1st Conference on Research Data Infrastructure
Natural Sciences
https://doi.org/10.52825/CoRDI.v1i.397

Published: 07 Sept. 2023

MaRDI

Building Research Data Infrastructures for Mathematics and the Mathematical Sciences

Renita Danabalan[1] [https://orcid.org/0000-0003-3324-6448], Michael Hintermüller[1] [https://orcid.org/0000-0001-9471-2479], Thomas Koprucki[1] [https://orcid.org/0000-0001-6235-9412], and Karsten Tabelow [1] [https://orcid.org/0000-0003-1274-9951]

[1] WIAS Berlin, Germany

Abstract: MaRDI is building a research data infrastructure for mathematics and beyond based on semantic technologies (metadata, ontologies, knowledge graphs) and data repositories. Focusing on the algorithms, models and workflows, the MaRDI infrastructure will connect with other disciplines and NFDI consortia on data processing methods, solving real world problems and support mathematicians on research data management.

Keywords: Research data, mathematics, research data infrastructures, semantic technologies, repositories

1 Introduction

At the heart of many scientific discipline practices lies the processing and analysis of data collected to gain actual scientific insights and/or discoveries. In general, this step can be understood as a sequence of data transformations acting on input creating the output. The output can be used to answer a specific research question and to support research findings. The input and output, together with the data transformations in-between are parts of the factual material necessary to validate research findings, thereby constituting the research data related to specific question.

Next, we illustrate how the concept of data transformations can be applied to research in mathematics. We start with the field of scientific computing that is related to numerical data. For example, solving a linear system of equations $Ax = b$ can be seen as the transformation of the system matrix A and the vector b, via a solver, to the solution vector x. In this case, we can reinterpret the solution process, e.g. the Gaussian elimination, as a transformation of input to output. While the idea of data transformation may be less obvious in other areas of mathematics, it can still be applied; e.g. a computer algebra system, such as Mathematica, ‘transforms’ formulae to formulae. The difference being the data is non-numeric and more complex comprising of both symbolic and exact data.

The diversity of data in mathematics can be attributed to the research process where objects are invented and their relations and their properties are discovered. Data can range from numerical and tabulated to non-numerical like formulae, symbolic data, models and documents. More importantly, computations with these objects leads to algorithms and research software corresponding to the data transformations as introduced above. The study of their performance is essential to mathematical research and data processing pipelines of the other disciplines where a similar approach to data transformation can be taken; novel findings are acquired by processing and analysing of data.

Although input data might look different, the concept and requirements remain the same. The access to objects, the study of their properties and relations requires standardised data formats, data interoperability and application programming interfaces. With this in mind, the Mathematical Research Data Initiative (MaRDI) will develop a robust Mathematical Research Data Infrastructure that allows for findability and accessibility of objects, investigation of performance of algorithms and a search engine to identify solutions to mathematical problems [1]. This is supposed to support mathematical research fulfilling the needs for data management [2]. Moreover, MaRDI would help to solve real world problems by translating them into mathematical ones. Here, several questions arise, e.g. existence of a mathematical model, availability of solving algorithms, input data or model validity. Bridging the MaRDI infrastructure to disciplines outside of mathematics reduces the amount of time required for finding existing models, algorithms and solvers.

