1 Binding letter of intent as advance notification or non-binding letter of intent

This is the binding letter of intent for the MaRDI 2019 NFDI consortium.

2 Formal details

Planned name of the consortium

Mathematical Research Data Initiative

Acronym of the planned consortium

MaRDI

Applicant institution

Weierstrass Institute for Applied Analysis and Stochastic (**WIAS**)
Mohrenstraße 39, 10117 Berlin
Director: Michael Hintermüller

Spokesperson

Michael Hintermüller, michael.hintermueller@wias-berlin.de. WIAS Berlin, HU Berlin

3 Objectives, work programme and research environment

- Research area of the proposed consortium: Mathematics 312-01

Mathematics investigates abstract mathematical objects and their properties or relations. Mathematical research generates new results by inferring from existing ones in a confirmable and reproducible way using the principle of logical reasoning. Therefore, sophisticated reuse is characteristic for mathematics. Mathematical results, however, are not represented by tables or numbers. Rather mathematical data is given by theoretical findings formulated and formalized in the language of mathematics. Findable and accessible mathematical objects and mathematical results are of paramount importance for developing the full creative power of mathematics.
Mathematics often has its origin in real application problems. This holds from the introduction of geometry 2500 years ago until today, where mathematical modeling, simulation, and optimization play important roles for understanding our complex world and for technological progress. Consequently, mathematical methods are nowadays not only used in the natural and engineering sciences, but also in economics, social, life, and environmental sciences, making it an essential key link between disciplines.
The growth of mathematics itself and its widespread use in other disciplines also leads to challenges: the complexity and the number of mathematical results have reached a volume that makes the handling by human experts difficult and the access for non-mathematicians often impossible. As an example, we mention the seminal paper by M. Hairer from 2014 on developing the theory of regularity structures, which despite the (typical mathematical) high compactness and complexity spans over 236 pages. Mathematics is created everywhere in science. Often this leads to a ”re-invention of the wheel’’ in different disciplines, without knowledge of each other or cross-referencing. The complexity required in today’s mathematics and its applications has reached a level which challenges confirmability.
Our approach 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. Hence, mathematics has to deal with all aspects of the FAIR principles.
The consortium will deal with the following issues, among others:

- **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 to be correct. However, with the advent of computer-based methods, complexity has increased beyond the grasp of a single researcher, and hence reproducibility of results has become an issue. Methods must be developed to handle these new challenges. Of particular importance is the link between mathematical results, software, and mathematical data (numeric, structured, symbolic, and metadata).
- **Confirmable Workflows**: Scientific results gain a lot of their value from the publication of the methods by which they have been obtained. In mathematics, formerly this role has been played by proofs, however since the advent of computer-aided mathematics, today the complete workflow needs to be documented, including types and versions of software used, program code, intermediate results, and others. The consortium aims to develop tools and standards for making this workflow easy and reliable.
- **Defining and Standardizing Mathematical Research Data**: Contrary to other disciplines, in mathematics, it is not a priori clear which types of data should be regarded as research data (RD). The consortium aims to develop a definition suited to the needs of working mathematicians along with the principles for handling them in a FAIR manner. The long-term goal is the scientific recognition of mathematical research data as a research accomplishment in its own right.

- **Development of Mathematical Services**: Standards are of no use if the burden to abide by them is placed entirely on the end user. Hence, services need to be developed that make the creation of FAIR research data easy and attractive.
- **Next-Generation Peer Review**: Correctness of computational methods, program code, and software cannot be ensured by traditional means of peer review. The consortium will work towards sustainability in this field by establishing standards and requirements for ensuring the correctness of these new kinds of research output.

Mathematical results are used in many disciplines and are the foundation of the cross-disciplinary methodologies mathematical modeling and simulation (MMS) as well as statistics and data science. The consortium aims to establish a common ground between the disciplines with respect to standards for models, software and data supporting the full data life cycle.
A central role is the interface to other disciplines. Mathematics can provide a precise formulation of the problems and the theory developed in other disciplines including the involved mathematical objects using mathematics as a common language. This covers models, parameters, data and numerical algorithms. On the other hand, it is important to link the mathematical objects 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 could provide a dictionary which helps to translate between the disciplines making results better findable, accessible and re-usable. This is not limited to models but also covers the involved data.

3.1 Task Areas

T1: Governance and Consortium Management

The 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 -flows, services, and interfaces to current research to reach a viable organizational and legal structure within the funding period.
Technically, the governance structure 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 Board in the consortium council (CC) by their co-spokespersons, where the strategic development of the objectives and structure of the consortium takes place. CC is the link to the members of the consortium. The third body in the structure is the general assembly GA constituted by the CC and all participants. 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 Scientific Advisory Board and a User Board. The boards will guarantee adherence to the FAIR principles within all developments in the consortium.

