36 months the MORwiki benchmark tool generates assessment results for all benchmarks and algorithms in the Wiki.

M-TA2-C4 After 48 month assessment results are recorded in the MaRDI portal, linked to both the benchmarks and algorithms.

Measure 2.4: Description and Design of FAIR CSE Workflows

Synopsis and relevance Algorithms from numerical mathematics and data are the backbone of simulations in engineering problems, where, practically, different numerical methods are chained to design workflows. This measure analyzes general and particular components and provides an abstract multi-layered description of CSE workflows: Each component will be characterized through an input/output description so that model, code, and data can be used interchangeably and, in the best case, redundantly. For that, we develop suitable meta data and a low level language for the descriptions of general CSE workflows. The findings and experiences are distilled into a guideline, an electronic documentation tool, and example applications that are integrated in the MaRDI portal.

Tasks

Develop a framework for the abstraction of workflow components Each component will be characterized in terms of its input to output relation. Here, *the input* stands for parameter that set up the current part of the workflow and *the output* denotes the (intermediate) results that are passed on to the next component. In this way, a component can be described in a multi-layered fashion, namely via a mathematical/physical model, via a model that has been inferred from data, via a software that computes the output for a given input, or via plain data. For any description there is associated meta data that embed the components in an ontology of mathematical objects or that carry practical information like accuracy of the algorithm or reliability of the data. The basis for the analysis will be application oriented use cases of other NFDI consortia (in cooperation with **T4**) and CSE examples considered in the MPI-DCTS.

Analyze examples and produce show cases for a systematic description of CSE components with models of different kind, code, and data used equivalently and – for best reproducibility – redundantly. One benefit of combining data and code lies in the flexible treatment of the associated simulation data. For example, the storage requirements of huge time series can be reduced by replacing the full data by parts and associated code that can provide the missing points on demand. Also, simulation parts that are defined as the result of empirical statistics can provisioned with the relevant code and statistical information and further improved as needed. Another benefit of the input/output perspective is the interchangeability of the concrete realization so that, e.g., for reproduction, a closed-source implementation can be substituted by an open-source equivalent. For the needed meta data description, efforts will be coordinated with other NFDIs via **T4**.

Design a prototypical lab notebook that supports the application of the abstraction of CSE workflows: Such a notebook provides a complete and possibly multilayered and redundant documentation of the workflow and instructions or direct interfaces to rerun particular components. In particular,

the abstraction of the components can be used resort to alternative implementations or realizations in a, say, reproduction scenario.

MaRDI integration and setup of example instances of CSE workflows in the lab notebook that are findable and accessible through the MaRDI portal (cp. **M5.4**) for inspection and reuse.

Deliverables

D-TA2-D1 A domain specific language for CSE workflows, that enables the abstract description of CSE workflows in terms of components and their realization. The implementation of this language is directed for automation of the analysis based on suitable meta data. (**X1: Core**)

D-TA2-D2 Documented example applications of the language to use cases. (**X2: Data**)

D-TA2-D3 An electronic lab notebook to be used for documentation and accessing CSE workflows in the abstract multilayered language. (**X3: Exchange**)

D-TA2-D4 A remotely accessible example instantiation of the lab notebook and an interface for integration in the MaRDI portal. (**X3: Exchange**)

Milestones

M-TA2-D1 After 12 months the conceptual analysis of the use cases and prototypical CSE workflows in terms of input/output description of the components is complete.

M-TA2-D3 After 36 months a set of suitable meta data for components of CSE workflows and a language to describe and realize CSE workflows by meta data have been defined.

M-TA2-D3 After 48 months a user interface for the language is available.

Services

- Knowledge graph of numerical algorithms with public API, integrated with MaRDI portal.
- Open interfaces developer platform.
- MORwiki benchmark tool as a demonstrator.
- Benchmark Framework

Embedding into NFDI

M2.1 contributes to the overarching objectives to “enable the development of mathematical services” (**O4**) and to “standardize semantics relations” (**O6**). While this measure will focus on algorithms from numerical mathematics where the need for improving the findability of algorithms and related data is most pressing, the established structures will be open for other branches of computational mathematics. Since this coincides with objectives from **T1**, a close cooperation is foreseen. The integration with the MaRDI portal an other services is carried out in collaboration with **T5**.

M2.2 contributes to the overarching objective to “interoperable mathematical research data” (**O1**) and will be implemented in cooperation with **T4**.

Measure **M2.4** contributes to the MaRDI objective **O3**, but also to “securing confirmable and reproducible results” (**O2**), yet due the abstraction, it also contributes to **O1**, with the ultimate goal of FAIR CSE workflows. This means the pursued objectives are shared with **T1**, but from a scientific computing perspective and an immediate application for the *Open Interfaces for Scientific Computing*

measure **M2.2**. In cooperation with **M3.3**, we will address reliability of workflow components inferred from data. Application relevant examples of other NFDI are considered in coordination with **T4**.

A close connection will be established with the FAIRmat consortium (Area C) that addresses the normalization of input/output data associated with different codes for the same physical system. FAIRmat will provide us with a case study where data normalization makes models and data interchangeable. We will support FAIRmat in the development of automated data normalization procedures based on abstract workflow descriptions and metadata. We also agreed on commonly organizing a workshop on metadata in 2022.

Based on their infrastructural use case *Matrix-Inclusion Microstructure Ontology* we will join efforts with the MatWerk consortium; see also **T4**. While the MatWerk use case will serve us as engineering application examples for **M2.4**, our unifying representation of data, model, and simulation will contribute to the MatWerk objective of developing an ontology to ensure interoperability of heterogeneous data for the description of matrix-inclusion materials.

An application of the concepts of the Task Area **T2** as a whole will be realized together with the PUNCH4NFDI consortium to address the enormous computational challenges in *Heavy Ion* research; see **T4** (Case Study 5). To streamline the related (cloud) computing workflow, we implement descriptions of the components that allow for interchanging of, for example, the optimization backends and thus, enhance reliability. The cooperation with PUNCH4NFDI will further include a consolidation of software data bases (**M2.1**), the definition of a benchmark case by PUNCH4NFDI (**M2.3**), and joint developments on standardized software APIs (**M2.2**).

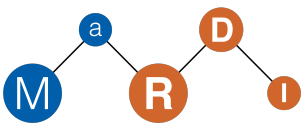
Requirements/own preliminary work (MaRDI Expertise)

The co-spokespersons heading this task area are diversely connected to many areas of scientific computing; Peter Benner is a director at the MPI DCTS, leading a group on basic research in scientific computing, Mario Ohlberger is a principal investigator of the cluster of excellence “Mathematics Münster” at the WWU.

The WWU has extensive experience in the development of scientific computing software and in building interfaces between numerical codes. The DUNE [Bas+08] distributed and unified numerics environment is a modular toolbox for solving partial differential equations built around a generic grid interface that allows the usage of third party grid managers such as ALBERTA [SS05], ALUGrid [Alk+16] or UG [Bas+97]. pyMOR [MRS16] is a software library of model order reduction algorithms that are formulated in terms of abstract interfaces for seamless integration of full order models realized by external partial differential equation solver packages. pyMOR’s interfaces will serve as basis for the development of PDE model interfaces in this measure.

Examples for interfaces for numerical models that are implemented and used by multiple software packages are rare. At a very low level, the BLAS [Law+79] and LAPACK [And+99] standards define well-established interfaces for the manipulation of vectors and matrices. With FlexiBLAS [KS14], MPI DCTS has developed a BLAS and LAPACK wrapper library that allows to exchange different BLAS / LAPACK implementations at runtime.

A database for model order reduction algorithms is, e.g., available in the MORwiki [MOR19]. While with swMATH [SWM], a database for mathematical software packages exists that enables access to



related articles via zbMATH and vice a versa.

Besides the use cases provided by the consortia MatWerk and FAIRmat, the basis for the analysis of the workflows are application oriented simulations of molecular structures (see [SS18]) as conducted in our partner group at the MPI DCTS, and the extensive experiences gathered at the MPI DCTS, see for example: [Gru+14; Yue+15; Ben+17; Ben+18; Ben+19].

Risks and Mitigation

The main risk for the success of this task area is that the developed services and standards might not be adopted by the scientific computing community.

T2 is coordinated by leading research groups in scientific computing and, in particular model order reduction. Measures **M2.1**, **M2.2**, **M2.3** build on services and software libraries already established by these groups in this field. Thus, success within the model order reduction community is very likely. To mitigate the risk regarding adoption in the greater scientific computing community, we have made outreach and community engagement activities important elements of all measures in this task area. Scientific computing is an interdisciplinary field with connections to various other scientific disciplines. By our collaboration with **T4**, in particular regarding **M2.4**, we ensure that our efforts are coordinated with related activities of other NDFI consortia.

T3: Statistics and Machine Learning

Research in Statistics and Machine Learning (ML) focuses on the development of broadly applicable methods for data analysis that solve prediction problems, support decision-making, and infer structure underlying a scientific phenomenon. These methods draw on a wide variety of computational techniques that include numerical methods from scientific computing (T2) and also symbolic computation as considered in computer algebra (T1). However, the data processed in Statistics and ML have the distinguishing feature of being *uncertain*, i.e., subject to *stochastic noise*. Separating this noise from the signal of interest is a key challenge that arises in virtually all branches of Science and Engineering. Addressing this challenge with methods that ensure that conclusions drawn in scientific studies are likely to persist in independent replication studies is a chief goal of Statistics and ML.

The MaRDI task area for Statistics and ML is concerned with the research processes that surround methods development in the field. These processes feature an interplay between experimental and theoretical work. Mathematical theory is frequently needed to design a particular method and to clarify possible optimality properties. However, theory is often only feasible under simplifying assumptions, and it is equally important that methods are subjected to careful simulation experiments that evaluate performance relative to existing competitors. Additionally, benchmarking on real-world data is crucial to ensure that new methods are indeed able to solve targeted practical problems.

In the context of MaRDI, and the NFDI more broadly, the task area seeks to address the needs that the Statistics and ML community has in the management of its research data. These research data range from literature, statistical models and algorithms to benchmark data sets and software. The task area will follow FAIR principles in initiating libraries of curated datasets that will be connected to software and research literature by providing an associated library of statistical analyses. The task area will further support FAIR development of new methods by creating workflows and a demonstration platform for how to evaluate, compare or also tune methods through empirical analyses and simulation studies. Finally, the task area will cooperate with journal partners to establish standards for quality control and reproducibility of numerical experiments in the scientific publication process.

Use cases. The goals of the task area will be pursued in the context of two focused use cases. At a high level, problems in Statistics and ML can be divided as being of supervised or unsupervised type. Selecting a representative from each branch, the guiding use cases for the task area will be:

1. Supervised Machine Learning for Regression Problems.
2. Unsupervised Model Selection in Graphical Modeling.

The first use case reflects the ubiquitous nature of regression problems, which are supervised as one learns from labelled data (in the form of values of an outcome variable). In this setting, experimental evaluation may also make reference to labelled examples. Model selection in graphical modeling, on the other hand, is an unsupervised task, where the data to be analyzed do not include ground truth labels. This makes the availability of data with additional domain information all the more important.

Leadership and Partners. Bernd Bischl, Professor for Statistical Learning and Data Science at the Ludwig-Maximilians-University Munich (LMU), will lead the work on Supervised Machine Learning. Mathias Drton, Professor of Mathematical Statistics at the Technical University of Munich (TUM), will

lead the work on Unsupervised Model Selection. The task area will be supported by the R Foundation, which leads the R project for Statistical Computing [R C20]. With strong community support, the R project provides a free software environment for statistical computing and graphics that is the basis for much of the academic research in Statistics and ML. The Biometrical Journal and the Journal of Statistical Software will be partners for efforts in next-generation peer review.

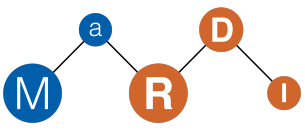
Four measures are planned in close cooperation between LMU and TUM:

- Measure 3.1: Library of Curated Benchmark Datasets (Service)
  - **Layers:** **X1: Core**, **X2: Data**
  - **Users:** Researchers in Statistics and ML that conduct data analyses as part of methods development. Practitioners from other fields that seek to publish datasets on a platform that facilitates subsequent statistical analyses.
- Measure 3.2: Library of Statistical Analyses (Service)
  - **Layers:** **X1: Core**, **X2: Data**, and **X4: Knowledge**
  - **Users:** Researchers in Statistics and ML as well as practitioners that are looking for demos of statistical analyses for adaptation to other problems or as training material. Researchers and practitioners that use the developed platform to publish their own data analyses.
- Measure 3.3: Empirical Analysis of Machine Learning Experiments (Demonstrator)
  - **Layers:** **X1: Core**, **X2: Data**, and **X4: Knowledge**
  - **Users:** Researchers that conduct numerical experiments for methods development, wish to publish experimental results or use such results for customization of methods.
- Measure 3.4: Standards for Peer Review of Numerical Experimentation (Specification)
  - **Layers:** **X3: Exchange**
  - **Users:** Authors, publishers, editors and reviewers of scientific publications.

Measure **M3.1** will use rich meta-information to make the compiled datasets FAIR and give a context of statistical tasks to be solved for each dataset. Measure **M3.2** will connect data to methods of analysis (software and literature) by establishing a library of statistical analyses of the datasets from Measure **M3.1**. Measure **M3.3** will provide statistical workflows for benchmarking of new versus existing statistical/ML methods, a problem for which it is crucial to account for the noisy/uncertain nature of the data considered in the field. Measure **M3.4** will establish standards for peer-review, with the goal of supporting both authors and reviewers to ensure that experiments reported in peer-reviewed publications meet a high standard of reproducibility.

Measure 3.1: Library of Curated Benchmark Datasets

Synopsis By its very nature, research in Statistics and ML requires datasets on which new methods can be illustrated and tested. This need has led to different projects that offer data repositories. A well-known example is the UCI Machine Learning Repository (hosted at the University of California, Irvine) [AN]. A more recent community effort is the OpenML project [Van+13]. However, the existing data repositories offer limited meta-information and typically lack curated libraries of benchmark datasets



that focus on specific data-analytic tasks or problem domains. This lack makes it difficult for users to find data that provide different levels of empirical information for a desired task.

In this measure, we aim to provide task-specific libraries of curated benchmark datasets following the FAIR principles and making them accessible through the MaRDI portal. For this purpose, we will use, adapt and extend the OpenML framework. This enables us to provide a convenient and unified infrastructure that also allows external users to create and publish custom libraries of curated benchmark data sets that focus on their data-analytic task or problem domain. In our work we are able to build on and expand recent work in [Bis+17] that adopted the OpenML framework to start the OpenML-CC18 dataset collection, which is a curated library specifically for classification tasks.

State of the art Many dataset collections exist on different websites and platforms. Among these, the OpenML project—a collaborative platform for sharing data and reproducible ML experiments—is closest in spirit to the measure planned here. OpenML provides a REST API with different software clients that allow importing datasets into popular software environments (e.g., R, Python, or Java) and uploading ensuing results. A growing number of users has led to over 20.000 datasets on OpenML (as of August 31, 2020), creating a situation where an overwhelming number of datasets is available for certain popular tasks (in particular, for classification through the OpenML-CC18 suite [Bis+17]), whereas few to none are designated for other tasks (e.g., of unsupervised learning). Besides general libraries, there are also dataset libraries focussing on specific issues. For example, the AutoML benchmarks from [Gij+19] are comprised of particularly difficult classification problems. While the existing projects make valuable contributions, it is evident that there is now a strong need for curated and FAIR libraries of benchmark datasets that support rigorous and standardized benchmarking of statistical methods across a broad spectrum of data-analytic tasks.

Target user group The services created in this measure support researchers developing new statistical methods/ML algorithms by allowing them to find and access well-selected and curated datasets for use in benchmark experiments and comparisons of algorithms at large scale. A second group of targeted users consists of practitioners from other fields that seek to publish own collections of datasets from their domain on a platform that facilitates subsequent statistical analyses.

Added value The measure provides libraries of benchmark datasets on OpenML with rich and customizable meta-information that facilitates Statistics and ML research on different levels. Findability of the library and specific datasets is facilitated through the MaRDI portal and easily queryable task-specific meta-information. Simple access, interoperability and reusability are ensured by establishing a unified infrastructure that allows to create, extend and retrieve libraries of datasets using the OpenML API and OpenML clients (e.g., R, Python, Java, C#).

Implementation Our work will be guided by two focused use cases: regression as a ubiquitous task of supervised learning, and model selection in graphical modeling as an unsupervised task. We will create extensions of the existing OpenML infrastructure (database, standards of meta-information, and client software) to support addition of datasets from different domains and use in different Statistics and ML tasks. We will then collect suitable datasets from the two aforementioned use cases and provide for each dataset appropriate meta-information, a wiki-like description, and a PID. This library

of datasets will be uploaded and made accessible on OpenML, and curated with the data description and meta-information. Specific guidelines will be developed to simplify and automate this process to provide other users the ability to roll out their own library of datasets. These guidelines and automated tools will be demonstrated in the context of concrete extensions of the library in the setting of neuroimaging and electro- and optophysiology, see Case Study 6 in task area **T4**.

Deliverables

D-TA3-A1 Library of datasets and associated meta-information for supervised learning and specifically for regression tasks.

D-TA3-A2 Library of datasets and associated meta-information for unsupervised learning and specifically for model selection in graphical modeling.

D-TA3-A3 Infrastructure to expand this library and create further customized dataset libraries (automation, standards).

D-TA3-A4 Integration of a library of datasets from neuroimaging and electro- and optophysiology for statistical data analysis.

Milestones

M-TA3-A1 OpenML workshop A sizeable group of dedicated researchers from all over the world contribute to the OpenML platform. In order to ensure its continuous development and maintenance, the OpenML core developers, which include TA co-leader Bernd Bischl, organize at least one workshop per year. This milestone consists of one such workshop focusing on connecting elements of the MaRDI portal including the measure described here to OpenML.

M-TA3-A2 Unified infrastructure to create dataset libraries Completion of a service for OpenML to create and maintain curated libraries of datasets for different tasks, to automatically generate basic meta-information, and to include semantic and task-oriented information.

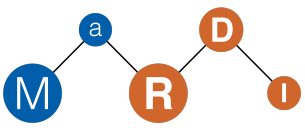
M-TA3-A3 Curated datasets for supervised learning Completion of a library with at least 10 curated datasets for regression tasks.

M-TA3-A4 Curated datasets for unsupervised learning Completion of a library with at least 10 curated datasets for model selection tasks in graphical modeling.

Measure 3.2: Library of Statistical Analyses

Synopsis Users access the library of datasets from Measure **M3.1** in order to perform statistical analyses. This second measure supports this step by conducting exhaustive statistical analyses of the datasets from the developed library. These analyses play the role of demos that connect the datasets to statistical/ML methods that solve the task(s) specified in the data's meta-information. Each analysis will be presented in a notebook that clarifies the aim of the analysis, gives its results, and showcases specific software (with focus on R and interfaces to OpenML). The notebooks will also link to literature describing the considered methods and software. Equipped with PIDs the notebooks will be collected in a library of statistical analysis on a searchable and extendable platform.

State of the art The Statistics and ML community has created many dataset collections and an almost overwhelming number of software packages, e.g., the R project for Statistical Computing [R



C20] currently features over 16,000 extension packages. While most software packages provide some limited demonstration examples, expert knowledge is required to navigate the available implementations and select software suitable for a specific tasks. There is thus a need to provide data- and task-oriented collections of demos that furnish entry points to the by now complex web of Statistics and ML methodology and software. A connection between dataset and analysis is provided by platforms like Kaggle [KAG], where users can present results of the method(s) they used to analyze the data. But these platforms typically lack connections to corresponding literature and limit a user's ability to comprehensively search for information on specific tasks or Statistics/ML methodology.

Target user group A first user group consists of researchers and practitioners looking for demos of statistical analyses for adaptation to problems in other domains or also training purposes. A second user group will use the platform as a medium for publishing their own statistical analyses. In the context of MaRDI's **T4**, both types of users are encountered in cooperation with NFDI consortia from other disciplines. To give one example, the PUNCH4NFDI consortium is interested in both general Statistics and ML resources for training of physicists and creation of physics-focused use cases.