2 The MaRDI layer architecture

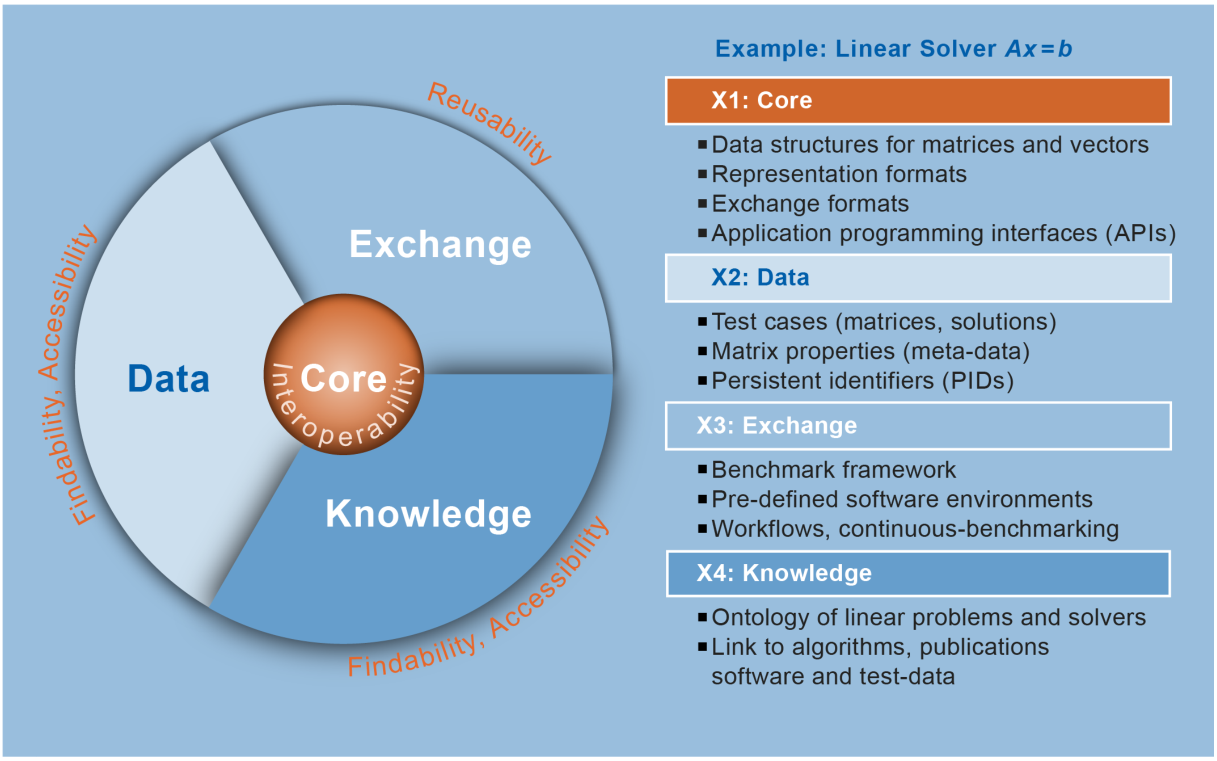
Reusability
Findability, Accessibility
Exchange
Data
Core
Interoperability
Knowledge
Findability, Accessibility
Example: Linear Solver $Ax=b$
X1: Core
▪ Data structures for matrices and vectors
▪ Representation formats
▪ Exchange formats
▪ Application programming interfaces (APIs)
X2: Data
▪ Test cases (matrices, solutions)
▪ Matrix properties (meta-data)
▪ Persistent identifiers (PIDs)
X3: Exchange
▪ Benchmark framework
▪ Pre-defined software environments
▪ Workflows, continuous-benchmarking
X4: Knowledge
▪ Ontology of linear problems and solvers
▪ Link to algorithms, publications software and test-data

Figure 1. MaRDI layer architecture and FAIR principles. The X1:Core layer enables the interoperability, the X2:Data and X4:Knowledge layer the findability and accessibilty and the X3:Exchange layer the reusability.

In order to achieve this goal, our layer architecture (Fig. 1) allows us to define the basic requirements of our infrastructure [1]. This benefits (a) method developers in running new algorithms on many test problems and also on special collections and (b) users in the generating performance data for selected algorithms or their implementations regarding various test problems from their application context. Additionally, 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.

Explaining the architecture using the example of the linear solver, the first goal of our X1:Core layer (Fig. 1) would to be develop standards for different types of matrices, e.g. full, banded or sparse matrices, which includes in-memory representation formats and file formats for input-output. Second, development of application programming interfaces (APIs) that would implement computations on matrices and vectors and to call the solver. To study the performance of algorithms, the user would run their solver on a set of test cases, consisting of matrices A, vectors b and solutions x. These are provided by X2: Data layer together with metadata schemas for the description of matrix properties (e.g., symmetric positive definite) and the solver (e.g., direct, iterative) as well as query functions for test problems with specific properties (e.g. symmetric, size). An example for such a database is the SuiteSparseCollection [3]. Our X3: Exchange layer, provides an environment that would allow the user to compare the performance, accuracy and efficiency of a solver used for specific matrices. Finally, bringing some structure and relation to the output of the exchange layer, the X4: Knowledge layer uses an ontology of linear problems and solvers to build a knowledge graph for linear problems by linking algorithms to publications, research software and test and performance data. The knowledge graph allows the user to find an appropriate solver for a specific problem or application. Moreover, the MaRDI layer architecture ensures the compliance with the FAIR principles, see Fig. 1.