T2: Mathematical Methods and Computer-based Experiments

Like all branches of science, mathematics relies increasingly on computer-based experimentation and results. Beyond computer algebra and numerical mathematics, this includes many disciplines in applied and pure mathematics as well as data science. Computerized mathematics handles heterogeneous data types such as input/output data of mathematical software, the chosen method, implementation details, standardized test cases.
We will establish the FAIR principles for computer-based experiments in communities around specific mathematical subjects. In community pilots we will implement services and platforms and distill experiences into actionable guidelines and best practices. Furthermore, enabling

comparability of competing algorithms and discernibly tracking the subject's state-of-the-art is vital for scientific and efficient research conduct.
All sciences use mathematics, which implies a particular responsibility for us. The signing universities and institutes understand this foundational role as a call to lead by example.

T3: Cooperation with Other Disciplines

The overall goal of **T3** is to develop a standardized language for describing the underlying mathematical models allowing for further development of metadata describing and interlinking publications, problem instances, programs, tools, and experiments. The establishment of an easily accessible workflow that allows scientists from other disciplines to link their domain-specific context and semantics to the mathematical objects is of great importance. This is not limited to the mathematical models but also covers the involved data and approaches. For the confirmable interlinking of data with software we collaborate with **T2** and **T4** and, at least initially, with seven representative NFDI consortia spanning the sciences, engineering and humanities, see Section 3.3.
The collaboration will be done bottom-up starting with the partner NFDI consortia, and building on top of existing strong personal links and previous work for instance in inter- and transdisciplinary Clusters of Excellence or Transregional Collaborative Research Centers as nuclei. **MaRDI** will not only define and standardize interfaces between mathematics and other disciplines, it will have to personally include representatives. A concept for advising and scouting multidisciplinary scientists into our platform will be developed, as prototypes of this type of inclusion.

T4: Confirmable Workflows

Scientific results gain a lot of their value from the publication of the methods by which they have been obtained. In mathematics, this role is traditionally played by proofs. However, since the advent of computer-aided mathematics, more ingredients are to be taken into account. The whole value added chain needs to be documented, including types and versions of software used, program code, intermediate results, formal proofs, and more.
The overarching goal is the confirmability and reproducibility of mathematical reasoning in the computer age. This requires to develop and specify workflows for setting up and documenting computer experiments in mathematics. Models, software, and data (both input and output) need to be interlinked.

T5: Data Culture and Community Integration

Professional associations, such as DMV, EMS, GAMM, and GOR, as well as the Leibniz-Network Mathematical Modelling and Simulation (MMS) connect the mathematical-scientific community internally and with scientists of other disciplines. The **MaRDI** consortium will use these communication channels to establish standards (**T6**), services (**T7**), and workflows (**T4**) for the community. To reach the wider international mathematical community, the Max Planck Institutes (Leipzig, Magdeburg), that are frequently visited by world leading mathematicians, will serve as a platform. Likewise, FIZ will employ the large reviewer network of zbMATH. These professional bodies will translate FAIR to guidelines and recommendations for working mathematicians to achieve a shared understanding and acceptance. The professional associations will use their meetings and organize dedicated meetings to discuss these issues. We plan to combine the experience of MFO with the infrastructure and mindshare of the Software/Data Carpentry Movement, which has been teaching tens of thousands of researchers and engineers the basics of engineering using Research Data Carpentries.
By integrating the Wikimedia community, we spread our data culture beyond the professionally organized mathematicians and ensure that interested citizens can be a part of FAIR mathematical research data in Germany and around the globe. **MaRDI** supports the Wikimedia 2030 vision of an

open ecosystem of free knowledge. We will make trusted mathematical research data available to the public and eliminate technical barriers to access or contribute to mathematics.

T6: Standardizing Mathematical Objects, Models, Metadata, and Ontologies

Any large-scale FAIR Research Data infrastructure is necessarily built on standardized data formats, APIs, and query languages. The **MaRDI** consortium will drive the development of standards for mathematical data as a cross-cutting topic for the whole NFDI. It is important to realize that mathematical services (see **T7**) need access to the full semantics of mathematical Concepts, Objects, and Models (COMs) to be effective. We will build on existing standards (OpenMath and W3C content MathML for symbolic mathematical data -- i.e.formulae) and community best practices (LaTeX, presentation MathML, and Office Math for mathematical documents). **T6** will standardize licensing, legal frameworks, legal metadata, and licensing annotation workflows for the three categories of data above.