Added value The created library connects data, methodology, software, and literature. It makes methods and software findable and accessible from a task-oriented point of view. The use of R notebooks ensures reproducibility and interoperability. The notebooks in the library constitute a resource for designing training courses in Statistics and ML methodology.

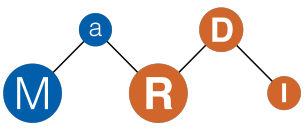
Implementation The statistical analyses will focus on the use cases also considered in Measure **M3.1**. Software implementations from the R project will be employed by using and extending existing interfaces between R and OpenML [Cas+17]. R and its markdown capabilities will be used to create notebooks, which will be given PIDs. Connections to literature will be established using the zbMATH [ZBM] and swMATH [SWM] services provided in MaRDI. The platform presenting the analyses will be created using suitable repository and wiki-technology, and extended to allow for user feedback on notebooks.

Deliverables

D-TA3-B1 Library of statistical analyses for regression tasks.
D-TA3-B2 Library of statistical analyses for model selection in graphical modeling.
D-TA3-B3 Infrastructure to expand this library and alongside extensions to the data library from Measure **M3.1**.
D-TA3-B4 Demonstration of library extension in the setting of statistical analyses for neuroimaging and modeling problems.

Milestones

M-TA3-B1 Unified infrastructure to analyze library datasets Completion of a service to access task-relevant datasets hosted on the OpenML platform and to analyze these data using routines implemented in R.
M-TA3-B2 Notebooks for supervised learning Completion of a library with at least 10 notebooks for regression tasks.



M-TA3-B3 Notebooks for unsupervised learning Completion of a library with at least 10 notebooks for model selection tasks in graphical modeling.

M-TA3-B4 Journal publication Publish a scientific article that presents the dataset libraries from Measure **M3.1**, the library of statistical analyses from the present measure, and the established infrastructure to access and to create extensions of these libraries.

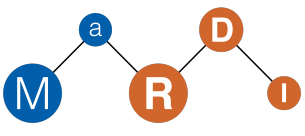
Measure 3.3: Empirical Analysis of Machine Learning Experiments

Synopsis Benchmark experiments that apply different algorithms to a collection of datasets play a key role in the development of ML algorithms. The results of such experiments include valuable meta-information about algorithm performance for different data characteristics and track algorithm settings (e.g., chosen hyperparameters). Careful statistical analyses of these results are required to draw general conclusions, e.g., about the importance of hyperparameters or the ranking of algorithms. In this process, already the design of benchmark experiments is challenging. Indeed, many decisions must be made before conducting the experiment, including for instance the choice of a specific design of experiments (DoE) for selection/sampling of hyperparameters. Other questions pertain to the number of selected datasets or repetitions necessary to draw significant conclusions from the experiments. Finally, the statistical analysis of the benchmark results has to reflect the chosen DoE.

This measure supports researchers in the process of benchmarking on real-world data. As a concrete use case, we will conduct exhaustive experiments based on the dataset libraries compiled in Measure **M3.1**. In this context, we will extend the OpenML infrastructure to allow the community to access prior benchmark results and publish their own results. We will develop new tools to simplify the aggregation and analysis of benchmark results as needed for a systematic experimental comparison of ML algorithms or more complex ML pipelines approaches. We will also summarize common pitfalls in the design of benchmark experiments and their statistical analysis, and list best practices and concrete guidelines to avoid these pitfalls.

State of the art The TA co-leader Bernd Bischl initiated the R package mlr [Bis+16] and is a co-author of mlr3 [Lan+19]. These packages offer a unified interface to dozens of ML algorithms, making the algorithms easily applicable to dataset libraries. In combination with parallelization frameworks, as offered by the batchtools package [LBS17], the software tools provide an ideal foundation to perform large-scale benchmark experiments on high-performance computing systems. Several general guidelines for the design [Hot+05; Web+19] and statistical analysis [Dem06] of the results of simple benchmarks exist. However, there is a lack of guidelines for more complex benchmarks, e.g., related to complex ML pipelines, which involve many hyperparameters of mixed and nested type and, thus, require advanced methods from design of experiments. Existing tools to simplify the analysis of benchmark results are either outdated [Eug] or specialized to subfields [Sae+19]. Hence, there is a need for supporting researchers in the process of benchmarking through the creation of tools for (interactive) analysis of results and suitable guidelines (e.g., on sequential adaptive benchmarking design for efficient use of computational resources).

Target user group This measure aims at ML researchers who want to conduct benchmark experiments and publish their results in machine-readable form. It also addresses researchers interested



in analyzing benchmark results to gain new insights and learn from previous experiments.

Added value Providing standardized guidelines and convenient tools that support researchers in the process of benchmarking will facilitate the design and analysis of complex benchmark results. Producing machine-readable results of benchmark experiments will facilitate collaborative research and foster interdisciplinary collaborations.

Implementation We will perform large-scale benchmark experiments focussing on the dataset libraries compiled in Measure **M3.1**. We will extend the existing OpenML infrastructure to publish and access benchmark results of individual experiments along with further meta-information and a PID. Additional tools to facilitate the analysis of benchmark results will be developed. We will compile a list of guidelines and best practices for the design and analysis of more complex benchmark experiments.

Deliverables

D-TA3-C1 A collection of large-scale benchmark results based on the library of datasets from Measure **M3.1**, which will be machine-readable and easily accessible on OpenML.

D-TA3-C2 A software tool that facilitates and supports researchers in the process of benchmarking, e.g., R packages focussing on the analysis of benchmark results.

D-TA3-C3 A list of guidelines and best practices for the design and analysis of more complex benchmark experiments.

Milestones

M-TA3-C1 Consistent interface for experiment results of different tasks Create a consistent technical infrastructure and interface on OpenML that allows publishing experimental results from various data-analytic tasks. This is needed as each task usually includes different meta-information and produces different types of results.

M-TA3-C2 Perform experiments for supervised learning Feed the OpenML platform with exhaustive experiment results performed on the library of datasets from Measure **M3.1**.

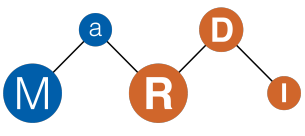
M-TA3-C3 Software tool Develop software for easier (descriptive) analysis of benchmark results.

M-TA3-C4 Compile a list of guidelines Completion of a list of best practices and key guidelines for analyzing and designing complex benchmark experiments.

Measure 3.4: Standards for Peer Review of Numerical Experimentation

Synopsis The main mode of documentation in the development of statistical/ML methods is the publication of peer-reviewed scientific articles. A key aspect of such articles are numerical experiments that explore the behavior of proposed methods. It is thus of crucial importance that such experiments are subjected to a rigorous peer review just like the textual description of methods and associated mathematical results.

In this measure, we will work with a journal partner to establish standards for peer-review of numerical experiments as well as the software code with which the experiments are conducted. Specifically, we will create a “badge system” that indicates different levels of (long-term) reproducibility of numerical experiments and the validity of software, and also highlights specific obstacles to full reproducibility



(e.g., confidentiality of medical patient data, or high demands in computation time). Our work will be guided by existing initial efforts by partnering journals as well as new initiatives such as the Open Science Framework (OSF), a cooperative effort of North American universities.

State of the art The publishers of journals and conference proceedings in Statistics and ML are largely aware of the challenges arising in peer-review of papers presenting extensive numerical experiments. Indeed, some journals have already implemented steps towards more rigorous review processes of software and data. Two examples are the *Biometrical Journal* (a regional journal of the International Biometric Society) and the *Journal of Statistical Software* (an open-access journal). The Biometrical Journal has established Reproducible Research (RR) editors that check submitted data and code after the article has been accepted in the regular review process [HSE16]. Authors are highly encouraged to submit code and data, but unless a new software is promoted, submission of code/data is not mandatory. The *Journal of Statistical Software* requires referees to review both the manuscript as well as the software, which has to work as indicated and needs to be well documented.

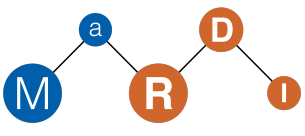
At a general infrastructure level, the Open Science Framework (OSF) of the U.S.A.-based Center for Open Science promotes openness and reproducibility of data and code. OSF works with a badge system for scientific articles that indicates transparency of methods and open access to data [Blo+20]. Despite their voluntary nature, prior research indicates the effectiveness of badges [Kid+16].

Added value This measure lays the groundwork for a next generation peer review process that enables reviewers to understand how numerical experiments were conducted and how output included in a paper was produced. It enhances the findability and accessibility from the reviewers' point of view. The badges generated in the developed peer review system provide readers immediate information on the quality of supplied code, documentation and data, independently of the paper.

Implementation Building on the RR initiative of the Biometrical Journal (BJ), the standards of the Journal of Statistical Software (JSS) and of the OSF, as well as our Measure **M3.3**, this measure will develop a peer review system that uses badges to indicate the transparency and openness of methods/software and data, and the correctness and reproducibility of numerical experiments. To this end, the task area will liaise with the BJ (see the journal's letter of support). In cooperation with the journal's RR editor Fabian Scheipl, we will select manuscripts that will be considered for pilot validation through badges. Upon author consent, a MaRDI PostDoc will serve as a technical editor that takes over the RR review process. For all badges, items must be made available on an open-access repository with a persistent identifier in a format that is time-stamped, immutable, and permanent, like OpenML or OSF.

In addition to the requirements of the BJ the technical editor checks MaRDI standards and adds badges indicating the achieved level of openness and reproducibility, while accounting for limitations such as confidentiality. In parallel, the MaRDI PostDoc will accompany the review process of selected submissions to JSS and work on adoption of the badge system in its software-centric setting of JSS.

In a second step the PostDoc will develop guidelines for referees. To further support reviewers, a technical infrastructure will be provided for checking submitted code. As the majority of code submitted to the BJ is written in R, we will focus on verification of R code, counteracting compatibility issues that



may be caused by future evolution of the R package system.

Deliverables

D-TA3-D1 Publications that follow the MaRDI standard and illustrate the use of the badge system

D-TA3-D2 Guidelines for reviewers

D-TA3-D3 Scientific paper describing design of the system, experiences from an initial time period and potential to extend the approach to other journals.

Milestones

M-TA3-D1 Review process following MaRDI standards 10 review processes each for the Biometrical Journal and the Journal of Statistical Software from the perspective of the guidelines from Measure M3.3.

M-TA3-D2 Guidelines for reviewers Creation of guidelines for reviewers.

M-TA3-D3 Peer-review process following MaRDI standards First peer-review processes at the Biometrical Journal completed with referees following the specified guidelines.

Services

Library of datasets Guided by two selected use cases, a library of datasets and associated meta-information will be established on the OpenML platform. An infrastructure will be provided to expand this library and create further customized dataset libraries (automation, standards).

Library of statistical analyses Again guided by the two selected use cases, a library of notebooks will be initiated that link data, methods, software and literature. An infrastructure will be established to expand this library alongside extensions to the data library from Measure M3.1. A service will be provided to access task-relevant datasets hosted on the OpenML platform and to analyze these data using routines implemented in R.

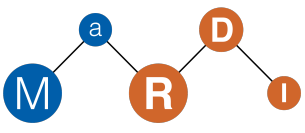
Benchmark results for real-word data Comprehensive benchmark results for the library of datasets from Measure M3.1 will be established. These results will be machine-readable and easily accessible on OpenML through a consistent interface that allows publishing experiment results from various data-analytic tasks. A platform will be set up, that collects and links all statistical analyses (including those from M3.2) for the performed experiments, including a comprehensive description of the experiments.

OpenML workshops The workflow for extending the libraries of datasets and statistical analysis as well as the benchmark results needs to be communicated. Besides proper online documentation, this will be done through workshops organized in the OpenML workshops series. As also described in T6, we seek to attract a diverse set of participants to these workshops.

Peer review publications following MaRDI standard A badge system for peer-review of numerical experiments based on MaRDI standards. Publications that illustrate the use of the badge system, and guidelines for reviewers.

Embedding into NFDI

Measure M3.1 on curated datasets immediately contributes to objective O1. Through the data analyses conducted in Measure M3.2 a semantic relation between data and statistical methods is estab-



lished, thus, addressing objective **O6**. Objectives **O2** and **O3** are targeted by Measure **M3.3**, which establishes workflows and ensures reproducibility of numerical experiments. Concerned with peer review, Measure **M3.4** implements objective **O5** in the context of Statistics and Machine Learning. Finally, the task area's measures contribute to **O7** through material for training courses and workshops.

Integration into the NFDI is realized via **T4**. In particular, Case Study 6 and Case Study 8 are related to NFDI-Neuro with respect to applications to Neuroscience, BERD@NFDI and PUNCH4NFDI in connection with methods from machine learning.

MaRDI Expertise

This MaRDI task area is co-led by Professors Bernd Bischl and Mathias Drton. Bernd Bischl brings valuable expertise from the OpenML project, which builds open source tools and workflows to discover, share and analyze data from diverse domains [Van+13]. He has also worked on connecting the OpenML project to the R project for Statistical Computing [Cas+17]. Mathias Drton and his PhD students have contributed a number of packages to the R project, including, e.g., [MDS20; FDW19]. WIAS has a longterm experience with R, neuroimaging, statistical analysis, and reproducible research, see [PT19].

Risks and Mitigation

The task area faces two primary risks:

- underestimated workload for implementation of one or more of the planned measures,
- poor adoption of the services created in the task area by the Statistics and ML community.

Aware of the risk of underestimating workload, the measures in the task area are designed in the form of pilots and demonstration examples rather than exhaustive databases. This permits a risk mitigation strategy that prioritizes completion of a more modest number of pilot and demonstration examples over immediate broad coverage of many topics. Furthermore, the focus on two complementary use cases leaves the flexibility to prioritize one of the use cases in any particular measure.

Our mitigation strategy regarding community adoption is to work with guiding examples and positive-natured incentives (e.g., badge system in the measure on peer review) as opposed to imposing requirements on researchers. Furthermore, outreach efforts will be coordinated through MaRDI **T6**.

T4: Cooperation with Other Disciplines

General description of the task area

Mathematics is undoubtedly at the core of many, if not most, scientific disciplines. Although each scientific discipline has its own research questions, structurally these questions can be very similar. Mathematical formalism makes the structural similarity visible and thus leads to a transferability of solution methods, concepts and ideas. As a simple, guiding example, the equation $\partial_t u - \Delta u = f$ with the Laplace operator Δ, is used to describe the momentum equation in fluid dynamics if u denotes pressure; or heat conduction if u denotes temperature; or a special case of the Maxwell equations in electrostatics; and many more. The mathematical solution theory of this family of equations leads to so-called harmonic functions, which in turn can be found in other disciplines, such as the description of sound waves. The goal of **T4** is to create a bi-directional bridge between mathematical developments in **T1**, **T2**, **T3**, and other disciplines. This is achieved mainly along the identified cross-cutting topics 'workflows', 'algorithms' and 'models', which consequently form the key data: Models, methods, algorithms, and their implementation in findable software, are closely intertwined and connected. On top of these obvious bridges towards other disciplines, there is still an enormous wealth of opportunities to be tapped with disciplines that are not directly linked to mathematics in the common understanding. The envisioned *MaRDI Platform for Interdisciplinary Exchange*, cf. Measure **M4.3**, also allows mathematicians to find novel application domains, test cases, and benchmark problems for their research.

Following the NFDI vision to develop a sustainable research data management strategy across disciplines, in **T4** we apply a bottom-up strategy based on Case Studies. This is initially done in a prototypical fashion and in a bilateral way between mathematics and other disciplines, realized through strong ties to partners in other NFDI consortia which partially already started to develop own databases, ontologies and services. The overarching conceptional challenge is to bridge different 'vocabularies' and to identify similar mathematical concepts and schemes, which directly implies that all four layers from **X1: Core** to **X4: Knowledge** need to be harmonized and interlinked across disciplines. It is thus of utmost importance that disciplinary research data stemming from other fields are linked in a stronger, more systematic way to the underlying mathematical concepts and corresponding realizations, that are developed within MaRDI.

As an example, a typical workflow in other disciplines that uses mathematical concepts in order to solve a certain problem can be described, along with the arising typical questions, as in Figure 7: A real world problem is simplified to e.g. experiments/surveys, and then described by equations/inequalities subject to initial/boundary conditions, together called the *model*. The model combined with concrete input data forms an *instance* of the problem. An algorithm/method/solution, exactly or approximately, transforms input data (the instance) into output data (the solution). The solution method must be validated and, finally, the solution itself must be interpreted with respect to its original context.

The selection of the disciplines and the choice of the Case Studies are based on the importance of mathematical models, methods and algorithms in each discipline. The chosen disciplines exhibit strong institutional and bidirectional links to other NFDI consortia, which constitutes a base condition for the success of **T4**. This selection is certainly not fixed for the entire funding period: Conversely,

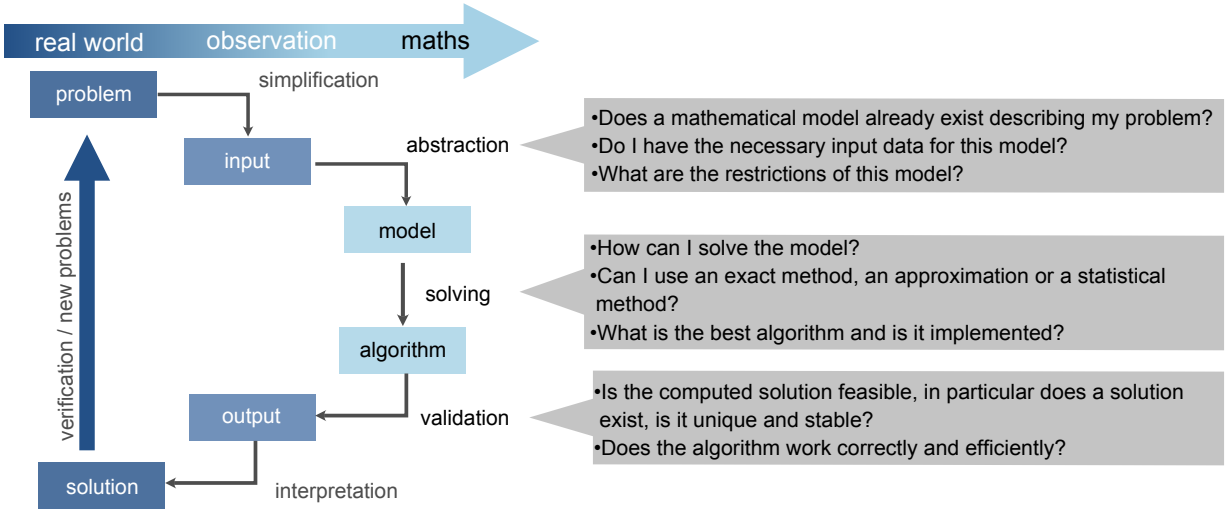


Figure 7: Typical modeling-simulation-optimization workflow: from real world problem to model.

we plan to take advantage of four partner consortia being ahead of us in terms of funding, using them as prototypes to add mathematical research data to research data handling that is already being established. The outcome of these success stories will determine the interaction with the other partners. Finally, we plan to allocate dedicated flexible funds towards the end of the funding period to include other disciplines, see **M4.4** for details and Section 3.4 for the organizational structure and budget handling.

Case Study 1: Engineering, Material Science and High Performance Computing (with NFDI4Ing, NFDI-MatWerk and NFDIxCS)

Modern industries (like automotive, chemical, biotechnical, and health industries) maximize their profit by making use of solving multi-scale and multi-physics problems. These types of problems are often interdisciplinary by construction and in addition, they not only exhibit, but also involve pressing challenges within modern mathematics, e.g., in applied analysis and scientific computing. Multi-scale and multi-physics problems are diverse: The engineering of synthetic microstructures, ceramics, polymers and bio-genic materials make use of data-driven and data-calibrated simulations of microstructured materials. Moreover, pressing societal challenges call for porous media groundwater and climate modeling. Equally pressing are personalized rehabilitation/medicine and a foundation research understanding of activation processes of movement and body control.