3 MaRDI task areas

The work in MaRDI is divided into 3 pillars by data categories (Fig. 2): exact and symbolic data (task area (TA) 1:Computer Algebra), floating point data (TA2: Scientific Computing) and data with uncertainties (TA3: Statistics and Machine learning). Our fourth pillar (TA4: Interdisciplinary Mathematics) translates real world problems into mathematical ones through the use of mathematical models solvable by methods from TA1-TA3. Though data types in every TA differ and require individual solutions, the algorithms, mathematical models and workflows are cross-cutting. These cross-cutting topics allow for synergies between the MaRDI task areas and with other disciplines and NFDI consortia. The services, including semantic technologies (metadata, ontologies, knowledge graphs) and data repositories, developed by the task areas will be integrated into the MaRDI portal, our central access point developed by TA5. In parallel, MaRDI's TA6 will engage with and involve the community in the development and use of our services to promote a culture shift towards FAIR science.

Funding

This work was supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative".

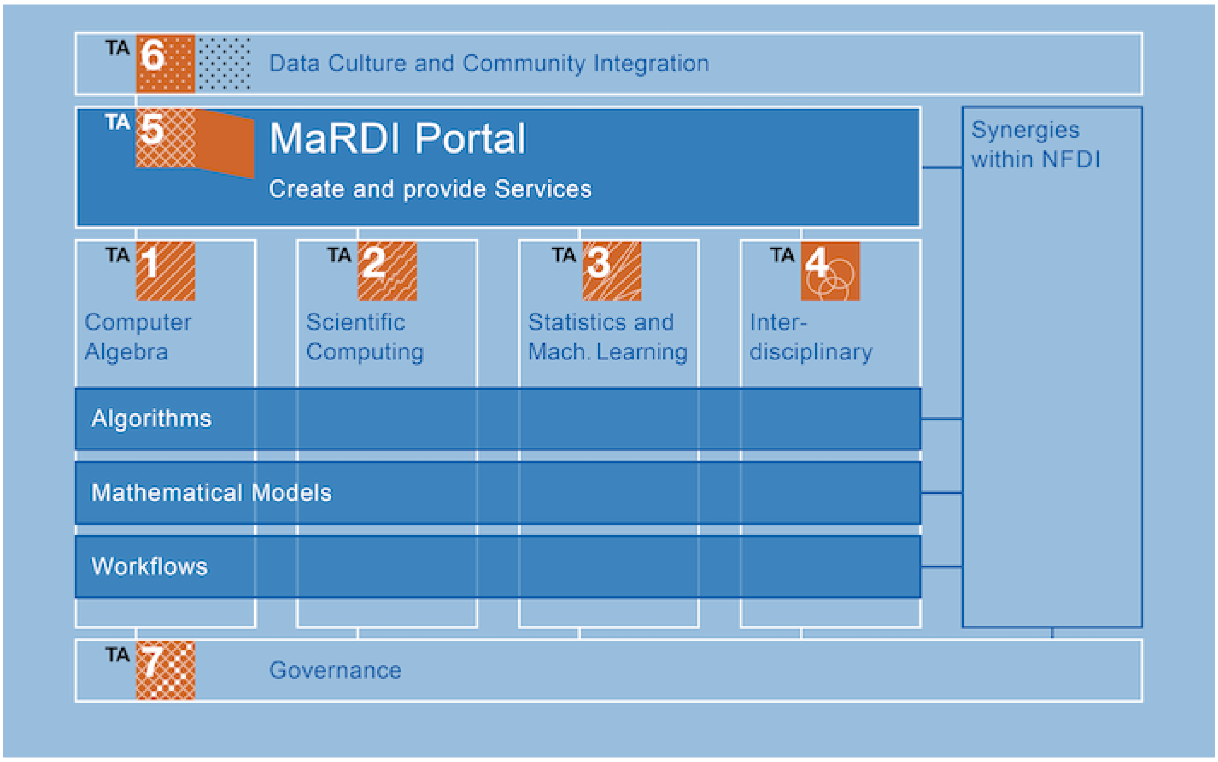

Figure 2. MaRDI organisation. Four task areas address different data categories in mathematics and interdisciplinary research. Task area 5 is dedicated to the MaRDI portal and TA6 deals with the community integration.

Acknowledgements

We thank Thomas Bender for creating the figures.

References

[1] The MaRDI consortium, *MaRDI: Mathematical Research Data Initiative Proposal*, May 2022. DOI: 10.5281/zenodo.6552436. [Online]. Available: https://doi.org/10.5281/zenodo.6552436.

[2] T. Boege, R. Fritze, C. Görgen, *et al.*, *Research-Data Management Planning in the German Mathematical Community*, 2022. arXiv: 2211.12071 [math.HO].

[3] T. A. Davis and Y. Hu, "The University of Florida Sparse Matrix Collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, Dec. 2011, ISSN: 0098-3500. DOI: 10.1145/2049662.2049663. [Online]. Available: https://doi.org/10.1145/2049662.2049663.