T7: Math Data Infrastructure and Services

In contrast to the **T6**, this task area acquires, semantifies, and indexes mathematical data. The goal of this task area is to deliver mathematical knowledge as a service. For one, this includes user-centered access to mathematical data in its various forms. Second, it provides machine-readable interfaces to third parties offering mathematical data via API requests or in bulk form. As a first step, we will homogenize and connect data from existing services such as formulae search engines, mathematical question answering system, mathematical plagiarism detection systems, computation engines, formulae recommender systems. In particular, data from arXiv, EuDML, MathOverflow, OEIS, Wikidata, swMATH, zbMATH, and many others will be accessible via the representation standards. Our commitment is to work together with the content providers to build cross-content services and to support them in adapting the metadata standards.

3.2 Use of existing infrastructures, tools and services

arXiv.org is a preprint portal for physics (45%), mathematics (30%), computer science (18%), etc. It currently comprises more than 1.5 Million articles, almost all of which come with LaTeX sources. The MPG Digital library is involved in the organization and FAU has converted "all" documents to HTML5 as a basis for semantics extraction and search. Together with zbMATH, this gives us a representative data set of mathematical full-texts that can be used for outreach, system evaluation, and benchmarking.

The **Digital Library of Mathematical Functions** 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.

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.

MathWebSearch is a mathematical formula search engine jointly developed by FAU and FIZ for the semantically querying mathematical documents with content MathML markup (e.g. generated from LaTeX via 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.

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.

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.

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.

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.

3.3 Interfaces to other proposed NFDI consortia

MaRDI and **NFDI4MSE** will collaborate with respect to the development and usage of representations of mathematical objects, in particular with respect to mathematical multi-scale modelling and as a solid foundation for the ontologies developed at NFDI4MSE. We will connect experimental and simulation data of real materials from NFDI4MSE with mathematical models and meta data in order to established confirmable workflows. Ideally, this will be done for use cases that build upon best practise examples from NFDI4MSE.

MaRDI and **PaHN-PaN** will collaborate with respect to data integration and annotation with meta data in order to make experimental data from PaHN-PaN accessible for data analysis with mathematical tools, such as machine learning or statistical data analysis. We will also work on showcases to use mathematical models and confirmable workflows for data based simulation in PaHN-PaN. The consortia will also collaborate with respect to building up a cross cutting platform for sustainable development of research software. One important point here is that sustainability of software never comes from the software itself but only through providing sustainable infrastructure for continuous long time deployment.

MaRDI and **FAIRmat** will collaborate on metadata (see **T6**). FAIRmat comprises the **NOMAD repository**, where computational data in materials science, resulting e.g. from density functional theory (DFT) are collected. Metadata for those data are already available, and could serve as a starting point for this activity. It is planned to harmonize metadata generation for Use Cases developed in both consortia (e.g., on battery materials) and the FAIRmat consortium invited **MaRDI** to hold the third NOMAD metadata workshop in 2021 jointly.

MaRDI, **NFDI4Ing** and **NFDI4Chem** will closely collaborate, including but not limited to reproducible science, the sharing of mathematical models, the generation and description of input data sets from experiments and measurements, and simulation software developed for analysis and quantifiable predictions. For example, Principal Investigators in the Cluster of Excellence EXC 2075 (Data-Integrated Simulation Science) are represented in these consortia, and they have already established a transdisciplinary open data and software hub within the Cluster and for the research community, that can serve as an initial prototype and best practice example.

MaRDI and **NFDI4Culture** will collaborate in the standardization of research data (**T6**). In contrast to most other disciplines research data are not restricted to array data, but involve complex objects. In mathematics these are purely ideal objects like elliptic curves, PDE-based models, or theorems/proofs. In the humanities these are cultural artefacts with all their material, conceptual, discourse, and reception historic properties. We will focus on symbolic object representations and ontology-based metadata approaches. With the cooperation we expect good coverage in these (research) data categories that can evolve into a cross-cutting standard applicable to all of NFDI. This cooperation will build on the well-established cooperation in the digital humanities at FAU.

MaRDI and **NFDI4RSE** will collaborate in the area of software carpentries and the curation of the swMATH database. **MaRDI** will contribute to the pool of carpentry instructors and develop course programs for mathematical research software engineering with a focus on confirmable workflows and data APIs. Details on the level of cooperation will be discussed at a later stage, as NFDI4RSE won't apply in 2019.