All these topics are highly diverse in terms of workflows (**X3: Exchange**, e.g., modeling-simulation-optimization, cf. **M2.4**), models, methods, algorithms and implementations. They include Direct Numerical Simulation, a wide range of competing numerical methods like particle methods, or grid-based discretizations; contain data-driven closure relations realized, e.g., via machine learning (cf. **T3**), phenomenological relations, stochasticity and uncertainty, and often ad-hoc ‘rules of thumb’ and other heuristics. Some of the methods require vast computing resources. As the target applications are societally challenging, no ‘gold standard’ in terms of mathematical approaches or benchmark references has been determined yet, making standardization and enabling comparability even more essential.

In **T4** we establish the link to an RDM strategy for High Performance Computing jointly with NFDIxCS

through the fluid dynamics codes SU2[18], Flexi[19] and DuMuX[20]. Flexi is, for instance, frequently allocating approximately 10 % of the annual resources at the High Performance Computing Center in Stuttgart, a member of the Gauss Initiative together with München and Jülich. Through this cooperation, we expand the collected and analyzed metadata by actual runtime, scalability and energy efficiency observations, in addition to the more mathematical metadata like convergence behavior. Thus, this Case Study emphasizes the importance of mathematics not only across disciplines, but also in the sense of cross-cutting topics across several NFDI.

By teaming up the expertise of the ITWM (Kaiserslautern), the Clusters of Excellence 'SimTech' (Stuttgart) and 'Mathematics Münster' and the MPDC in Magdeburg, a very broad impact in terms of Scientific Computing applications (**T2**) is expected.

Case Study 2: Theoretical and Bio-Chemistry (with NFDI4Chem)

Analytical chemistry constantly generates experimental data (in large scale facilities of top-level research, e.g. BESSY II). In most cases an analysis of the data is carried out on the basis of traditional physical models, e.g. of quantum theory. The advent of machine learning methods, which is feasible through the use of large databases, can only come about through a jointly created and shaped data platform between the specialist disciplines. We will cooperate to make the data from NFDI4Chem available accordingly. However, also less noticed mathematical advances (non-negative matrix factorization, Markov State models) could lead to a new conception of natural science ideas. MaRDI uses the description and structure of the chemical data to make these mathematical methods findable for other disciplines. We build on existing test cases and data sets that have been developed in the interdisciplinary cluster of excellence SALSA.

Another cooperation with NFDI4Chem, realized through the SimTech Cluster of Excellence, targets technical biochemistry, through a focus on experimental and computational enzyme catalysis, which is relevant, for instance, in drug design problems. Here, RDM is already well-established, and the challenge is to interlink existing databases of molecules and enzymes, and existing simulation software, with the Database of Numerical Algorithms from **M2.1**.

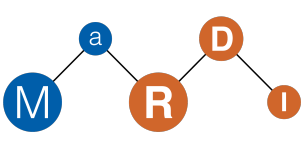
Case Study 3: Digital Humanities (with NFDI4Culture)

Sometimes societal challenges are "solved" with mathematical models. Such an example has been shown in the 'avoidance of animal experiments' [Vec+19; Spa+17]. The strongest link between cultural studies and mathematics is often to enable and interpret statistical analyses of digitalized data (a link to **T3**). However, our interdisciplinary exchange in MaRDI will also be dedicated to the 'application of non-statistical mathematical models in cultural studies'. NFDI4Culture traces social changes by changes in their cultural assets, which is the link to the goals of the project area EF5 in the Excellence Cluster MATH+ providing models to MaRDI.

Case Study 4: Digital Catalysis (with NFDI4Cat)

Catalysis and chemical engineering are fields of high strategic importance for solving pressing challenges concerning climate change and fight against air, water and soil pollution as well as supply of

[18] https://su2code.github.io, lead developers from ITWM, Kaiserslautern
[19] https://www.flexi-project.org, lead developers from Stuttgart
[20] https://dumux.org, lead developers from Stuttgart



sustainable energy, materials and chemicals production at the same time. Both are highly interdisciplinary in nature, and digitalization of catalysis ('Digital Catalysis') is in progress on an industrial scale. NFDI4Cat supports this development by providing services and tools for research data management in digital catalysis following the FAIR principles. This includes the development of metadata standards and ontologies for data produced and employed in digital catalysis. As one concrete example of global pressing issues, NFDI4Cat focuses on the reduction or complete avoidance of CO_2 emissions in industrial processes.

MaRDI co-spokesperson Peter Benner (MPI DCTS) serves as co-spokesperson in NFDI4Cat, too, and this joint case study of both consortia is built on his contributions to NFDI4Cat. This involves in particular the multiscale modeling and simulation as well as the control and optimization of Power2Gas and Power2X plants as typical catalytic processes necessary for supporting the energy transition in Germany. Here, CO_2 is employed for methanation and other processes to produce basic chemicals. A typical example in this context is the optimization and control of a methanation reactor [Ben+17]. Here, we have a full workflow from modeling to simulation to optimization and control of the process. Therefore, we will use this to describe how a confirmable workflow in numerical experiments involving several software components and data sources can be implemented. The standards that will be developed for describing the computational pipeline and data are exemplified with this case study. In further steps, this prototypical best practice example will then be tested and employed for more advanced computational experiments related to Power2Chemicals processes. Metadata and ontologies used in these workflows will be developed in cooperation of both consortia and with the FAIRmat consortium that also overlaps with both initiatives in particular in these aspects.

Case Study 5: Optimization and Computer Algebra in Particle Physics and Heavy Ion Research (with PUNCH4NFDI)

In heavy ion research large scale non-smooth optimization problems have to be solved and classical gradient based algorithms cannot be applied. Examples are studies of the quark-mass dependence of hadron masses in quantum chromodynamics or amplitude analyses in hadron physics. While a scalable implementation of an evolutionary algorithm is already available within the GENEVA software framework, it is desirable to develop and benchmark further scalable non-smooth optimization algorithms for applications on large scale clusters. In general, within **M2.4**, PUNCH4NFDI and MaRDI will set up workflows for algorithm development and evaluation. Such workflow exemplifies the usage of infrastructure developed in **M2.1**, **M2.2** and **M2.3** to advance heavy ion research.

Within elementary particle theory, computer algebra methods are intensely used to perform large-scale analytic calculations. Examples of relevant algorithms concern anti-differentiation in the analytic calculation of Feynman integrals by using special higher transcendental functions, Mellin-Barnes techniques, the analytic solution of large systems of ordinary differential and difference equations, multi-summation in rings and fields, Almqvist-Zeilberger algorithms, relations in associated special function spaces (sums and integrals) and of the related special numbers. MaRDI will provide FAIR data formats (see **M1.2**) and guidelines for validated computational workflows (see **M1.1**) to improve the re-useability within the scientific community.

MaRDI and PUNCH4NFDI share a common interest in linking publications, models, computational

methods, relevant software, and the calculated data (e.g., mathematical expressions generated up-to the order of tera-terms within PUNCH4NFDI). PUNCH4NFDI will contribute the domain-specific knowledge and the original content to the database, and MaRDI will integrate this information into its knowledge graph (see **M2.1**). Information from the HEP-INSPIRE[21] community hub and the HEP-forge[22] development environment shall be used.

Case Study 6: Neurosciences (with NFDI-Neuro)

Neurosciences produce a variety of heterogenous data. Representative examples are activity descriptions from electro- and optophysiology or EEG, MEG, EMG, as well as (f)MRT scans. In particular in the field of computational neurosciences the simulation, analysis and validation of workflows are of major interest. In this Case Study we focus on imaging processing problems, e.g., in magnetic resonance imaging for the detection of temporal signals or structural changes [HHP18; Tab+19]. Further, we aim at signal processing and machine learning, e.g., in neuroscience applying EEG signals for determining biomarkers for classification; in physiological resilience observations of Internet Gaming Disorders; or clinical outcome data like death, recovery, achieving special chronically stable state.

Another field is mathematical modeling, e.g., by systems of ordinary differential equation or neural networks, for the purpose of neuron and brain simulation. This requires the description of the model, its execution and parameter estimation from data, similar to SBML in system biology. Such a specification is a prime example for mathematical research data. A systematics of those model systems is missing, such that two identical or similar models in two publications may differ significantly in their mathematical nomenclature, their graphical representation of the network architecture, or their parameterization. MarDI and NFDI-Neuro will cooperate on the development of executable FAIR specifications of neuron and brain models. MaRDI can contribute descriptions of the mathematical properties of the model systems and NFDI-Neuro will contribute domain-specific knowledge. Both can be integrated into the MaRDI knowledge graph.

Case Study 7: Decision Science (with NFDI4MobilTech)

Decision Science deals with the evaluation and optimization of decisions in the presence of conflicting goals stemming from different interest groups. Such decisions have a huge variety of applications, e.g., the optimization of production processes, logistics, design of supply chains, applications in health, environmental and social sciences. Often, methods of Operations Research are used, like combinatorial optimization, integer programming or continuous optimization. We concentrate on integer programming and use optimization of public transport as initial prototype. This includes planning the basic network with its stops, stations and direct connections, designing lines, computing a timetable and dealing with operational questions such as vehicle- and crew scheduling. All these steps require specific input data, and all of these steps can be modeled as decision problems, where goals can be to save costs, to offer passengers good connections with few transfers and low traveling times, or to provide a system which is robust against delays. This is often unknown to researchers in traffic engineering, and researchers in optimization are not aware of the requirements and problems in public transport planning. Concretely, NFDI4MobilTech and we will jointly specify an interface to

[21] https://inspirehep.net
[22] https://www.hepforge.org/

provide traffic and mobility data for use in mathematical models. The data formats of the interface will cover all modes (public and private transport) with their supply data (road and rail networks,traffic control data, timetables etc.) and their demand data (traffic and passenger counts, origin-destination-matrices, etc.). The interface will provide functionality to search specific meta data in our repositories according to classifications and ontologies. Therefore, this interface between NFDI4MobilTech and MaRDI will be a good prototype how to provide appropriate interfaces to other domain communities.

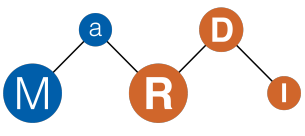
Case Study 8: Data-driven Mathematics, AI and Machine Learning (with NFDI4DataScience, PUNCH4NFDI and BERD4NFDI)

Modern data-driven methods, Artificial Intelligence and Machine Learning are crucial in several scientific fields and in various application domains. This case study serves as demonstrator for the application and use of MaRDI's **T3** and the corresponding measures by linking and bridging to different fields and consortia. For instance, methods from Statistics and increasingly also Machine Learning play an important role for research in data-intensive fields such as particle physics. Topics of interest are statistical analysis beyond Gaussian uncertainties, e.g., detector resolution, application of machine learning and neural networks to particle tracking, particle detection and also for permanent detector monitoring. PUNCH4NFDI and MaRDI **T3** will cooperate to extend the data library from **M3.1** and the library of statistical analysis from **M3.2** with examples from particle physics, and to use the resources created in these measures as a basis for statistical training of physicists.

Moreover, MaRDI and the consortium BERD4NFDI will cooperate on interdisciplinary topics in machine learning led by Bernd Bischl, who is co-spokesperson in both consortia. Both consortia aim at advancing machine learning approaches and at enabling researchers to successfully apply them to specific research questions. While MaRDI contributes its expertise in algorithms, their implementation and empirical benchmarks, BERD4NFDI focuses on the application of these tools to data and research questions in economics and social sciences, and provides data suites, use cases and algorithmic challenges for MaRDI. A further field of collaboration is the assessment of data quality in heterogeneous and unstructured data.

MaRDI also closely collaborates with NFDI4DataScience, through a jointly defined bridge between the consortia: On the one hand, NFDI4DataScience will contribute example data sets and machine learning tasks from language technology, biomedical sciences, information sciences and social sciences to the libraries planned in MaRDI's **T3**. Additionally, for the use and implementation of benchmarking platforms, cooperation and a continuous exchange concerning realization and possible interfaces are planned. On the other hand, NFDI4DataScience can profit from MaRDI's expertise on machine learning pipelines for practical applications. Furthermore, MaRDI will contribute its micro standards for metadata to describe characteristics of algorithms and benchmark data to the joint development of metadata commons for Data Science. Via this bridge, aspects of both mathematics and computer science in the field of data science and machine learning can be efficiently combined and complement one another in a fruitful and synergetic way.

Summarizing the Case Studies, we have defined four measures, expanding and extending on all other Task Areas:



- Measure 4.1: Documentation and Analysis of Interdisciplinary Workflows
- Measure 4.2: Standardization of Mathematical Descriptions across Disciplines
- Measure 4.3: MaRDI Platform for Interdisciplinary Exchange
- Measure 4.4: Transfer beyond Case Studies

We emphasize that in contrast to **T1**, **T2**, **T3** so far, our measures are all designed (and required) to span all layers:

– **Layers: X1: Core**, **X2: Data**, **X3: Exchange**, and **X4: Knowledge**

To emphasize the interdisciplinary aspects of mathematics and the goals of the NFDI, certain deliverables are also included in the partner initiatives' proposals. However, due to space constraints, we do not break down the following detailed description of the measures to the level of individual cooperations between all involved partners, in particular in terms of Milestones and Deliverables. Instead, we deliberately choose to illustrate our ideas and approaches with explicit and representative examples from the range of the Case Studies, building on the ideas of all other Task Areas.

The guiding scheme for the first three measures is as follows: For each Case Study, we proceed first in a top-down manner, starting with the analysis of the disciplinary workflows. Once these workflows are understood, and preliminarily described in a way that is understandable by both mathematicians and scientists from the partner disciplines, we can proceed in a bottom-up approach along the four layers. Finally, in the last measure and during the second half of the funding period, we open the initial limitation to the selected Case Studies and partner consortia. Budgetary, this is realized by a 'flexible funds' component, based primarily on bidirectional projects between partner consortia as described in Section 3.5.

Measure 4.1: Documentation and Analysis of Interdisciplinary Workflows

Synopsis and relevance Based on the Case Studies, we initially survey, analyze, and categorize the use of mathematical concepts and workflows in the partner consortia, and their respective state of FAIRness, based on typical workflows like modeling-simulation-optimization. The goal is to extract existing commonalities in terms of workflows, models, algorithms and implementations; along with their appropriate metadata; and thus to set the stage for the following measures through a thorough state-of-the-art analysis of their heterogeneous use.

Detailed Description The Case Studies with NFDI4Ing, NFDI-MatWerk, NFDIxCS, NFDI4Cat and PUNCH4NFDI (and to some extent also NFDI4Chem) are primarily oriented towards *'classical' models based on differential equations* and their numerical approximation. Hence, the abstract workflow developed in **M2.4** can (initially) be directly instantiated. We use the above-mentioned Poisson problem $-\Delta u = f$ to describe our approach to identify commonalities, but emphasize that the same reasoning holds for more complex and/or hierarchical models (e.g., multiphase flow with NFDI4Ing, or microstructured materials with NFDI-MatWerk), see also the *Database of Mathematical Models* below in **M4.2**. Such models are already included or planned to be included in the metadata model descriptions of the partner initiatives. We will thus identify different 'physical' meanings of the variable u, e.g., fluid pressure, electric current, …, in preparation for the **X4: Knowledge** layer in **M4.2**.

Furthermore, we will identify different formats for the data f (layers **X1: Core** and **X2: Data**), ranging from measurements (point clouds) via machine-learned closure relations (evaluable formulas) to fully dynamic fully data-driven realizations as a software component. In terms of algorithms (the realization of the Δ operator, more generally the **X3: Exchange** layer), we can initially focus on selected numerical schemes that are implemented in the codes of the MaRDI partners and the codes of the external partners, e.g., the BETTY archetype in NFDI4Ing. This will lead to the identification of lower-level commonalities: For example, the mathematical object 'linear system of equations' (cf. the guiding example on page 29) appears after discretization of the model. Certain discretization schemes rely on the rigorous mathematical analysis of polynomials, touching algebraic aspects of **T1**. In terms of the layers **X1: Core** and **X2: Data**, a multitude of different data and (meta)data formats have to be collected, analyzed and classified. For instance, established matrix storage formats that still lack interoperability in software (**X1: Core**, conversion routines) are addressed on the basis of Case Studies, but also the accessibility of discipline-specific datasets (**X2: Data**, integration of data repositories, like different maturity of curated imaging experiments in porous media or turbulent flow). These two layers also include models, methods (numerical schemes) and algorithms (their implementation) across the disciplines. Furthermore, the workflows themselves are subject of our analysis (**X4: Knowledge**).

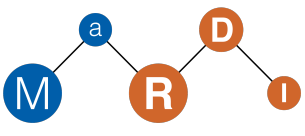
Other Case Studies, e.g., those related to NFDI4Culture, NFDI-Neuro, BERD@NFDI, and NFDI4DataScience, are oriented more towards *data-based modeling aspects*, as the underlying data are in most cases empirical. The data can often be categorized as descriptive or structured with respect to space and/or time. Typically, the workflow in these domains is to statistically analyze the data, to recognize patterns, to derive a model explaining the experimental findings, and to iteratively add missing parameters or statistics, e.g., by experimental design, linking directly to **T3**. As illustrative example: After inference of the Markov chain parameters of sequential Raman spectroscopy data, tipping points and broken ergodicity can be analyzed. From that analysis further experiments can be proposed. Time series are also used in decision sciences, business, medicine, and archaeology, so this approach can be transferred to a broad range of applications in these disciplines. In the layers **X1: Core** to **X3: Exchange**, we learn from our cooperation partners their standardization ansatz. Then, we collect the meta data from our cooperation partners and implement software tools that automatically convert the descriptive or unstructured data into standard data formats. The goal is to map as many as possible experimental findings to mathematical workflow standards, e.g., by introducing a standard data format describing time series data.

The above approach leads to the following **Deliverables and Milestones**, note that they are partially budgeted in **T1**, **T2** and **T3** for some of the Case Studies:

M-TA4-A1 Profound analysis and documentation in collaboration with NFDI4Ing, NFDI4Chem, NFDI4Culture and NFDI4Cat (first 6 months)

M-TA4-A2 Profound analysis and documentation for the other Case Studies and partner consortia (first year)

M-TA4-A3 State-of-the-art analysis beyond Case Studies (rolling)



Different NFDI consortia start at different times. For our Deliverables this is a strong benefit: If the disciplinary aspect (e.g., the comprising the model, the experimental data, the instance of an algorithm) has already been sketched, and if the mathematics has also matured descriptions, then adding mathematical abstractions, metadata and ontologies (**M4.2**) is substantially more realistic to accomplish. We are aware however that this is a timely issue, since standardizing in an interdisciplinary way is challenging if a disciplinary terminology and vocabulary has already been fully established. Because of this dependency we start **M4.2** immediately after **M4.1** and measure their common success.

Measure 4.2: Standardization of Mathematical Descriptions across Disciplines

Synopsis and relevance After analyzing and collecting the state of the art in **M4.1**, we first design a model database complementing the other databases developed in **T1**, **T2**, **T3**. Then, we interlink these databases, and the databases, ontologies and metadata schemes developed in the partner disciplines, through a hierarchical definition of micro-schemata and micro-ontologies. This leads to an intermediate interdisciplinary 'spiderweb' of interlinkable, searchable and findable data, metadata and model descriptions spanning **X1: Core** to **X4: Knowledge**, and we carefully ensure that all efforts are compatible with the envisioned (interdisciplinary) knowledge graph architecture of **T5** and **M4.3**.