4 Cross-cutting topics

MaRDI has identified the following cross-cutting topics

- Governance models
- Knowledge transfer, collaboration, and synergies
- (Meta)data lifecycle and quality, standards and best practices
- Data APIs
- Software/data interaction, data reduction/compression, research software engineering
- Data workflows
- Federated data infrastructures and repositories
- Integrated services, ontologies, and federated discovery systems
- Legal and ethical framework enabling FAIR data through licenses, privacy, contracting
- Community integration, data culture, carpentry, training

In general **T3** is explicitly designed to stimulate knowledge transfer and synergies through collaborations with disciplines employing mathematical modeling and simulation or data science and their corresponding NFDI consortia.

MaRDI aims to contribute to the NFDI with respect to the following topics:

Standards for the representation mathematical research data and confirmable workflows. The quality of data and metadata is of utmost importance for mathematics, since research usually relies on involved derivations and computations where even minor inaccuracies can quickly propagate and lead to wrong results. Additionally, mathematical data often come along with high longevity - e.g., formulas, functions, or mathematical models will usually be reused for a very long, possibly even indefinite time. This requires an intense focus on data quality, provenance, lifecycle, and standardization. Mathematical data in this sense are frequently not “big” but rather “deep” within their description. **MaRDI** aims at making sustainable procedures for them available, including confirmable workflows for mathematical data (**T4**). **MaRDI** will also standardize mathematical data representations (**T6**) from a mathematical perspective. The dynamic cooperation with other NFDI consortia ensures that these standards are compatible and synergistic with all applications.

Next generation number formats. The ubiquitous employ of computers to tackle scientific problems is also pervasive in mathematics. This manifests in the design and implementation of algorithms, driving scientific computing, but also the sound development of (next generation) machine-readable number formats, utilized from office spreadsheets to supercomputer simulations.

Confirmable and Reproducible Workflows for Computer Experiments and Simulations. Hence, the interplay of software and data is essential for **MaRDI** and addressed specifically in **T2**. To make the research data resulting from computational experiments accessible to reuse by others than the creator performing the computations, and also to guarantee usefulness as input to other software packages (interoperability), the data needs to be enriched with semantic annotations and metadata (see **T6**), and the systems need to be equipped with data APIs (application programming interfaces) based on these. Such an approach also serves the purposes of data reduction and compression; e.g., for many simulations, it is hopeless to store all generated data. Instead, reusable conservation of the corresponding mathematical model, its discretization, and the entry points would serve much better for reproducibility. This will be a joint topic with many other consortia and include quite heterogeneous data: symbolic, concrete, and narrative data (aka ontologies, knowledge graphs, documents, etc.). Expertise from existing services within **MaRDI**

(e.g., the connection of software with publications) will be made open for reuse for other disciplines.

Federated repositories for heterogenous and data. The heterogeneous nature of the involved data requires viable federated infrastructures involving sophisticated data descriptions. As described above, solutions for mathematical data will usually be of interest for application areas. Specifically, mathematical data are typical for “long tail” research data. Repositories adapted to them would likely be able to serve the needs of other data of this kind.

Mathematical formula search. Services for mathematical data from **T7** like formula search will be employed which are also of immediate interest for reuse in cooperating areas.

Licensing and Legal Aspects. The research group Intellectual property rights (**IGR**) of FIZ supports **MaRDI**, as well as several other NFDI consortia w.r.t. legal aspects. Experts for licenses and Open Access/Open Science requirements in distributed systems support the mathematical community, which has been a driving force toward an accessible, efficient, fair, and transparent system of scientific information. Standards for licenses, privacy, contracting developed in **T6** in close interaction with the community represented in **MaRDI** will enable the reuse in other sciences according to the FAIR principles.

Data culture and training. Last, but not least, a cross-cutting concern is the training of the research community in the NFDI standards, data sets, services, and workflows. The **MaRDI** consortium will investigate setting up a **Research Data Carpentry (T5)** to leverage the great success of the **Software/Data** carpentries in training research communities. If this is successful, we will offer this platform to the NFDI as a whole.

The international mathematical research community is highly connected. The manifold international relations in the **MaRDI** consortium will be employed for not just the international impact and compatibility of German mathematics, but science in general through the propagation and recognition of research data infrastructures, standards, and services developed within the NFDI framework (**T5**).

MaRDI already initiated the cooperation with seven consortia: **FAIRmat**, **NFDI4MSE**, **NFDI4Culture**, **PaHN-PaN**, **NFDI4ING**, **NFDI4Chem,** and **NFDI4RSE** (see above) which usually derives from several of these aspects.