Detailed Description Standardizing disciplinary data and metadata is already a tremendous challenge, and the reason why the NFDI exists. Harmonizing data across disciplines adds a combinatorial explosion, as, e.g., identified by the European Commission's Directorate-General for Research European Open Science Cloud (EOSC) strategic implementation plan (see Section 4.2). The vision of being able to combine datasets over disciplinary borders can only be realized if there is a backbone network of metadata, ontology and vocabulary services that is able to connect datasets by interlinking standardized descriptions. In the description of specific research results, models, methods, algorithms and workflows as identified in **M4.1**, **M2.4** and others, there is a trade-off between the necessary flexibility and specificity of the description on the one hand and the interoperability of this description by standardization and machine readability on the other hand. A possible solution – as pursued in the metadata concept of some of the already running NFDI consortia and suggested by the EOSC recommendations – consists of the individual composition of standardized micro-schemata that are interlinked to each other. To enable specificity and interoperability, these schemata can be defined in a hierarchical way, by inheriting properties of the parent concepts and refining them in child concepts. In this way, different concepts can be related via their common ancestor while entailing all necessary specific information, see Figure 8.

In order to build a database for mathematical models corresponding metadata schemata are essential. For instance, the model $-\Delta u = f$ may appear as an 'Equation Type' (submodel, **X1: Core** in terms of vocabulary and semantics on the (micro-) ontology level and in the *Model Database*) in a complex fluid dynamics simulation, and a software realization of a multigrid method tailored to such elliptic models (**X3: Exchange**, see **M2.1** and **M2.4**) may appear in some electrostatics problem, and the core relation (in terms of applicability of the method to the problem) between the two must be represented properly, across disciplines (**X4: Knowledge**). The same holds, e.g. for spectroscopic data in Computational Chemistry, which must be interlinked with matrix factorization techniques. These

examples highlight the challenge of combining and unifying metadata descriptions, vocabularies and ontologies.

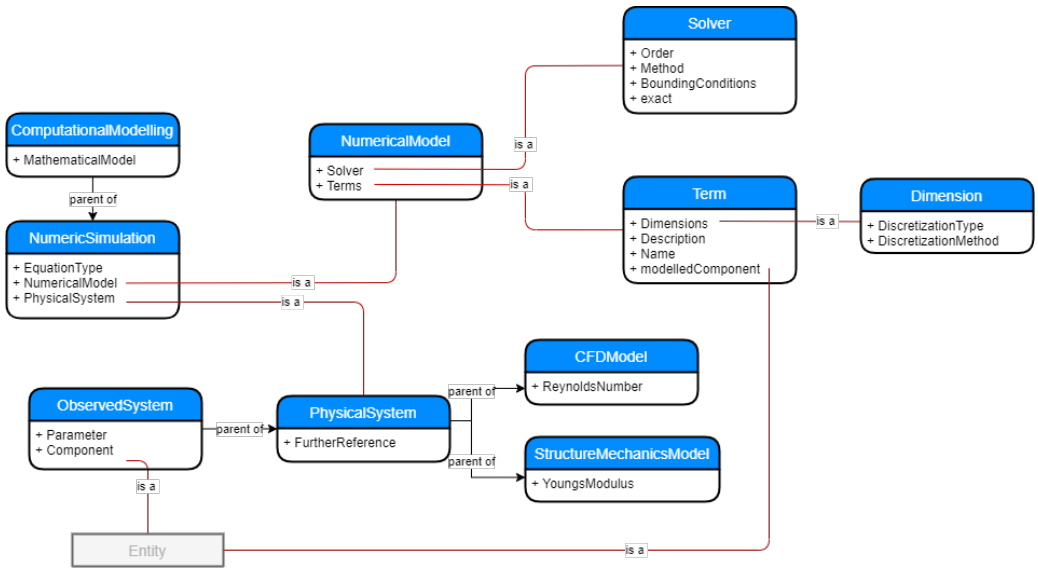


Figure 8: Example of interrelated concepts for the description of a Fluid Dynamics simulation.

In consideration of the wide variety of already existing schemes and standards, and in terms of backward compatibility, an all-encompassing all-agreed-upon ontology seems too ambitious and unrealistic. Instead, we follow the EOSC recommendation to interlink different standards by mappings or interrelating ontologies, in a hierarchical way. As outlined in Section 4.2, we aim at using the Wikibase and Wikidata technologies as much as possible to avoid dead-ends and to prepare for interoperable (sub-)knowledge graphs.

Beyond their metadata description, mathematical models have to be considered as entities of their own class, which can be uniquely identified, cited, and categorized, rather than found as plain text with a mixture of mathematical notation and common language [KT16]. Those vague references potentially lead to ambiguity, cites different original work, incompleteness, and “re-invention of the wheel”. The recognition of mathematical models as part of mathematical research data will be established by creating a semantic digital corpus of mathematical models. For the description of the mathematical models itself, we aim at a modeling-oriented lightweight markup. In contrast to a pure plain-text description, the markup generates machine-interpretable relations between the formal description entities and express their mathematical and domain-specific semantics.

Deliverables and Milestones To achieve this goal, we again differentiate between models, data, methods, concepts, and workflows, as outlined in **M4.1**. As models are at the core of defining commonalities between disciplines (e.g., $-\Delta u = f$ occurs in many fields), one deliverable is the design of a *model database*, in which we gather the results, e.g., from **M4.1**, **M2.4**, **M3.3**, and all case studies. For instance, we will abstract heat conductivity, electrostatic potentials, and drug diffusion into one shared object ‘elliptic second-order PDE model’ [KT16], and interlink it with particular instantiations in the partner disciplines.

LaTeX with additional semantic macros can represent mathematical models in a model-oriented

markup language. This notation allows encoding entities containing the model's main characteristics, such as the equation, the domain, boundary conditions, material laws, and the (physical) quantities used in the model.

Based on this model description, LaTeXML generates XML formats and MathML representations optimized for further machine processing. These resources can be used for information retrieval via engines such as MathWebSearch, and interlinked presentations via Web or Wiki pages. Such an approach is, e.g., employed by NIST for the Digital Library of Mathematical Functions (DLMF). Our approach is to develop MaRDI's semantic LaTeX macros in a way that these model representations can be integrated into the MaRDI knowledge graph automatically. The model description will re-use models as building blocks to describe coupled systems within a hierarchy of models. For specific applications like circuit simulation, systems biology, process simulation, and neural networks formalized modeling languages such as SPICE, SBML, OpenModelica, or concepts from machine learning pipelines like TensorFlow or scikit-learn exist. We will integrate these executable model specifications by semantically mappings the involved quantities and mathematical concepts.

Further, we will interlink the model database and the developed micro-schemata with the *Database of Numerical Algorithms* (see **T2**, **M2.1**) comprising concrete software realizations of solution schemes suitable for a specific problem. Third, we will enrich this hierarchical web-of-schemata with purely mathematical results like existence proofs for solutions of certain models instances, or convergence order proofs for numerical schemes. Combining the suggested micro-approach with publications (and their data) is a key success metric for this part of MaRDI, along with a documentation of open research questions.

Finally, we will make actual disciplinary measurements and simulation data available to, e.g., method developers, to be used for validation. This is closely linked to **M2.3**.

The implementation of these ideas for the Case Studies leads to the following

Deliverables

D-TA4-B1 Condensation of micro-schemata, ontologies and vocabularies for the Case Studies (first year), initial designation of unified and unifiable identifiers (second year)

D-TA4-B2 Semantic macros for LaTeX for model descriptions and mappings to MaRDI knowledge graph

D-TA4-B3 Templates for model description in LaTeX

D-TA4-B4 Processing pipeline to human-readable (PDF, HTML5) and machine-actionable model representation in XML and MathML using LaTeXML for integration in MaRDIs Wikibase.

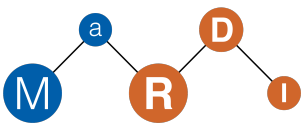
D-TA4-B5 Design of the hierarchical *MaRDI Database of Models*

D-TA4-B6 Interlinking all MaRDI Databases in terms of then-chosen interdisciplinary technology

Milestones

M-TA4-B1 Initial corpus of 100 model descriptions from literature encoded

M-TA4-B2 Semantic macros for physical quantities, physical constants, physical concepts etc. necessary to enrich initial corpus



M-TA4-B3 Semantic macros for metadata vocabularies and ontologies necessary to categorize initial corpus
M-TA4-B4 Semantically enriched initial corpus
M-TA4-B5 Repository for mathematical models integrated with processing pipeline
M-TA4-B6 Report and guidelines on model representation, processing tools and data base
M-TA4-B7 First model encoding workshop with external participants

Measure 4.3: MaRDI Platform for Interdisciplinary Exchange

Synopsis and relevance The goal of this measure is to bridge the gap between the disciplinary approaches in terms of a technically feasible platform. The main challenge is to homogenize the existing and developed micro-schemata, so that both mathematicians and researchers from other disciplines can navigate the MaRDI knowledge graph, e.g., by 'randomly' choosing a starting point in the hierarchy of interconnected models, methods, solvers, software realizations, …, and find suitable components for their problem, by following the interconnecting links either 'upwards' towards abstractions towards a more general view, e.g., model, or 'downwards' towards a more specialized solver implementation, or, and this is the key **X4: Knowledge** layer, 'sideways' to related problem statements, test problems, application domains, ….

Detailed Description The outcome of **M4.1** and **M4.2** is, as seen from a higher level NFDI perspective, already promising, but not accessible to all users from Mathematics or other disciplines. There is a strong risk that research projects, in terms of the NFDI, are stuck in their, discipline-inspired, micro-ontologies. This is rather discouraging, as it prevents truly interdisciplinary connections. Designing, specifying and developing micro-schemata and micro-ontologies (along the entire **X1: Core** to **X3: Exchange**) can only lead to added benefits in terms of generating knowledge transfer (**X4: Knowledge**) if the underlying technology is accessible both by human users (from Mathematics and the partner disciplines) and is machine-readable. In close cooperation with **T5** and the already ongoing efforts in the partner consortia, we have thus decided to build upon the existing WikiBase knowledge graph and the WikiData descriptions as the underlying format, building upon WikiData for research data and metadata. For this vision to be successful, technical mappings have to be devised on top of the semantic mappings developed in **M4.2**, that are compatible with **M5.3**. This is a common challenge faced by all partner consortia: The already started partner consortia approach the problem similarly to us, with various approaches to describe micro-schemata, vocabularies and ontologies and the goal of creating a knowledge graph based on WikiData terminology, as pursued by **T5** and the companion projects in partner consortia. Through a joint effort, the necessary translations between the outcome of **M4.2** will emerge. The role of **T4** is to (co-) moderate this development. Technically, the *MaRDI Portal for Interdisciplinary Exchange* will be a part of the MaRDI portal developed in **T5**.

The above approach leads to the following **Deliverables and Milestones**, in close cooperation with **T5**:

D-TA4-C1 Establishment of the database of mathematical models as a service within the MaRDI portal and integration in MaRDI's WikiBase instance, see **T5**

D-TA4-C2 Mapping of disciplinary ontologies, vocabularies and schemata to WikiData-terminology, and opening such schemata to interdisciplinarity

D-TA4-C3 Terminology service for ontologies and vocabularies for use beyond MaRDI

D-TA4-C4 Integration of MathWebSearch to Database of mathematical models

D-TA4-C5 Integration with publications, software and data sets

D-TA4-C6 Functionality for peer-reviewing of model descriptions by editors for quality control

Measure 4.4: Transfer beyond Case Studies

On the one hand we extend the Case Studies within our partner disciplines. On the other hand we expand our collaborations and the roll-out of the service infrastructure to other disciplines and further newly established NFDI consortia. The impact of the implemented services on the communication between mathematics and other applied disciplines is also evaluated in this measure.

Communication-promoting and also inhibiting processes during the interdisciplinary work are examined in detail, in particular whether or where the use of the MaRDI services, prototypes and the central portal provide possibilities for an improvement of interdisciplinary communication. We will ask our collaborators, and new partners as established during the funding period, to use the MaRDI portal in order to identify possible mathematical models for addressing their own challenges. These experiences are documented and used to improve the MaRDI services.

Storing and handling rather conceptual than relational data is a cross-cutting challenge for MaRDI. We will extend the low-barrier portal as far as possible to include this type of conceptual questions and to support it as a long-term service, paying attention to data and service life cycle questions.

We emphasize that the selection of MaRDI partner consortia, as built upon in **M4.1**, **M4.2** and **M4.3**, should be seen as an initial selection, driving the case-study based approach in **T4**. This final measure is designed to partially keep MaRDI open to new developments. New links may be built between consortia, new consortia will appear in the subsequent founding round, and ideally also some current Case Studies will be finished within this funding period. Hence, **M4.4** is equipped with flexible funds, to enable the flexibility of connecting research data handling in Mathematics with more disciplines and NFDI consortia than initially accounted for. For example MaRDI already agreed on a cooperation with FAIRmat, see **M4.2**. We expect further case studies from the Leibniz-Network "Mathematical Modeling and Simulation" coordinated by WIAS. Many member institutions are also involved in other NFDI consortia and research data initiatives.

Services

T4 provides services on all levels, from **X1: Core** to **X4: Knowledge**: These include micro-schemata, -vovabularies and -ontologies that are harmonized across disciplines, the *Database of Mathematical Models*, and finally the *MaRDI Portal for interdisciplinary Exchange*. All services will be integrated in the MaRDI portal.

Embedding into NFDI

T4 provides the embedding MaRDI into the NFDI as a whole. We initially focus, in a bottom-up strategy, on selected Case Studies to cooperate with other NFDI consortia on integrating mathematical research data, aiming at unifying the description of identical mathematical concepts in different application domains. Vice versa, Mathematics benefits from improved FAIRness in the partner disciplines. We are aware that our Case Studies are only a starting point, and have thus designed **M4.4** to incrementally open MaRDI up to more disciplines and to more partners.

Requirements/own preliminary work (MaRDI Expertise)

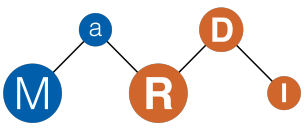
Due to the multitude of different Case Studies, we here only list contributions explicitly that are relevant for the budget justification. More references of the MaRDI consortium members can be found in the description of the Case Studies above.

In Engineering (first Case Study), several larger software frameworks exist, that allow the solution of a wide range of partial differential equation models rather than just a concrete instance. In the research project 'Web of Models' [Gai+11] within the Rhineland-Palatinate research center CM^2 – Center for Mathematical and Computational Modelling (TUK, DFKI, Fraunhofer ITWM) a web-based platform was established with a first approach of a standardization procedure for Navier-Stokes type models in combination with a web-based Python interpreter to visualize the models. The 'Web of Models' addressed the aspects informal description, mathematical description, software support, and metadata. Since 2016, WIAS together FIZ Karlsruhe and Michael Kohlhase (FAU) started a working group establishing mathematical models as research data. Starting with case studies in semiconductor device simulation concepts for the representation of mathematical models that is human-readable but also machine-actionable [KT16; Kop+18; Koh+17] allowing for different levels of formalization have been developed.

Similarly, within the Dipl-Ing BMBF project [IS18] at USTUTT, an initial system for the description of simulation data and processes in engineering has been developed which (semi-)automatically captures structured metadata, stores them in a scheme called EngMeta [IS19; SI19; SI20], and provides an open platform interface for interaction with the metadata and data. With EngMeta, ways to document the whole research process from the perspective of the engineering sciences. To be able to describe mathematical models standardized with metadata, and to transfer these results to mathematics, additional metadata fields and annotations are necessary. EngMeta is currently actively being used in NFDI4Ing.

In physics, chemistry and health and neuroscience, typically only well-established mathematical models and algorithms are applied so far, and existing databases are limited to selected models and statistical data. In contrast to engineering, the introduction of mathematical models in health and neuroscience is a relatively new trend. The neuro-imaging community is exemplary for its developments of open databases like openneuro.org or from the Human Connectome Project and corresponding metadata standards like BIDS [Gor+16]. MaRDI has a profound expertise for neuroimaging problems and neuroscientific modeling [PT19].

In decision sciences, integer programming is a standardized modeling language to describe optimization problems, available in many solvers such as gurobi, CPLEX or mosel. The DFG-funded



research unit FOR 2083 on ‘Integrated public transport optimization’ is an effort to provide a common data structure between mathematicians and researchers in traffic engineering. This includes a standardized way to describe input data and parameters, and a common evaluation of output data [Fri+17; Ptn] as well as a library of algorithms for planning in public transportation [Sch+], all of them being open source.

The use of mathematical models for approaching questions in social sciences is discussed controversially. The transition from conceptual questions to numerical methods is still a giant step. The mathematical assessment of waste water treatment can be seen as an example how to bridge this gap [Web19]. MaRDI is the first consortium that will try to extend mathematical method databases to interconnect with questions from societal challenges. The H2020 project MSO4SC ‘Mathematical Modeling, Simulation and Optimization for Societal Challenges with Scientific Computing’ serves as a starting point, here we have developed mathematical methods to prevent animal testings [Vec+19; Spa+17].

Risks and Mitigation

The main risk is that by construction, **T4** relies on the success of the partner NFDI consortia. To mitigate this, we adopt a bottom-up, case-study based approach starting with partners that already receive funding, i.e., NFDI4Ing, NFDI4Chem, NFDI4Cat and NFDI4Culture. The lessons learned from these cooperations immediately influence the cooperation with the other partners.

One could argue that **T4** has a too narrow focus on just a few other disciplines. However, we are convinced of our unavoidable bottom-up approach. To mitigate the risk, we carefully selected the partner consortia in both depth and breadth of disciplines and existing data culture. We emphasize that this selection only holds for the first three years of the funding period. **M4.2** is explicitly designed to incorporate more partners based on the flexible funding allocation described in Section 3.4.

A major challenge is the decision to (initially) employ WikiBase as the underlying software to model knowledge graphs, and to map data and metadata to WikiData entries. However, as outlined in Section 4.1, the risk that different NFDI consortia use different solutions to model their knowledge graphs is common to merging the efforts of all consortia, and our approach based on Case Studies with a wide range of partner disciplines is designed to alleviate this issue. Note that this is not an issue when integrating data sets or software from other disciplines, as these typically have their own DOI already and can simply be linked into the knowledge graph(s).

T5: The MaRDI Portal

Overview

The aim of this task area is to develop, implement and maintain an user-friendly way to make the world-wide (digitally) available mathematical knowledge, research data and services accessible to the scientific community – and in particular the services developed within the MaRDI consortium. The main motivation for this is that most of the available knowledge, research data and services are hosted as individual solutions in a silo-like fashion. This hinders a FAIR use of data repositories over larger branches of the mathematical community and beyond. The MaRDI portal will offer a solution to this issue and will become a one-stop contact point for the mathematical community and beyond.

To allow a global search within the MaRDI portal, the newly developed and also already existing external resources will be consolidated and integrated into the MaRDI knowledge graph and thereby made accessible through a low barrier unified user interface and machine accessible interfaces (APIs) – and also through the overarching NFDI knowledge graph.

Besides the search functionality, the MaRDI portal will also allow to store large research datasets and host mathematical services such as algorithm or workflow execution. This functionality will provide a stable and sustainable environment to researchers not having these resources at their respective institutions.

The main goals of this task area are:

- to develop and host a portal that will allow researchers to find and explore available information, research data and services from multiple sources in the mathematical domain – including those developed within the MaRDI consortium.
- to develop and host a storage service for mathematical research data.
- to develop the portal infrastructure for the front- and backend, including the necessary interfaces for integration of (external) data repositories and services.

In order **to achieve these goals** we will work on the following measures (with temporal overlap) which will be implemented jointly between the groups at FIZ and ZIB.

- Measure 5.1: MaRDI Portal Technology
  - **Layers:** **X1: Core**, and **X3: Exchange**
  - **Users:** Mathematical researchers, service developers, researchers from other fields, general public
- Measure 5.2: Data and Service Lifecycle Management
  - **Layers:** **X3: Exchange**
  - **Users:** Data providers, service providers, third party data consumer
- Measure 5.3: Standardization & Interfaces
  - **Layers:** **X4: Knowledge**
  - **Users:** Other NFDI consortia, funding agencies, aggregators

- Measure 5.4: Service Infrastructure
  - **Layers:** **X3: Exchange**
  - **Users:** Mathematical researchers from other task areas, other participants
- Measure 5.5: Distributed Computing and Storage Infrastructure
  - **Layers:** **X3: Exchange**, and **X2: Data**
  - **Users:** Information specialists
- Measure 5.6: Service Development, Integration and Maintenance
  - **Layers:** **X3: Exchange**, and **X2: Data**
  - **Users:** Data providers, service providers, third party data consumer

Measure 5.1: MaRDI Portal Technology

Synopsis Measure **M5.1** comprises the planning, implementation, management, and sustainable operation of a portal to access all MaRDI project-relevant information and resources. This also applies to the uniform web-based and API based access of the consortium's data repositories and information systems. Coordination of the platform occurs centrally through **T5**. However, the portal content is contributed and maintained by the respective task areas.

Relevance and Added Value The MaRDI portal will be the initial contact point for the scientific community for finding and using mathematical services and research data. Thus, within the portal the developed services and solutions of the MaRDI consortium (and potentially also from external providers) will be made available and accessible. This will be simplified by a low barrier unified user interface and supplementary machine accessible interfaces (APIs).

Implementation We will build the MaRDI portal based on a customized version the Wikimedia technology stack integrating the classical MediaWiki frontend and content management system and the Wikibase knowledge graph as data backend. Since the Wikimedia technology stack is designed as a general purpose knowledge management system, significant adjustments to the needs of the math community are required. This includes in particular: persistent identifiers (PIDs), authorization and authentication (OAuth2, shibboleth), implementation of the mathematical knowledge graph and multiple math-specific extensions on all levels, e.g. for the search functionality.

Deliverables

D-TA5-A1 Reporting Including reports about *Implementation planning* (quarterly), *Implementation progress* (annually), *MaRDI portal data statistics* (annually) and the *Security* (annually).

D-TA5-A2 Digital MaRDI Portal The developed front- and backend offer the main functionality for operation of the MaRDI portal, in particular the implementation and search of the knowledge graph.

Milestones

M-TA5-A1 Initial portal online. After 6 months the portal software is installed, the backup and restore functionality is tested, a continuous integration pipeline is established, and two instances (one for development and one for production) are set up.

M-TA5-A2 Permissions set up. After 12 months all relevant players can log in to the portal and upload, view and modify data - according to the planned access permissions as defined by the MaRDI board.

M-TA5-A3 API access possible. After 18 months read, insert, update and delete is possible via APIs allowing for a fast population of the MaRDI knowledge graph.

M-TA5-A4 Math specific data-types implemented. After 24 months additional math specific data types are implemented and accessible from within the MaRDI knowledge graph. After 30 months, formulae in the portal can be searched in the same way as normal text.

M-TA5-A5 External identifiers integrated. After 36 months data in the MaRDI portal can be mapped to external data sources, in particular also to data and metadata of other NFDI consortia.

M-TA5-A6 Federated NFDI queries possible. After 42 months a unified query layer to perform federated queries with other NFDI consortia data has been established and can be used from within MaRDIs portal graphical user interface.

M-TA5-A7 Portal performance improvements In month 52 – additional search indexes will be implemented and essential slow running queries will be optimized – based on the MaRDI use cases to increase overall performance. This process will work iteratively together with query use-case providers.

M-TA5-A8 Portal documentation. At the end of the project the MaRDI portal is fully documented, snapshots of the software are preserved, and the changes to Wikimedia's technical infrastructure are submitted to the foundation repositories.

Measure 5.2: Data and Service Lifecycle Management

Synopsis Research data usually undergoes (state) changes during its life cycle: from data production, through some modified versions, to some kind of final version that needs to be stored in a long-term storage facility while supporting all FAIR principles accordingly. In this measure, we will establish and implement a data life cycle plan that is the basis of all research data related tasks. To this end we will homogenize all services, data, and APIs that will be made available through the MaRDI portal. Ultimately, this will support the assurance of data quality standards and ensure FAIR compliance.

Relevance and Added Value A major aspect of MaRDI is focussed on the appropriate processing of mathematical data and the provision of a corresponding infrastructure. It is important to acknowledge that mathematical research data goes through different states during its life cycle: starting with raw data which is processed and aggregated in various ways and then stored and made available to others, which - again - might process, aggregate and make it available and so on. Long-term storage constitutes the penultimate station of this life cycle, which must be ensured, before potential planned deletion of data closes this cycle.

All current and future MaRDI partners as well as all co-applicants and members from other NFDI consortia will benefit from the implementation of this measure. The establishment of a mathematical standard on how research data can be managed over its entire life cycle is a prerequisite for the sustainability of mathematical research data. Eventually this prerequisite will be established in this

measure which will be a fundamental long-lasting value for the mathematical data culture.

Implementation A senior software engineer (on postdoc level) at FIZ will coordinate, develop and implement service and research data life cycle management strategies for MaRDI, i.e., mathematical research data available from the MaRDI portal. In the beginning of the project, the position will focus on the development of data and service life cycle protocols, service reliability standards as well as API versioning and standardization. Here, we will orientate ourselves on well established best practice solutions such as semantic versioning and long term support releases as well as the current experience of the NFDI consortia started in 2020. In a later phase, the focus will shift towards homogenization of the MaRDI service and data life cycle plans with other NFDI consortia. In addition, the implementation and enforcement of data and service life cycle plans will become an increasingly challenging task as the number of research datasets and services grows. However, clear standards and a strict enforcement of the protocols is essential to ensure high data quality and a pleasant experience when navigating through the connected MaRDI services. All data and service life cycle plans will be tested on zbMATH and swMATH data as well as on the services developed by other MaRDI task areas. Moreover, developer meetings, workshops and hackathons will be organized to pool knowledge, guide training and disseminate standards and best practices.

Deliverables

D-TA5-B1 Reporting Including reports about *Data FAIRness* (annually), *Service Reliability* (annual), and the *Final report about data and service incidents*.

D-TA5-B2 Data Life Cycle Plan The data life cycle plan is the basis of all research data related tasks.

D-TA5-B2 The definite guide for working with mathematical research data. The final report will be openly published as a handbook for future initiatives. In contrast to the annual incident reports, it will have the form of a short and comprehensible handbook.

Milestones

M-TA5-B1 Data live cycle template available. After 6 months a first version of the MaRDI data life cycle template will be available describing the requirement and processes to get data published, updated and eventually archived in the MaRDI ecosystem.

M-TA5-B2 Service live cycle template available. After 12 months the template for orchestration, maintaining, patching, fixing, versioning and shutting down services in the MaRDI ecosystem will be published.

M-TA5-B3 First service life cycle plan available After 18 month the first service life cycle plan based on the swMATH service will be published.

M-TA5-B4 Service reliability measures developed After 24 months key performance indicators for measuring the service performance and mechanism to automatically acquire the required data will be available.

M-TA5-B5 Data and service cycles harmonized with other NFDI consortia. After 42 months we will release a revised revision of the data and service lifecycle templates after consultation and harmonization with other NFDI consortia.

M-TA5-B6 Modular data and service lifecycle strategy released. After month 54 the third version of the data and service lifecycle template will be released.

M-TA5-B7 Data adoption and donation strategy established. At the end of the project the last milestone will be a strategy for data donation or the adoption of other datasets in case of a continuation of the project.

Measure 5.3: Standardization & Interfaces

Synopsis The measure **M5.1** aims to coordinate efforts towards standardization related to the integration of MaRDI services and resources into the MaRDI portal. Moreover, we will define interfaces, anthologies and best practices for the curation of the MaRDI knowledge graph.

Relevance and Added Value Standardizing the MaRDI services and resources requires close interaction with **T1**-**T4** moderators and addresses not only technical and domain-independent standards. Rather, also standardization specific for the mathematics domain related to unique identification (persistent identifiers) as well as for representation is needed and will be developed and implemented. The MaRDI knowledge graph (KG) is a central object within the MaRDI portal that enables data integration and sophisticated search over the diverse information sources. In this measure we will define the necessary interfaces, anthologies and best practices for the manipulation and curation this object. Close collaboration with teams dealing with knowledge graphs from other MaRDI partners, but also other disciplines and NFDI consortia is crucial and will be implemented. This will ensure that the MaRDI knowledge graph is not only a closed self-contained object, but a part of the overarching NFDI knowledge graph. The knowledge graph itself will observe W3C standard technologies in order to enable overarching federated search access on the distributed research data collections internally, across the consortium, and beyond.

Implementation For the implementation, we will apply best practices from the Linked Open Data (LOD) community, W3C standards and the general purpose Wikidata KG. In close collaboration with **T7**, we will integrate an ontology / LOD working group which is derived from the user forum. To ensure the diversity and variety we will invite experts and stakeholders for instance from Wikimedia Deutschland as well as relevant players from other NFDI consortia.

For our standardization efforts, we already collaborate closely with international standardization organizations and work on standards for mathematics and research software. In this measure we will continue our efforts and represent the needs of the MaRDI consortium in these standardization bodies. Moreover, we will organize the congruence of highly problem specific standards with more general standards for mathematics and research data in general, collaborating closely with **T1**-**T4**.

Deliverables

D-TA5-C1 Reporting Including reports about *Initial technology*, *The MaRDI knowledge graph distributed operation* and *Standardization* (annually). *Interface Descriptions* Definitions and implementations about interfaces, anthologies and best practices for the curation of the MaRDI knowledge graph.

D-TA5-C2 Long term data integration strategy statement. The long term data integration strategy statement will elaborate on a sustainable long-time strategy for merging the MaRDI knowledge graph with the overarching NFDI knowledge graph.

Milestones

M-TA5-C1 Ontologies implemented. After month 12 we will have designed, implemented, and evaluated required ontologies (in close cooperation with T1), math specific metadata and data types.

M-TA5-C2 MaRDI metadata standard for scientists and publications established. After month 24 the metadata standard that will be used by all data in the MaRDI knowledge graph is established and communicated to all consortium members.

M-TA5-C3 MaRDI metadata standard for software and institutions established. After month 24 The metadata standard for software and institutions is established based on international standards to annotate software.

M-TA5-C4 Process documentation on MSC assignment to MaRDI artifacts established. After month 36 the process description for mathematics subject classification (MSC) is established and communicated to all consortium members. This allows to optimally classify mathematical research data in its hierarchical classification system.

M-TA5-C5 Crosswalks to competing standards established. After month 48 a process description to allow integration of relevant datasets (with reduced features) that do not comply with MaRDI standards is established and communicated to the MaRDI consortium.

M-TA5-C6 Guide on standards for mathematical research data published. A guide with a special focus on the representation of mathematics in general purpose knowledge graphs and their mathematics specific extensions is published at the end of the project.

Measure 5.4: Service Infrastructure

Synopsis Within this measure we will develop the infrastructure needed to realize and integrate services described in T1-T4. The developed infrastructure will also be used to run the actual MaRDI portal (which is a set of services from this point of view).

Relevance and Added Value The MaRDI portal will allow access to a variety of services, such as portal services (e.g., the user interface or the search functionality) or user-defined services coming from the mathematical community, for example a mathematical algorithm that can act on a given matrix. Traditionally, these services are realized as an executable stand-alone software that needs to be executed on the target platform. However, this approach obviously heavily depends on the available IT environment and can even fail to run, if the software developer's IT and the target IT are not compatible.

Implementation To circumvent the above described difficulties, our approach will be based on operating system (OS)-level virtualization (also called containerization). This means, that the provided service is encapsulated in a light-weight functional module (container) that contains the actual software together with all its dependencies. This gives the service provider maximum freedom in the way how a service can be developed and then be integrated into the MaRDI portal. Within this measure,

we will first evaluate available frameworks such as Docker (as the current de-facto standard), Mesos (since it is well suited for distributed computing frameworks such as Apache Flink) containered, and others for their applicability for the specific requirements in the mathematical community.

An important component for the integration of the containerized services is the definition of the interfaces via which the service can communicate with the underlying service infrastructure. One of the main tasks in this measure will be to design and widely evaluate these interfaces such that services from all over the community can be readily supported.

In the second phase of the project we will build mechanisms to allow remote service execution. Thereby, external service providers can also be integrated into the MaRDI portal while keeping the actual service executing on the providers' hardware. This way, the MaRDI portal allows the integration of a very large number of services without the need to provide large underlying IT resources (see also measure **M5.5**).

Deliverables

D-TA5-D1 Reporting & Training Internal reports will be published regularly: implementation planning (quarterly), implementation report (annually), statistics report (annually), security report (annually). Further, we will hold regular user meetings and technical trainings.

D-TA5-D2 Service infrastructure The service infrastructure offers the basis functionality for all other MaRDI internal and external services. A running instance will be hosted at ZIB.

Milestones

M-TA5-D1 Technology Evaluation done. After 6 months the technology review is done and some initial service prototypes have been implemented and evaluated using the most promising technology candidates. (This will done in cooperation with measure **M5.6**.)

M-TA5-D2 Service Infrastructure ALPHA done. After 9 months a service infrastructure with a minimal feature set has been implemented to allow other teams to adapt to the used technology and develop very early service prototypes. (This will done in cooperation with measure **M5.6**.)

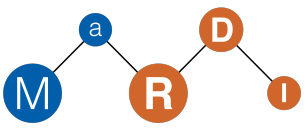
M-TA5-D3 Service Interface Definition done. After 12 months the service interfaces have been defined and evaluated with other development teams from the MaRDI consortium.

M-TA5-D4 Service Infrastructure FINAL done. After 30 months the full-feature system of the service infrastructure is up and running.

M-TA5-D5 Remote Service Execution done. After 40 months the system for remote service execution (MaRDI-RSE) is up and running. (This will done in cooperation with measure **M5.5**.)

M-TA5-D6 Remote Services Integration. After 55 months the main remote services from the MaRDI consortium are integrated based on the MaRDI-RSE system. (This will done in cooperation with measure **M5.5**.)

M-TA5-D7 Documentation finalized. At the end of the project (month 60) the full documentation is available and the final software codebase is preserved.



Measure 5.5: Distributed Computing and Storage Infrastructure

Synopsis Within this measure we will develop and integrate the needed infrastructure for distributed computing and distributed storage which is needed for running the available services within the MaRDI consortium. This includes the technology needed for running the portal, long-term storage for research data and integration of external data-storage facilities. We will also support and moderate the processes leading to integration of other partner resources.

Relevance and Added Value Services and datasets available through the MaRDI portal will be contributed by numerous members of the MaRDI community as well as by external providers. One of the main goals of the MaRDI portal is to enable unified access to MaRDI resources and services that are hosted at the central MaRDI site and on external sites. Thus, the system will be designed to allow – transparent to the user – remote execution of services on the individual provider's site to avoid transfer and synchronization issues for large dynamic datasets that are also stored on the respective provider's site.

Implementation To support the above described scenario, the MaRDI portal – and thus the underlying infrastructure – must allow the integration of services and datasets that are physically located at a remote site. The main component of this system will be the development of protocols and interfaces enabling exchange of control sequences from the MaRDI portal to the external systems. Not only must the main (portal) system be in synchronization with the remote systems at all times but also downtime of the connected systems must be detected and handled accordingly, to avoid user frustration. Additionally, mirroring and spill-over functionality will be implemented to allow load-balancing of service and data access throughout the connected locations. This will be based on available open source technology for peer-to-peer storage techniques for distributed file systems and distributed object storage systems, e.g. for S3-alike systems such as minIO. In the long-run of the project we will enable integration of locally available queuing systems that might already be in place in the participating compute and storage facilities, e.g. based on Slurm. Storage and Service infrastructure will be implemented in coordination with measure **M5.2**.

Deliverables

D-TA5-E1 Reporting & Trainings We will regularly publish internal reports, such as: implementation planning (quarterly), implementation report, statistics report and a security report (all annually). Further, we will hold regular user meetings and technical trainings.

D-TA5-E2 Distributed Service Infrastructure This system provides the necessary interfaces and end-points to allow integration of services executed on external hardware outside of the MaRDI core portal.

D-TA5-E3 Distributed Storage Infrastructure This system allows integration of external storage facilities that can be used transparently by the MaRDI portal and services.

Milestones

M-TA5-E1 Technology Evaluation Done. After 6 months the technology review is done and initial service prototypes using distributed storage have been implemented and evaluated.

M-TA5-E2 Distr. Storage Infrastructure ALPHA Done. After 9 months a prototype for a distributed storage infrastructure with a minimal feature set has been implemented to allow other teams to adapt to the used technology within their own prototypes.

M-TA5-E3 Distr. Storage Infrastructure FINAL Done. After 30 months the full-feature system of the distributed storage infrastructure is up and running.

M-TA5-E4 Long-term Storage Facility Available After 36 months a system for long-term data storage is up and running, based on the distributed storage infrastructure, including integration of external services such as RADAR and ZIB's tape archive.

M-TA5-E5 Remote Service Integration After 40 months the system for remote service execution (MaRDI-RSE) is up and running. (In cooperation with measure **M5.4**.)

M-TA5-E6 Documentation. At the end of the project (month 60) the full documentation is available and the final software codebase is preserved.

Measure 5.6: Service Development, Integration and Maintenance

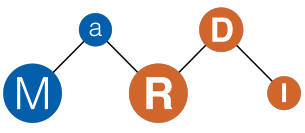
Synopsis This measure is concerned with the provision of the actual content for the MaRDI portal from the view-point of a service provider. This will be implemented with two perspectives in mind: (1) we will support all other task areas (e.g. scientific computing or statistics) to integrate their planned services into the MaRDI portal. (2) We will develop new services in an agile fashion (see below). Further, we will support and moderate the integration of other services that will be developed within the MaRDI consortium.

Relevance and Added Value One part of this measure will focus on the integration of the exemplary services from the other MaRDI task areas. In addition to that, several other services will be developed in an agile fashion. By this we mean that the decision which services will be integrated next will be made through a committee that will meet quarterly. In these meetings, representatives from all task areas will create a list of those services that should be developed next. This allows to dynamically react to changes in prioritization of the essential features that the MaRDI portal should provide. Further, it allows to dynamically react to changes in the (human) resources. For example, developers from other teams could join the service development team if they have available resources and thus help to increase the number of services that can be deployed.

Implementation The first round of services will be focused on the integration of established data sources, such as zbMath, swMath, arXiv, EuDML, OEIS, Wikidata and Encyclopedia Mathematica. Here, we will work together with the respective content providers to build cross-content services and to support them in adapting the metadata standards and delivering mathematical knowledge that is encoded according to best practices and standards in cooperation with measure **M5.3**. Within this measure we will also develop other needed core services, e.g. a formula search engines, a mathematical question answering system, a mathematical plagiarism detection systems, computation engines, or a formulae recommender systems.

Deliverables

D-TA5-F1 Reporting & Trainings We will regularly publish internal reports, such as: implementation



planning (quarterly), implementation report, statistics report and a security report (all annually). Further, we will hold regular user meetings and technical trainings on the topic of best practice for service development.

D-TA5-F2 MaRDI Portal Services Various services will be implemented, integrated and made available through the MaRDI portal. A programmers best-practice handbook will be published containing exemplary use-cases and tips and tricks for best performance.

Milestones

M-TA5-F1 Technology Evaluation Done. After 6 months the technology review is done and some initial service prototypes have been implemented and evaluated using the most promising technology candidates. (This will done in cooperation with measure **M5.4**.)

M-TA5-F2 Service Infrastructure ALPHA Done. After 9 months a service infrastructure with a minimal feature set has been implemented to allow other teams to adapt to the used technology and develop very early service prototypes. (This will done in cooperation with measure **M5.4**.)

M-TA5-F3 More and more services implemented. We aim to release a new service every month. However, due to the agile nature of this measure, no planned milestones can be formulated.

M-TA5-F4 Documentation. At the end of the project (month 60) the full documentation is available and the final software codebase is preserved.

Embedding into NFDI

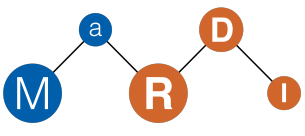
Within this task area we contribute to all 6 MaRDI objectives **O1** – **O6** by either directly offering services that support the respective objectives (e.g. execution of workflows for **O3**) or support the individual task areas by developing standards, interfaces, specifications and guidelines (e.g. for interoperable mathematical research data for **O1**). Further, measure **M5.6** directly contributes to objective **O4** (“Enable the development of mathematical services”).

On the global level, we complement the efforts of **T4** and **T6** to coordinate with other NFDI initiatives on the data and metadata level. The overall goal of the MaRDI knowledge graph is not only being embedded into the overarching NFDI knowledge graph but to become a subgraph of one logically centered (and physically distributed) NFDI knowledge graph. This enables federated queries across multiple NFDIs.

Requirements / preliminary work

The mathematical information retrieval (MathIR) community has developed several services in the past decades [GSC15; Sch17]. One of the main goals is to connect symbolic and narrative mathematical research data to be human readable for interpretable mathematics. A recent example where members of this consortium have contributed are the swMATH and zbMATH systems, where mathematical software from swMATH is linked to the appropriate corresponding publications in the zbMATH system. This interplay of math-specific and general data processing services is – in general – well researched [SLM13; SRN18; ST19; ZY15] and will be the foundation for the work in this task area.

FIZ and ZIB have extensive knowledge and experience in building and maintaining large software projects, in storing and analyzing of very large research data-sets, in ontology and knowledge graph



engineering as well as in the design, implementation, and maintenance of general large scale scientific information infrastructures.

Risks and Mitigation

Within this task area we found three main risks to be relevant (in brackets: likelihood / severity): competing standards (low, less critical), IT problems, such as hardware downtime or loss of data (low, critical), and slow response time of the MaRDI portal (high, medium).

We designed the measures in this task area to include well established technical and procedural techniques including project management and design specific areas. For example, specific counter-measures to slow response time of the portal will be a major activity of measure **M5.1**. Here, several options are available to help if technical measures are insufficient to handle the presumably high load of data and queries, such as appropriate adjustments of the data and service lifecycle management plans.

T6: Data Culture and Community Integration

Developing and establishing a common data culture is the core objective of MaRDI. All other objectives rely heavily on the acceptance of MaRDI by the mathematical community in Germany and internationally. This task area is designed to address this challenge for the German community. The network of MaRDI partners and participants will help to establish an international standard for FAIR data due to the global nature of scientific collaborations. Some communities in the mathematical sciences already are developing a data culture (scientific computing, machine learning, AI research, statistics, etc.), and MaRDI has dedicated task areas to set standards within these disciplines and to connect to professional societies like GOR, GAMM, and IMS.

The primary target group for this task area is the mathematical community at large. A particular focus lies on those disciplines that are currently not working on a data culture and maybe are not even aware that they produce data as a valuable resource that should be available in a FAIR way. Doing this, we take gender diversity into account. Today, the mathematics community is still struggling to offer adequate career perspectives for promising young female researchers. We will take dedicated measures towards reaching gender balance in all our activities. Ultimately, FAIR mathematics needs to be fair with respect to gender aspects. Therefore, we propose to work closely with learned societies—in particular the DMV and the EMS—as well as building support for data in mathematics on an institutional level via the library system and IT support services at universities and research institutions. This defines another target group for this task area: scientific service specialists who will support and disseminate the common data culture. In this way, we aim to establish a FAIR data culture on a fundamental level in the mathematical community. Obviously, the establishment of a specific common data culture cannot stop here, and there is a wider audience to address. This includes neighboring disciplines as well as a general academic audience or public stakeholders. All of these will be addressed as our third target group.

MaRDI will promote the FAIR principles and support the establishment of the common data culture following two main ideas: showcasing best practice examples via the MaRDI portal and collecting user feedback to incorporate new features into the portfolio. This approach will be successful only if a majority of mathematicians see an advantage in using MaRDI for their own research agenda. We understand the mathematical community as customers of MaRDI and they need to see an advantage in using MaRDI over handling data in their usual way. Funding agencies and publishers increasingly demand compliance with FAIR principles. A unique selling point of MaRDI is that it helps mathematicians to fulfill these demands.

While it is necessary to raise awareness of the existence of MaRDI, the key goal is for scholars to see an immediate personal benefit in MaRDI. To make this happen requires a dedicated strategy to market MaRDI and in particular the MaRDI portal. This includes speaking to mathematicians about the data that are relevant for their own research.

Another important aspect is to gather and integrate the needs and visions of the community that we aim to serve. This is the second goal of our task area: We aim to build a broad network of MaRDI allies across all mathematical disciplines to integrate the mathematical community into the development

process of MaRDI, e.g. by collecting feedback on MaRDI (like feature requests, technical issues, etc.). The MaRDI portal plays an important role here and the development of the portal must therefore integrate community feedback (like use cases) that we collect and communicate.

We propose three pillars for achieving these goals (cf. Figure 9):

1. Mathematical Data Consultancy
2. MaRDI Workshops
3. Interactive Dissemination

These three pillars carry our measures, which we identify with the specific target groups mentioned above. The tasks and milestones explain the concrete steps that we plan to take to achieve our goals. The pillars are based on Measure **M6.0**: Internal Communication and Dissemination Coordination, where we anchor the overall strategic and organizational aspects of **T6**. Measure **M6.0** will also link the topical task areas **T1**-**T4** to this task area and coordinates the exchange between the communities in computer algebra, scientific computing, statistics, and other relevant areas.

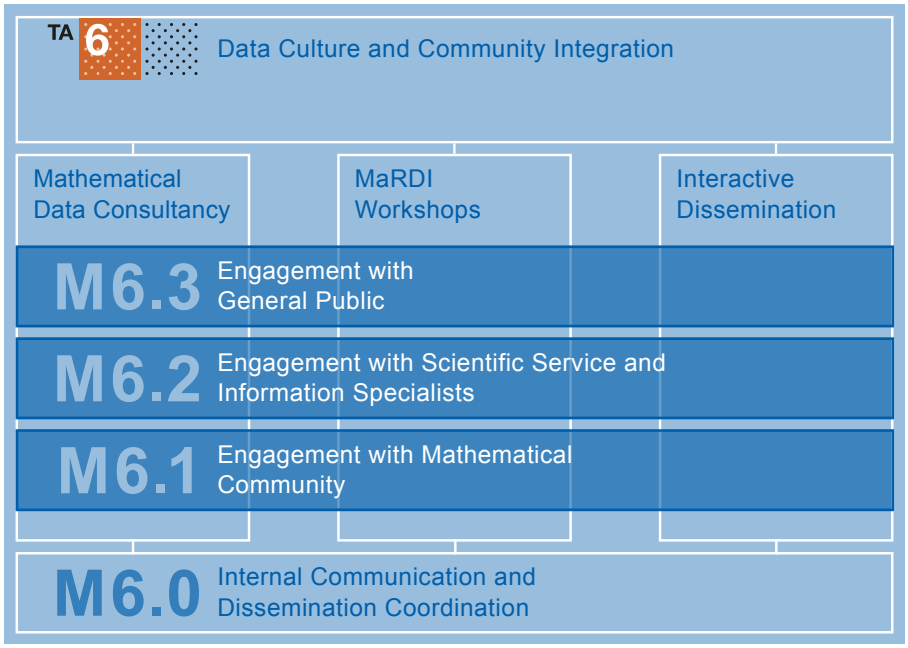


Figure 9: **T6**—Target groups (measures) and methodological pillars.

Before we describe our measures and how they reflect our pillars in more detail, we will briefly discuss the overall team working in **T6**. For more concrete job descriptions and budget issues, we refer to subsection **??**. MiS will lead the overall task. MiS will hire a Dissemination Coordinator (*DisCo*) and contribute the working time of a professional graphic designer to the project. FUB will hire a Mathematical Data Consultant (*MDC*) who will closely cooperate with the DMV Media Office whose leader commits working time to the project as an in-kind contribution. IMAGINARY will reinforce its team with a MaRDI Communicator (*MCom*) and MFO will employ a part-time MaRDI Librarian, partly as an in-kind contribution (*MLib*).

The Mathematical Data Consultancy will be led by the DMV in close collaboration with FUB (MDC and in-kind contribution). MiS (in-kind) and MFO (MLib and in-kind) also contribute to this pillar. Being the association representing all of Mathematics, DMV has more than 5000 members. Using this network, the Mathematical Data Consultancy will easily get in touch with target groups 1 and 3. MiS and MFO both have strong mathematical libraries and will use their professional network to reach out to target group 2.

The MaRDI Workshops pillar is sustained by IMAGINARY (MCom), MiS (DisCo and in-kind), and MFO (in-kind). MFO is a conference center with considerable experience in hosting research workshops, MiS is a world-class research center that also runs an extensive visitor program, and IMAGINARY has a lot of experience in providing participative training in mathematical topics to a diverse audience. All partners will contribute parts of their conference and training infrastructure to this task area.

The development of sophisticated digital content communicating mathematical research is a non-trivial task. IMAGINARY has established itself as a central hub to create and share interactive dissemination material within the mathematics community. IMAGINARY additionally strives to engage a larger audience towards understanding mathematical research and has 12 years of experience in more than 400 dissemination activities in 60 countries. Based on this development work, DMV/FUB (MDC) and MiS (in-kind) will support IMAGINARY's interactive dissemination team (Com and in-kind).

Measure 6.0: Internal Communication and Dissemination Coordination

Synopsis All dissemination activities need to be coordinated and kept track of. This measure will take care of a constant flow of information between TA6 and TA 1-5 regarding dissemination activities. The coordination supports the organization of workshops and further suitable outreach activities along the timeline of MaRDI activities and developments. It will also assist in connecting the MaRDI consortium with the target groups 1 and 2.

Added Value The overall concept of this task area relies on a well-established flow of information within the team members but also between the **T6** members and the other task areas—in particular with members of task area **T5**. This information flow will be organized in this measure.

Implementation This measure will be led by the task area lead Bernd Sturmfels (Director MiS). Weekly online meetings of all members of the task area will guarantee the information flow within the task area. (Deliverable D-TA6-O1) During these meetings, the overall dissemination strategy will be further developed and adapted. Constantly, DisCo and MDC will update the other team members about ongoing developments in the project. Vice versa, DisCo will collect feedback the members received during the past week and prepare feedback reports to share with the respective task area. (Deliverable D-TA6-O2)

This measure will also provide organizational support for all other measures.

Resources The tasks in this measure are mainly executed by DisCo. MiS contributes graphics and IT support and IMAGINARY also contributes IT support.

Deliverables

D-TA6-O1 Regular meeting of team members.
D-TA6-O2 Feedback reports to members of **T1**-**T5**.

Measure 6.1: Engagement with Mathematical Community

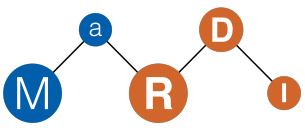
Synopsis Wide acceptance of FAIR principles and good integration of MaRDI's services in the mathematical community is the foundation on which MaRDI has to build. So this measure is clearly es-

sential and our main priority. On the one hand, we will inform the mathematical community about MaRDI's goals, progress, and features through diverse channels. On the other hand, we will make every effort to collect community feedback (from general points of view on research data in mathematics to concrete feature requests for the MaRDI portal), again by diverse choices. We will be present at mathematical congresses (DMV Jahrestagung, European Congress of Mathematics, International Congress of Mathematicians) and have MaRDI representatives and members at workshops and conferences through an ally program to cast a broad net and to collect input from a wide variety of working mathematicians.

Added Value Active participation and engagement of the mathematical community in building the common data culture are essential for MaRDI's success. By the means of this measure, we guarantee that FAIR principles become widely accepted among mathematicians and that the research data infrastructure provided by this project becomes standard among mathematicians.

Implementation This measure will be led by Rainer Sinn, currently head of the Discrete Geometry group at FUB which is hosting the DMV Medienbüro. Our first main goal is to inform the mathematical community at large about FAIR principles and mathematical research data. We will achieve this goal by disseminating information with the support of the networks that are part of the MaRDI consortium. In particular, the DMV and the EMS will play an important role here. We will showcase best practice examples and MaRDI services at mathematical congresses (DMV Jahrestagung, European Congress of Mathematics) at a MaRDI station and information desk. This MaRDI station will be supported by the interactive dissemination team. We will make these best practice examples available online and advertise it on the portal www.mathematik.de (which is hosted by the DMV and maintained by the DMV Medienbüro). We will produce content for the DMV Mitteilungen, starting with a presentation of MaRDI as a project and FAIR principles and later conducting interviews with researchers who published mathematical research data for the first time in their career with a focus on people who work in areas that are perceived to be "pure" like topology or number theory. This can lead to striking and unexpected benefits. For instance, memory errors on computers were detected by experimental computation of class groups of number fields, which shows the use of mathematical research data from number theory by Buhler and Harvey [BH11]. We will interview researchers with such results to increase awareness of the diversity of mathematical research data and the importance of FAIR principles. (Deliverables D-TA6-A1/-A2) This first goal is mainly the responsibility of MDC.

Our second main goal is to gather feedback from the mathematical community in order to communicate it to the MaRDI developers (in particular, the portal team **T5**) so that the diverse needs can be met and increase acceptance of MaRDI. We will achieve this goal by informal meetings with research mathematicians at congresses and through a structured program. The structured program includes surveys prepared by the experienced support at MFO. They will be regularly conducted at MFO during their weekly workshops and schools and may also be available online. We will conduct surveys at mathematical congresses like the DMV Jahrestagung. (Deliverable D-TA6-A3) With the MFO support, we will reach a large and diverse group of mathematicians and expect new insights with respect to MaRDI services. Within the program, we will also organize MaRDI workshops together with other task areas. These workshops will be attended by researchers in the respective fields with research



talks as usual at mathematics conferences but they will also have MaRDI specific content. This will include presentations of best practice examples and FAIR principles as well as hands-on activities with interactive information and MaRDI portal experts. (Deliverable D-TA6-A4)

An important tool for our activities is our MaRDI ally program. We will recruit young and established researchers from various mathematical disciplines to support MaRDI in their expert communities as MaRDI allies. They will present MaRDI services and mathematical research data from their community to the community at expert conferences (Deliverable D-TA6-A5). They will be in regular contact with MDC to stay informed about developments in MaRDI (through online tools).

The MaRDI ally program will support our goal to raise awareness of research data in mathematics and FAIR principles broadly in the mathematical research community through targeted and specialized information. In Measure **M6.3** we will describe the development of a MaRDI station—a tool designed to showcase MaRDI to a general audience. Obviously, we will also use this station in this measure.

Deliverables

D-TA6-A1 Online library of research data examples and interviews. (pillar I)
D-TA6-A2 Annual activity report. (pillar I)
D-TA6-A3 Annual surveys about awareness of FAIR principles. (pillar II)
D-TA6-A4 Attractive offer of workshops. (pillar II)
D-TA6-A5 Presence at least at one major mathematical conference per year. (pillar III)

Milestones

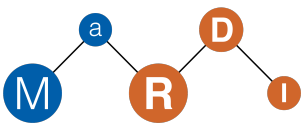
M-TA6-A1 Surveys show that 50% of participating mathematicians are aware of MaRDI.
M-TA6-A2 100 mathematicians participated in at least one workshop.

Resources The tasks are executed by MDC (focus on pillar I), DisCo (focus on pillar II), and MCom (focus on pillar III). MiS and MFO will support the activities accommodating workshops (in-kind). Visiting conferences and offering workshops require a travel and consumables budget. Details will be presented in subsection **??**.

Measure 6.2: Engagement with Information Specialists

Synopsis Establishing a sustainable data culture needs support on many different levels. Integrating library and IT infrastructures and information and research data management specialists working in scientific services in research institutions, therefore, is an important component of our strategy. They have a close connection to the research community, and it is their task to cater to the community's needs in order to support research, combined with the specialist knowledge of information professionals. As of now, there is no congruent approach to dealing with mathematical research data in research libraries or within the research data management at research institutions. Efforts in this community are scattered. There is no sustained existing network to jointly work on these questions.

We can overcome this by building on the gathered expertise and experience inside the MaRDI consortium—e.g. zbMath and FIZ, MFO, TUK—and making use of the existing cooperations between information specialists and researchers at MaRDI institutions and their respective networks



in particular in the Leibniz Association and the Max Planck Society. The aim of this measure is to build up a network of information specialists and research data managers working in (mathematical) libraries, research institutes, with academic publishers, funding agencies, etc. to incorporate their view and feedback into the MaRDI activities and to disseminate information about MaRDI and its developments into this community.

Added Value Having another vehicle to transport MaRDI developments directly to the community and to embed MaRDI services in the research support infrastructure. Incorporate additional expertise into the efforts for standardization in **T1**-**T4**.

Implementation This measure will be led by the scientific coordinator of MiS, Jörg Lehnert. We want to create a bridge between the mathematical community and this target group by regularly transferring information about MaRDI developments, collecting feedback, and disseminating it back to the teams in TA1-5. We will build a network, jointly work on ideas for standardization and other input for the efforts of TA1-5 and have a regular exchange of expertise.

Deliverables

D-TA6-B1 Directory of mathematical research data specialists at German universities and research institutions. (pillar I)
D-TA6-B2 Annual activity report. (pillar I)
D-TA6-B3 Annual workshop or network meeting of relevant stakeholders of this target group. (pillar II)
D-TA6-B4 Online training material for employees in scientific service. (pillar III)

Milestones

M-TA6-B1 Directory of mathematical research data specialists contains 50 experts.
M-TA6-B2 Online training material gives a comprehensive overview of research data management with MaRDI.

Resources The tasks belonging to pillar I are carried out mainly by MLib and MDC, those of pillar II by MLib and DisCo, pillar III is covered by MCom. Like above, visiting conferences and offering workshops requires a travel and consumables budget. Details will be discussed in subsection **??**.

Measure 6.3: Engagement with General Public

Synopsis MaRDI will drive how mathematicians think about and work with research data. This is a major change in culture and thus needs to be accompanied by a general awareness campaign and dedicated dissemination activities. These are targeted to a broader scientific, but also a general audience. Large scale awareness and acceptance of an open science and research data culture are fundamental to establish the basis for MaRDI to be successful and sustainable. The general dissemination activities will focus on the FAIR principles and the benefits of MaRDI for our society. It is planned to involve journalists from a variety of media, politicians, and (science) museum stakeholders and through them reach out to the general public.

Added Value MaRDI actively contributes to a general awareness of an open research data culture and establishes itself as a central driving force with a benefit for society.

Implementation This measure will be led by Andreas Matt, director of IMAGINARY.

A modular workshop and presentation system will be set up. This way, MaRDI can be present at small or large events, conferences, or workshops with informal presentations and reach out to a diverse audience. For instance, there is an increasing demand for soft-skills seminars in structured graduate schools and even in undergraduate education and we will develop a suitable module.

As a core engagement module, an interactive exhibit, the MaRDI station, will be developed to explain the project and its services and outcomes to various audiences. It is an open source and customizable software program for touch screens to be presented at workshops, conferences, events, and to be temporarily or permanently installed at research institutes, universities, or science museums. The station invites everybody to interact with curated MaRDI content in an attractive and engaging way. It is closely linked to the content and services of the MaRDI portal.

All activities are backed by the development and regular adaptation of interactive dissemination media for a physical and digital presence of MaRDI. They include printed flyers, posters, and roll-up banners as well as an introductory short animation film, images, and film clips for social media. A distribution strategy and network will make sure that all media are easily accessible and used. Established services and networks (Wikimedia/Wikipedia, ECSITE) are used and integrated.

Deliverables

D-TA6-C1 Annual activity report. (pillar I)
D-TA6-C2 Virtual MaRDI Research Data Workshops and Presentations. (pillar II)
D-TA6-C3 “Infrastructure” to book MaRDI lectures (with a pool of presenters and a variety of formats), online resources for self-information. (pillar II)
D-TA6-C4 MaRDI station (source code, installation, installation manual, presentation manual). (pillar III)
D-TA6-C5 Web animation film (1 minute). (pillar III)
D-TA6-C6 Digital resources to download (flyers, posters, roll-ups) for self-printing or digital use. (pillar III)

Milestones

M-TA6-C1 MaRDI station is installed in 2 science museums and shown at 4 conferences.
M-TA6-C2 First MaRDI lecture is booked.

T7: Governance and Consortium Management

The overall management of the consortium will be the responsibility of WIAS (spokesperson: Michael Hintermüller, the Director of WIAS). It will be supported by staff applied for within this proposal as well as by resources of WIAS (in-kind). Specifically, WIAS will set-up a MaRDI/NFDI secretariat for which it will provide corresponding office space along with standard technical equipment and infrastructure as well as some support personnel. This also includes setting-up and maintaining an appropriate web environment for information on MaRDI and its services. The overall management duties will be split into administrative ones and scientific/technical ones, respectively, and will be further specified below.

The general management structure of MaRDI aims at overseeing the internal topic development, task progress, and interplay with designated use cases, partly motivated or jointly developed with other NFDI consortia or disciplines. It also steers the agenda, structure, and timelines of the consortium concerning work-packages and work-flows, projects, services, and interfaces to current research in order to reach a long-term viable organizational structure within the funding period. Moreover, it organizes and supports the governance structure and duties of various boards within MaRDI. In addition, a series of activities within the consortium (such as regular meetings) and workshops in open space formats to involve user groups (including users external to MaRDI) as well as potential new participants or associated institutions, interest groups or individuals will be arranged and coordinated by the MaRDI management.

Technically, the governance of MaRDI is based on a federated structure of responsibilities. The task areas with their respective leading spokesperson form the consortium board (CB). Together with the elected spokesperson of the consortium, the CB will constitute the executive and representative core of the MaRDI structure. The co-applicants of MaRDI join the CB in the consortium council (CC) by their co-spokespersons, where the strategic development of the objectives and structure of the consortium takes place. The CC is the link to the members of the consortium (in the sense of this application, i.e., the applicants and co-applicants). The third body in the structure is the general assembly (GA) constituted by the CC and all participants (in the sense of this proposal). The GA is also open to associated members of MaRDI. In order to foster and perpetuate cooperations with other NFDI consortia enabling joint developments, we will integrate representatives of partner consortia as guests on all levels of the structure (BC, CC, GA). The council will be advised by an international Advisory Board (AB) and a User Forum (UF). The boards will, among others, guarantee adherence to the FAIR principles within all developments in the consortium. MaRDI's governance will further implement a board for internal and external interoperability. While the former guarantees inner coherence of MaRDI activities and services, the latter is clearly geared towards the entire NFDI network in order to develop common standards for securing work- and data-flows across NFDI consortia. The governance structure and interrelations between the associated boards are depicted in Figure 10.

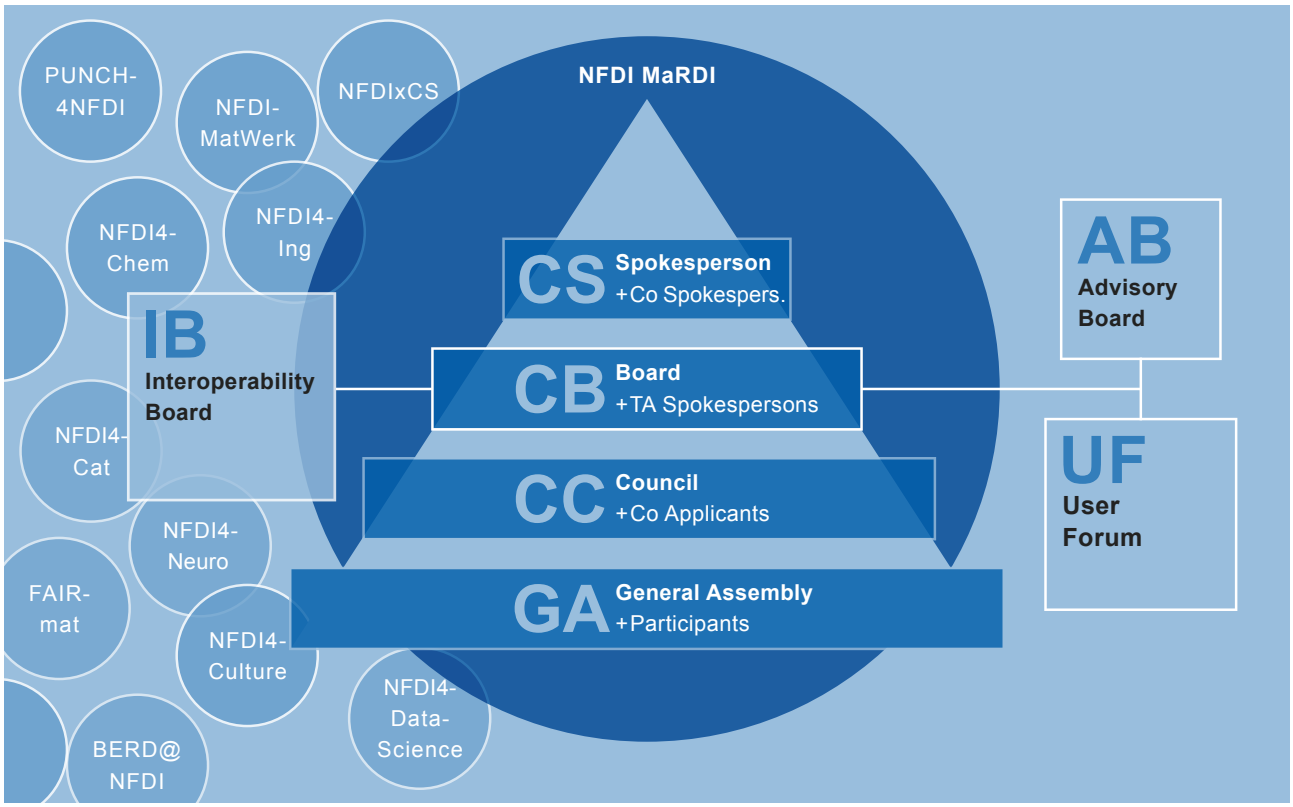


Figure 10: Center: MaRDI internal board structure including the group of the consortial speaker + co-spokespersons (CS). Left: Cooperating NFDI consortia and interoperability board. Right: External advisory and user platforms.

Designated board member (CB)	Task Area
Michael Joswig	T1
Peter Benner	T2
Mathias Drton	T3
Anita Schöbel	T4
Christof Schütte	T5
Bernd Sturmfels	T6
Michael Hintermüller	T7

Next we describe the various responsibilities and duties of the foreseen boards in more detail.

The CB is responsible for administering and endorsing the Ts, along with associated timelines and cooperation structures, as well as other MaRDI structures. This includes reflecting on the work development within current Ts whose progress is typically due to dynamically assigned projects, introducing new projects or Ts, or dismissing existing ones. It will also be responsible for overseeing the shaping of services which result from MaRDI activities. More specifically, the CB will run a Technical Committee whose task is the certification of MaRDI services and products prior to their integration into the digital portal. If such a certificate cannot yet be issued, guidelines for improvement of the respective

item will be given. Also, decisions on training activities for a wide dissemination of proper workflows, policies, synergies as well as services will be taken.

The CC is responsible for maintaining and strategically developing the overall vision of MaRDI and for ensuring the adherence to guiding principles such as the FAIR-principles. It oversees the general development, strategy, and achievements (i.e., successes and needs for further development) as well as sustainability of MaRDI activities.

The largest body of MaRDI is the GA. It consists of all members of MaRDI and is a forum of exchange between various groups, most notably including users of MaRDI outcomes. Through this consortium the internal structural decisions will become transparent to a wide community (note here that MaRDI participants are active interest groups or learned societies such as GAMM, GOR, and EMS; also MaRDI is strongly interlinked with Clusters of Excellence with emphasis on mathematics and its applications and which are funded within the current German Excellence Strategy). Vice versa, through the GA, strategic decisions of the consortium will be informed by feedback from and exchange with user communities.

On all levels, designated representatives of other NFDI consortia will be integrated as guests. This approach is aimed at fostering the creation of a robust network of NFDI consortia (in particular with relation to mathematical data) which shall become the backbone structure for national research data.

MaRDI will also be guided by advice from external experts boards. The AB will provide valuable recommendations and perspectives concerning the overall development of MaRDI. It shall consist of researchers, policymakers, possibly selected industrial partners and others involved in mathematics related science and information technology. Through its wide ranging experience, the AB will assist in achieving the long-term goals and objectives of MaRDI. The board will also be valuable in discussions for shaping the form of MaRDI within NFDI along with its partners and stakeholders for securing optimal long-term services.

The UF will be composed of representatives of MaRDI related research communities and data repositories. It represents a designated interface to users of MaRDI services and it is an essential mechanism for feedback on the objectives and tasks as well as the consortial composition of MaRDI. Expressing needs and concerns, as well as discussing potential future developments regarding mathematical research data management and infrastructure, also in view of the entire NFDI, will be important functions of the interface between the UF and the MaRDI team. While there will be a focus on national members in the UF, we also plan to involve several strategic partners from international organisations with experience or active agenda along open access, FAIR principles and user-motivated services in connection with (mathematical) research data, such as EMS, SIAM, or NIST.

Work in this task area is structured into three measures:

- Measure 7.1: Establishing MaRDI Administrative Structure and Governance
- Measure 7.2: Strategic Embedding and Organization

Measure 7.1: Establishing MaRDI Administrative Structure and Governance

Initially, the entire MaRDI-structure will be set-up. This includes a secretariat, web platform for information and exchange, the communication tree for and with the task areas (T) as well as the legal

contact point for general issues regarding FAIRness, licenses and mathematical data use. Another very important task is to rapidly set-up the MaRDI board structure, where initial emphasis is given to CB, CC, and setting up an information dispatcher for exchange with the MaRDI community in the GA. Simultaneously, the AB will be recruited and set up. Afterwards, also in exchange with stakeholders and scientific societies and groups connected to MaRDI, the UF will be formed.

On the organizational side, the entire financial administrative set up will be fixed at WIAS. Obviously, this has to rely on written cooperation agreements between co-applicants and participants. For this, the underlying letters of commitment form the basis.

Once in full operating mode, **T7** will run the entire governance including board meetings and realizations of board decisions. It will administer the internal review and strategy mechanisms. Moreover, it will provide the legal services for MaRDI members and potential requests on mathematical data rights from other NFDI consortia.

On the broader NFDI landscape, **T7** interchanges MaRDI progress, services and potential cooperation projects with other consortia. As mentioned above, in order to facilitate and foster such an exchange MaRDI foresees guest positions for representatives of other consortia in MaRDI boards. But, of course, MaRDI will also actively exchange with other consortia.

Finally, **T7** also holds an agile budget. These funds are indispensable as it is already evident now that there will be dynamic needs from various task areas as the strategic work of MaRDI unfolds. Here, we exemplify two such needs. For instance, MaRDI establishes research data management (RDM) services, tools and infrastructures that will raise legal questions (e.g. use of terms, liability, copyright for databases, recently in the EU introduced provisions for data mining, implementation of the FAIR principles). For such tasks MaRDI will provide legal support as soon as needed. When legal standards are developed at a later point, MaRDI can also include legal aspects in the education and training activities for researchers and for RDM staff (**T6**). The legal contact point can then review and contribute legal content to the training concepts provided by MaRDI. On the other hand, the work on the digital portal will experience periods of heavy duty, in particular at the points in time when prototypes from **T1**, **T2**, **T3** become available and need to be converted into broader services within the portal. This work would be in addition to the routine development work on our agile portal solution.

Measure 7.2: Strategic Embedding and Organization

This measure is the main block that coordinates and orchestrates the strategic work of MaRDI. This concerns the selection and realization of work projects, work packages, milestones and timelines, depending on status reports from the various constituents of MaRDI and in alignment with board decisions. Most importantly task area **T7** also defines, given the exchange with the CB and CC under consultation of the AB, the services to be focused on and implemented in MaRDI's central digital portal. Of course, also the monitoring of the strategic decisions in the CC and their realization fall into this measure.

In order to facilitate the flow of information and to continuously develop the MaRDI community (internally as well as externally), **T7** organizes at least twice a year a GA and further an annual MaRDI meeting. While the latter is meant as a forum of exchange by exhibiting MaRDI's work progress and strategic lines also inviting other NFDI consortia and scientific communities to present their respec-

tive progress and status on topics connecting to mathematics, the GAs merely serve the purpose of an open internal exchange of the MaRDI team in order to identify obstacles on the way, to highlight achievements and to exchange information of general interest. **T7** will also set-up and maintain several internal communication channels ranging from regular news bulletins via Email to blogs.

The embedding and networking of MaRDI, however, does not stop at the aforementioned level. Rather, **T7** will also strategically embed MaRDI not only within the NFDI, but also within and between scientific societies (both nationally as wells internationally). It will also make sure that connections to developments such as Open Cloud Europe, FORCE11 or other then newly established initiatives will be made, and, wherever useful, **T7** will maintain and develop such contacts.

For measure **M7.1** and **M7.2** WIAS will implement a position of a scientific coordinator and one position for the administrative work.

Appendix

1 Bibliography and list of references

arXiv.org is a preprint portal for physics (45%), mathematics (30%), computer science (18%), etc. The MPG Digital library is involved in the organization. The SIGMathLing(*)[23] Special Interest Group on Mathematical Linguistics (maintained by FAU) maintains and distributes various narrative data sets, in particular, the HTML5+MathML translation of the 1.5M arXMLiv articles, mathematical word embeddings, and fragment classification datasets. Together with EuDML and zbMATH, this gives us a representative data set of mathematical full-texts that can be used for outreach, system evaluation, and benchmarking.

DataVerse is an open source web application to share, preserve, cite, explore, and analyze research data.

DepositOnce(*) is a repository for research data and publications hosted at TUB.

The **Digital Library of Mathematical Functions (DLMF** provides machine-readable semantics that allows for formulae search and interactive display of additional metadata. This includes links to definitions for the symbols and identifiers used in the formula, references to proofs and sketches of proofs when proofs are not available on the literature, as well as hyperlinks to related concepts.

The **Distributed and Unified Numerics Environment**(*) DUNE is a free and open source software framework for solving partial differential equations. It is developed since more than 15 years as a collaborative effort of several universities and research institutes. In its name, the term “distributed” refers to distributed development as well as distributed computing. The distinguishing feature of Dune is its flexibility combined with efficiency. The main goal of Dune is to provide well-defined interfaces for the various components of a PDE solver for which then specialized implementations can be provided.

The European Digital Mathematics Library (**EuDML**)(*) is the largest collections of open access publications in mathematics. Developed and maintained by a consortium of several European institutions, including EMS and FIZ, it comprises currently about more than 265,000 research articles from five centuries, and provides DML functions and APIs.

The **Encyclopedia of Mathematics(*) (EoM)** is an open access resource designed specifically for the mathematics community. With more than 8.000 entries, illuminating nearly 50.000 notions in mathematics, the EoM is the most up-to-date reference work in the field of mathematics. In cooperation with the EMS and its publishing house EMS Press it become freely availbe. The EMS monitors any changes to articles and has full scientific authority over alterations and deletions.

GAP(*) (Groups, Algorithms and Programming) is a computer algebra system for computational discrete algebra with particular emphasis on computational group theory. It is free and open-source under the GNU General Public Licence.

LaTeXML is a software, developed by NIST, which converts LaTeX documents to XML, which can be used for machine interpretation and publication on the web. It started in the context of the *Digital Library of Mathematical Functions* to convert LaTeXsources with sepcial semantic annotions to generate the DLMF-website[24].LaTeXML generated websites present the semantic annotations from the

[23] Contributions by MaRDI members are marked by an asterisk (*).
[24] https://dlmf.nist.gov

LaTeX source as popups and hyperlinks to display definitions and type information for individual mathematical operators as well as constraints, relations to other formulae, and references to the literature.

MathHub(*) is a portal developed by FAU for mathematical knowledge representations of different levels of formality and semantically enhanced documents that use the underlying representations for user adaptivity and interaction. Resources hosted on MathHub include the major theorem prover libraries (300.000+ theorems) in OMDoc/MMT, semantically annotated course materials, and a multilingual mathematical lexicon.

MathDataHub(*)[25] is a portal for managing collections mathematical objects. The definition, representation, and properties of the objects are specified formally, and from this MathDataHub derives Database Schema, APIs, User Interface, and search facilities. Resources hosted on MathDataHub include the Small Groups Library, various graph/maniplex data sets, and additive bases.

MathWebSearch(*) is a mathematical formula search engine developed by FAU and used by FIZ as part of zbMATH for the semantically querying mathematical documents with content MathML markup (e.g. generated from LaTeXvia LaTeXML). The system has been in active use in zbMATH and other mathematical information systems (e.g. arXiv.org since 2015).

The **MORwiki**(*) (Model Order Reduction Wiki) was initiated in the year 2013 by the MPI Magdeburg, and has, by 2019, collected sixty interactive wiki article pages written by more than fifty contributors. At the heart of the MORwiki are three main sections: Benchmarks, methods and software. The benchmarks are test problems by which different algorithms can be compared. The methods are summaries of the mathematical algorithms, and the software section collects their respective implementations.

The **MPI MIS' Eberhard Zeidler Library**(*) performs optical character recognition on the indices of its digitized books and linking the indexed terms with their respective references in the full-texts.

OpenML(*) is a collaboration platform through which scientists can automatically share, organize and discuss machine learning experiments, data, and algorithms. It's free to use and all available empirical data and metadata is under the CC-BY license.

The **OSCAR**(*) project develops a comprehensive open source computer algebra system for computations in algebra, geometry, and number theory. In particular, the emphasis is on supporting complex computations which require a high level of integration of tools from different mathematical areas.

polymake(*) is open source software for research in polyhedral geometry. It deals with polytopes, polyhedra and fans as well as simplicial complexes, matroids, graphs, tropical hypersurfaces, toric varieties and other objects. Two key design features are particularly relevant to the proposal: the extendible polymake type system is serialized and formalized (as a RELAX-NG XML schema); there is a built-in interface to the database **polyDB**.

pyMOR(*) is an open source software library for building model order reduction applications with the Python programming language. In the joint DFG project "pyMOR - Sustainable Software for Model Order Reduction" by WWU and MPI DCTS, tools and computing infrastructure are developed that enable cloud-based scientific computing with pyMOR and other software in the web browser, facilitating the exchange of experiments between researchers.

[25] https://data.mathhub.info

R is a language and environment for statistical computing, which is freely available under the GNU General Public License. R notebooks will be used to create a library of reproducible statistical analyses. Data from the OpenML platform will be analysed and connected to the notebooks through the API provided by the R package OpenML(*).

RADAR(*) (Research Data Repository) is a not-for-profit and discipline-agnostic research data repository, which guarantees the availability of published research data for at least 25 years. It provides a generic and interoperable metadata schema which can be complemented with discipline-specific information. RADAR helps implementing FAIR principles. The repository has a Core Trust Seal certification and is listed on re3data.

Singular(*) is a computer algebra system for polynomial computations, with special emphasis on commutative and non-commutative algebra, algebraic geometry, and singularity theory. It is free and open-source under the GNU General Public Licence.

The **Small Groups library**(*) provides access to descriptions of the groups of small order. Groups fundamentally capture the concept of symmetry; they are listed up to isomorphism. For instance, this includes all 423 164 062 groups of order at most 2000 (except 1024).

The **swMATH**(*) database, maintained by FIZ and ZIB, establishes a connection between scientific publications and mathematical software. It provides software metadata and semantic information such as links to the home pages, the Internet Archive, licensing terms, versions, MSC classifications, authors, as well as software usage and citations in publications.

Wikibase Wikibase is an open-source knowledge base software developed for the knowledge base Wikidata. The key features of Wikibase include its simple but powerful data model, versioning, and multi-language support. There are several programming interfaces to access Wikibase. The data model of a Wikibase instance is mapped to the *Resource Description Framework*, so that the database can also be queried via SPARQL. Besides Wikidata, Wikibase is mainly used in the scientific and cultural sectors.

zbMATH(*), edited by the EMS, FIZ, and the Heidelberg Academy of Sciences, is the world's most comprehensive and longest-running abstracting and reviewing service in mathematics, which covers the complete research literature since 1868 by the effort of currently more than 7,000 mathematicians worldwide. Currently, zbMATH is in the transition process from the traditional subscription model to an information system providing open services, data and API.

Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN.

References from MaRDI partners

[Bas+08] P. Bastian et al. "A Generic Grid Interface for Parallel and Adaptive Scientific Computing. Part II: Implementation and Tests in DUNE". In: *Computing* 82.2–3 (2008), pp. 121–138. DOI: 10.1007/s00607-008-0004-9.

[Bau+17] U. Baur et al. "Chapter 9: Comparison of Methods for Parametric Model Order Reduction of Time-Dependent Problems". In: *Model Reduction and Approximation*. Ed. by P. Benner et al. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017, pp. 377–407. DOI: 10.1137/1.9781611974829.ch9.

[Ben+17] P. Benner et al. “POD-DEIM for Efficient Reduction of a Dynamic 2D Catalytic Reactor Model”. In: *Computers & Chemical Engineering* (2017), pp. 777–784. DOI: 10.1016/j.compchemeng.2017.02.032.

[Ben+18] P. Benner et al. “Comparison of Model Order Reduction Methods for Optimal Sensor Placement for Thermo-Elastic Models”. In: *Eng. Optim.* 51.3 (2018), pp. 465–483. DOI: 10.1080/0305215X.2018.1469133.

[Ben+19] P. Benner et al. “Computing the Density of States for Optical Spectra of Molecules by Low-Rank and QTT Tensor Approximation”. In: *J. Comput. Phys.* 382 (2019), pp. 221–239. DOI: 10.1016/j.jcp.2019.01.011.

[BEO01] H. U. Besche, B. Eick, and E. A. O’Brien. “The groups of order at most 2000.” In: *Electron. Res. Announc. Am. Math. Soc.* 7 (2001), pp. 1–4.

[Bis+16] Bernd Bischl et al. “mlr: Machine Learning in R”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5938–5942.

[Bis+17] Bernd Bischl et al. “OpenML Benchmarking Suites”. In: *arXiv* (2017), arXiv–1708.

[Cas+17] Giuseppe Casalicchio et al. “OpenML: An R package to connect to the machine learning platform OpenML”. In: *Computational Statistics* 32.3 (2017), pp. 1–15. DOI: 10.1007/s00180-017-0742-2. URL: http://doi.acm.org/10.1007/s00180-017-0742-2.

[DKK18] B. Drees, A. Kraft, and T. Koprucki. “Reproducible and comprehensible research results through persistently linked and visualized numerical simulation data”. In: *Optical and Quantum Electronics* 50.2 (2018), p. 59. DOI: 10.1007/s11082-018-1327-1.

[EMNW11] G. Engeln-Müllges, K. Niederdrenk, and R. Wodicka. *Numerik-Algorithmen. Verfahren, Beispiele, Anwendungen.* 10th revised and extended ed. Berlin: Springer, 2011, pp. xxii + 755. ISBN: 978-3-642-13472-2/hbk; 978-3-642-13473-9/ebook.

[FDW19] Rina Foygel Barber, Mathias Drton, and Luca Weihs. *SEMID: Identifiability of Linear Structural Equation Models*. R package version 0.3.2. 2019. URL: https://CRAN.R-project.org/package=SEMID.

[Feh+16] J. Fehr et al. “Best Practices for Replicability, Reproducibility and Reusability of Computer-Based Experiments Exemplified by Model Reduction Software”. In: *AIMS Mathematics* 1.3 (2016), pp. 261–281. DOI: 10.3934/Math.2016.3.261.

[Feh+19] J. Fehr et al. *Sustainable Research Software Hand-Over*. e-prints 1909.09469. cs.GL. arXiv, 2019. URL: https://arxiv.org/abs/1909.09469.

[GHJ16] E. Gawrilow, S. Hampe, and M. Joswig. “The **polymake** XML file format”. In: *Mathematical software – ICMS 2016. 5th international conference*. Cham: Springer, 2016, pp. 403–410. DOI: 10.1007/978-3-319-42432-3_50.

[Gij+19] Pieter Gijsbers et al. “An open source AutoML benchmark”. In: *arXiv preprint arXiv:1907.00909* (2019).

[Gru+14] S. Grundel et al. “Model order reduction of differential algebraic equations arising from the simulation of gas transport networks”. In: *Progress in Differential-Algebraic Equations*. Differential-Algebraic Equations Forum. Springer Berlin Heidelberg, 2014, pp. 183–205. DOI: 10.1007/978-3-662-44926-4_9.

[HHP18] M. Hintermüller, M. Holler, and K. Papafitsoros. “A function space framework for structural total variation regularization with applications in inverse problems”. In: *Inverse Problems* 34.6 (2018), pp. 064002, 39. ISSN: 0266-5611. DOI: 10.1088/1361-6420/aab586.

[JJK18] C. Jordan, M. Joswig, and L. Kastner. “Parallel enumeration of triangulations”. In: *Electron. J. Combin.* 25.3 (2018), Paper 3.6, 27. ISSN: 1077-8926. URL: http://www.combinatorics.org/ojs/index.php/eljc/article/view/v25i3p6.

[Koh+17] M. Kohlhase et al. “Mathematical Models as Research Data via Flexiformal Theory Graphs”. In: *Intelligent Computer Mathematics: 10th International Conference, CICM*. Ed. by Herman Geuvers et al. Cham: Springer International Publishing, 2017, pp. 224–238. DOI: 10.1007/978-3-319-62075-6_16.

[Kop+17] T. Koprucki et al. “Mathematical models as research data in numerical simulation of opto-electronic devices”. In: *2017 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*. 2017, pp. 225–226. DOI: 10.1109/NUSOD.2017.8010073.

[Kop+18] T. Koprucki et al. “Model pathway diagrams for the representation of mathematical models”. In: *Optical and Quantum Electronics* 50 (2018), pp. 70/1–70/9. DOI: 10.1007/s11082-018-1321-7.

[KS14] M. Köhler and J. Saak. *FlexiBLAS - A flexible BLAS library with runtime exchangeable backends*. LAPACK Working Note 284. The Netlib, Jan. 2014. URL: http://www.netlib.org/lapack/lawnspdf/lawn284.pdf.

[KT16] T. Koprucki and K. Tabelow. “Mathematical Models: A Research Data Category?” In: *Mathematical Software – ICMS 2016: 5th International Conference*. Ed. by G.-M. Greuel et al. Cham: Springer International Publishing, 2016, pp. 423–428. DOI: 10.1007/978-3-319-42432-3_53.

[KTK16] T. Koprucki, K. Tabelow, and I. Kleinod. “Mathematical research data”. In: *PAMM* 16.1 (2016), pp. 959–960. DOI: 10.1002/pamm.201610458.

[Lan+19] Michel Lang et al. “mlr3: A modern object-oriented machine learning framework in R”. In: *Journal of Open Source Software* 4.44 (2019), p. 1903.

[LBS17] Michel Lang, Bernd Bischl, and Dirk Surmann. “batchtools: Tools for R to work on batch systems”. In: *Journal of Open Source Software* 2.10 (2017), p. 135.

[MDS20] Giovanni M. Marchetti, Mathias Drton, and Kayvan Sadeghi. *ggm: Graphical Markov Models with Mixed Graphs*. R package version 2.5. 2020. URL: https://CRAN.R-project.org/package=ggm.

[MOR19] MORwiki community. *MORwiki – The Model Order Reduction Wiki*. 2019. URL: http://www.modelreduction.org.

[MRS16] R. Milk, S. Rave, and F. Schindler. “pyMOR – Generic Algorithms and Interfaces for Model Order Reduction”. In: *SIAM Journal on Scientific Computing* 38.5 (2016), S194–S216. DOI: 10.1137/15M1026614.

[Paf17] A. Paffenholz. “polyDB: a database for polytopes and related objects.” English. In: *Algorithmic and experimental methods in algebra, geometry, and number theory*. Cham: Springer, 2017, pp. 533–547. DOI: 10.1007/978-3-319-70566-8_23.

[Sch17] M. Schubotz. *Augmenting Mathematical Formulae for More Effective Querying & Efficient Presentation*. Epubli Verlag, Berlin, 2017. ISBN: 9783745062083. DOI: 10.14279/depositonce-6034.

[SG] A. Schöbel and S. Gramsch. Own database query at publication database Scopus on 23 July 2019: For each year from 1988 to 2018, the total number of peer-reviewed publications is determined for the four subject areas excluding mathematics. The figure shows the percentage of peer-reviewed publications using mathematical concepts. A document is classified as “using mathematical concepts” if title or abstract contains a keyword from the following list: computer simulation, mathematical models, algorithm(s), artificial intelligence, neural networks, numerical methods, parameter estimation, simulation, probability, geometry, Monte Carlo, regression analysis, genetic algorithms, computational methods, data mining, numerical model. compared to the total number in each subject area excluding mathematics itself.

[SLM13] M. Schubotz, M. Leich, and V. Markl. “Querying Large Collections of Mathematical Publications: NTCIR10 Math Task”. In: *Querying large Collections of Mathematical Publications. Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10*. National Institute of Informatics (NII), 2013, pp. 667–674. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/03-NTCIR10-MATH-SchubotzM.pdf.

[ST19] M. Schubotz and O. Teschke. “Four decades of TEX at zbMATH.” English. In: *Eur. Math. Soc. Newsl.* 112 (2019), pp. 50–52. DOI: 10.4171/NEWS/112/15.

[SWM] *Mathematical Software – swMATH*. URL: http://swmath.org (visited on 09/07/2017).

[Tab+19] K. Tabelow et al. “hMRI – A toolbox for quantitative MRI in neuroscience and clinical research”. In: *NeuroImage* 194 (2019), pp. 191 –210. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.01.029.

[Van+13] Joaquin Vanschoren et al. “OpenML: Networked Science in Machine Learning”. In: *SIGKDD Explorations* 15.2 (2013), pp. 49–60. DOI: 10.1145/2641190.2641198. URL: http://doi.acm.org/10.1145/2641190.2641198.

[Yue+15] Y. Yue et al. “Application of Krylov-type parametric model order reduction in efficient uncertainty quantification of electro-thermal circuit models”. In: *Progress In Electromagnetics Research Symposium (PIERS 2015)*. 2015, pp. 379–384. DOI: 10.17617/2.2223025.

[ZBM] *zbMATH the first resource in mathematics*. URL: http://zbmath.org (visited on 09/29/2019).

General references

[Alk+16] M. Alkämper et al. “The DUNE-ALUGrid Module.” In: *Archive of Numerical Software* 4.1 (2016), pp. 1–28. DOI: 10.11588/ans.2016.1.23252.

[Ame19] American Mathematical Society. *MathSciNet – Mathematical Reviews*. 2019. URL: https://mathscinet.ams.org.

[AN] Arthur Asuncion and David Newman. *UCI machine learning repository*.

[And+99] E. Anderson et al. *LAPACK Users' Guide*. third. SIAM. Philadelphia, PA, 1999. DOI: 10.1137/1.9780898719604.

[Bak16] M. Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604 (2016), pp. 452–455. DOI: 10.1038/533452a.

[Bas+97] P. Bastian et al. “UG – A flexible software toolbox for solving partial differential equations”. In: *Computing and Visualization in Science* 1.1 (1997), pp. 27–40. DOI: 10.1007/s007910050003.

[BBS16] D. H. Bailey, J. M. Borwein, and V. Stodden. “Facilitating Reproducibility in Scientific Computing: Principles and Practice”. In: *Reproducibility*. John Wiley & Sons, Ltd, 2016. Chap. 9, pp. 205–231. DOI: 10.1002/9781118865064.ch9.

[Blo+12] T. Blochwitz et al. “Functional mockup interface 2.0: The standard for tool independent exchange of simulation models”. In: *Proceedings of the 9th International MODELICA Conference*. 076. Linköping University Electronic Press. 2012, pp. 173–184.

[Bri+19] A. Brinckman et al. “Computing environments for reproducibility: Capturing the “Whole Tale””. In: *Future Generation Computer Systems* 94 (2019), pp. 854–867. DOI: 10.1016/j.future.2017.12.029.

[CVD02] Y. Chahlaoui and P. Van Dooren. *A collection of benchmark examples for model reduction of linear time invariant dynamical systems*. Tech. rep. 2002–2. Available from www.slicot.org. SLICOT Working Note, 2002.

[Dem06] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *Journal of Machine learning research* 7 (2006), pp. 1–30.

[Eug] Manuel JA Eugster. “benchmark: Benchmark Experiments Toolbox, 2011”. In: ().

[Fri+17] M. Friedrich et al. “Angebotsplanung im öffentlichen Verkehr - planerische und algorithmische Lösungen”. In: *Heureka'17*. 2017.

[Gai+11] J.-M. Gaillourdet et al. “WoM: An Open Interactive Platform for Describing, Exploring, and Sharing Mathematical Models”. In: *Knowledge-Based and Intelligent Information and Engineering Systems*. Vol. 6884. Sept. 2011, pp. 126–135. DOI: 10.1007/978-3-642-23866-6_14.

[Gor+16] K. J. Gorgolewski et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments”. In: *Sci Data* 3 (2016), p. 160044. DOI: 10.1038/sdata.2016.44.

[Her] M. Heroux. *The TOMS Initiative and Policies for Replicated Computational Results (RCR)*. URL: https://toms.acm.org/replicated-computational-results.cfm (visited on 10/10/2019).

[Her+] R. Herzog et al., eds. *OPTPDE — A Collection of Problems in PDE-Constrained Optimization*. URL: http://www.optpde.net.

[Her+14] R. Herzog et al. “OPTPDE: A Collection of Problems in PDE-Constrained Optimization”. In: *Trends in PDE Constrained Optimization*. Ed. by G. Leugering et al. Vol. 165. International Series of Numerical Mathematics. Springer International Publishing, 2014, pp. 539–543. DOI: 10.1007/978-3-319-05083-6_34.

[Her19] M. A. Heroux. “Trust Me. QED.” In: *SIAM NEWS* 52 (6 2019), pp. 5–7. URL: https://sinews.siam.org/DetailsPage/trust-me-qed.

[Hot+05] Torsten Hothorn et al. “The design and analysis of benchmark experiments”. In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 675–699.

[HSE16] Benjamin Hofner, Matthias Schmid, and Lutz Edler. “Reproducible research in statistics: A review and guidelines for the Biometrical Journal”. In: *Biometrical Journal* 58.2 (2016), pp. 416–427. DOI: 10.1002/bimj.201500156. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201500156. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201500156.

[HW09] M. A. Heroux and J. M. Willenbring. “Barely Sufficient Software Engineering: 10 Practices to Improve Your CSE Software”. In: *Proceedings of the 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering*. 2009, pp. 15–21. DOI: 10.1109/SECSE.2009.5069157.

[IS18] D. Iglezakis and B. Schembera. “Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart – Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING”. In: *Das offene Bibliotheksjournal* 3 (2018).

[JK06] W. Joppich and M. Kürschner. “MpCCI—a tool for the simulation of coupled applications”. In: *Concurrency and Computation: Practice and Experience* 18.2 (2006), pp. 183–192. DOI: 10.1002/cpe.913.

[JP] The Julia Project. *The Julia Programming Language*. URL: https://julialang.org.

[KAG] *Kaggle*. URL: https://www.kaggle.com/.

[Kid+16] Mallory C. Kidwell et al. “Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency”. In: *PLOS Biology* 14.5 (May 2016), pp. 1–15. DOI: 10.1371/journal.pbio.1002456. URL: https://doi.org/10.1371/journal.pbio.1002456.

[Klu+16] T. Kluyver et al. “Jupyter Notebooks - A publishing format for reproducible computational workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. 2016, pp. 87–90. DOI: 10.3233/978-1-61499-649-1-87.

[KR05] J. G. Korvink and E. B. Rudnyi. “Oberwolfach Benchmark Collection”. In: *Dimension Reduction of Large-Scale Systems*. Ed. by P. Benner, D. C. Sorensen, and V. Mehrmann. Vol. 45. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, 2005, pp. 311–315. DOI: 10.1007/3-540-27909-1_11.

[Law+79] C. Lawson et al. “Basic linear algebra subprograms for FORTRAN usage”. In: *ACM Trans. Math. Software* 5 (1979), pp. 303–323. DOI: 10.1145/355841.355847.

[Mat] *MATLAB*. The MathWorks, Inc. URL: http://www.matlab.com.

[MCG] Office of Management, Budget Circulars, and Guidance. *OMB Circular A-110*. URL: https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A110/2cfr215-0.pdf (visited on 10/10/2019).

[Med19] MediaWiki. *MediaWiki — MediaWiki, The Free Wiki Engine*. 2019. URL: https://www.mediawiki.org.

[Mip] *MIPLIB 2017*. http://miplib.zib.de. 2018.

[OD] The Octave Developers. *GNU Octave*. URL: http://octave.org.

[Per19] J. M. Perkel. “Workflow systems turn raw data into scientific knowledge”. In: *Nature* 573.7772 (2019), pp. 149–150. DOI: 10.1038/d41586-019-02619-z.

[Ptn] *Collection of open source public transport networks by DFG Research Unit "FOR 2083: Integrated Planning For Public Transportation"*. 2018. URL: https://github.com/FOR2083/PublicTransportNetworks.

[R C20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: https://www.R-project.org/.

[Sae+19] Wouter Saelens et al. "A comparison of single-cell trajectory inference methods". In: *Nature biotechnology* 37.5 (2019), pp. 547–554.

[SBB13] V. Stodden, J. Borwein, and D. H. Bailey. "Setting the default to reproducible: Reproducibility in Computational and Experimental Mathematics". In: *SIAM News* 46 (2013), pp. 4–6.

[Sch+] A. Schiewe et al. *LinTim - Integrated Optimization in Public Transportation. Homepage*. http://lintim.math.uni-goettingen.de/. open source.

[SE19] Springer and European Mathematical Society. *Encyclopedia of Mathematics*. 2019. URL: https://www.encyclopediaofmath.org.

[SI19] B. Schembera and D. Iglezakis. "The Genesis of EngMeta – A Metadata Model for Research Data in Computational Engineering". In: *Metadata and Semantic Research*. 2019, pp. 127–132. DOI: 10.1007/978-3-030-14401-2_12.

[SRN18] P. Sojka, M. Ruzicka, and V. Novotný. "MIaS: Math-Aware Retrieval in Digital Mathematical Libraries". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM*. ACM, 2018, pp. 1923–1926. DOI: 10.1145/3269206.3269233.

[SS05] A. Schmidt and K. G. Siebert. *Design of Adaptive Finite Element Software – The Finite Element Toolbox ALBERTA*. Springer, 2005. URL: http://www.alberta-fem.de/.

[SS18] E. Schulze and M. Stein. "Simulation of Mixed Self-Assembled Monolayers on Gold: Effect of Terminal Alkyl Anchor Chain and Monolayer Composition". In: *The Journal of Physical Chemistry B* 122.31 (2018), pp. 7699–7710. DOI: 10.1021/acs.jpcb.8b05075.

[Str13] B. Stroustrup. *The C++ Programming Language*. 4th. Addison-Wesley Professional, 2013. ISBN: 9780321563842.

[VR95] G. Van Rossum. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995. URL: https://ir.cwi.nl/pub/5007.

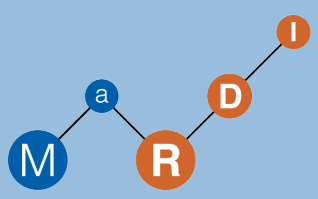
[Web+19] Lukas M Weber et al. "Essential guidelines for computational method benchmarking". In: *Genome biology* 20.1 (2019), p. 125.

[Wik04] Wikipedia contributors. *Wikipedia, The Free Encyclopedia*. 2004. URL: https://www.wikipedia.org.

[Wil+] Robert Wilson et al. *ATLAS of Finite Group Representations*. URL: http://brauer.maths.qmul.ac.uk/Atlas/v3/ (visited on 09/18/2019).

[Wil+16] M. 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.

[ZY15] Q. Zhang and A. Youssef. "Performance Evaluation and Optimization of Math-Similarity Search". In: *Intelligent Computer Mathematics - International Conference, CICM*. Vol. 9150. Lecture Notes in Computer Science. Springer, 2015, pp. 243–257. DOI: 10.1007/978-3-319-20615-8_16.


M
a
R
D
I