1st Conference on Research Data Infrastructure

Natural Sciences

https://doi.org/10.52825/CoRDI.v1i.323

Published: 07 Sept. 2023

MaRDIFlow: A Workflow Framework for Documentation and Integration of FAIR Computational Experiments

Pavan L. Veluvali[1] [https://orcid.org/0000-0001-8804-0338], Jan Heiland [1] [https://orcid.org/0000-0003-0228-8522], and Peter Benner [1] [https://orcid.org/0000-0003-3362-4103]

[1] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

Abstract: Numerical algorithms and computational tools are essential for managing and analyzing complex data processing tasks. With ever increasing availability of meta-data and parameter-driven simulations, the demand and the need for reliable and automated workflow frameworks to reproduce computational experiments has grown. In this work, we aim to develop a novel computational workflow framework, namely MaRDIFlow, that describes the abstraction of multi-layered workflow components. Herein, we plan to enable and implement scientific computing data FAIRness into actionable guidelines for FAIR computational experiments.

Keywords: FAIR, Computational Workflows, Reproducibility, MaRDIFlow

1 Introduction

Scientific computing has been a cross-disciplinary topic at the borders of applied mathematics, computational sciences and engineering (CSE), as well as other scientific domains involving numerical computations. Likewise, algorithms from numerical mathematics and data are the backbone of simulations in engineering problems, where, practically, different numerical methods are chained to design workflows.

Over the last two decades this has led to the accumulation of significant and frequent difficulties in maintaining solutions and in enabling collaborations within disciplines. We know that the ability to reproduce original research results is contingent on the availability of the original data and methods. As a result, in the past many years, efforts have been devoted towards the development of computational workflows for various scientific applications [1].

In general, a computational workflow is defined as a step-by-step description for accomplishing a scientific objective expressed in terms of tasks and their data dependencies [2]. The complex multi-step methods that are typically used for data collection, data analytics, predictive modeling, and simulation in turn lead to the development of new products. As a research data management (RDM) tool they can also be stored, retrieved for modifications, and subsequently reused in different scenarios with user-defined patterns. Currently, computational workflows offer graphical interfaces with high-level mechanisms for composition as well as traditional text-based programming interfaces.

However, a key challenge for computational scientists is building a framework for creating, maintaining, and accessing reusable workflows [3]. While tools and programming interfaces are important aspects in computational science and engineering, integrating them into a workflow framework can further express details about meta data and task dependencies, respectively. Nonetheless, motivated by demands of the mathematical community and other disciplines, MaRDI (Mathematical Research Data Initiative) [4], the consortial initiative of mathematical sciences, aims to set standards for the design of confirmable workflows.

In this regard, as a part of the MaRDI consortium on research data management in mathematical sciences [4], we present a novel computational framework, namely MaRDIFlow, that focuses on automation of abstracting meta-data embedded in an ontology of mathematical objects while negating the underlying execution and environment dependencies into multi-layered descriptions.

2 MaRDIFlow

The overall objective of our workflow framework is to provide a programming environment that simplifies the effort required by users or scientists to orchestrate a FAIR computational experiment. In MaRDIFlow, the workflow components are considered as abstract objects which are in turn described by their input to output behavior and as well as by their corresponding metadata. Through metadata and by matching the I/O interfaces, the objects are chained together to form a computational workflow. Herein, input stands for the parameter that sets up the current part of the workflow, and output denotes the final/intermediate result which is passed on to the next component. In this way, a workflow component can be described in a multi-layered fashion, namely via a mathematical/physical model, via a model that has been inferred from data, or via plain data, see Figure. 1.

For a systematic description of CSE components, we plan to incorporate models of different kind, code, and data equivalently and redundantly. One of the important benefits 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.

In order to present the description of individual workflow components, we produce show cases with different meta-data for redundant and reproducible mathematical models. As a minimum working example we incorporate a two-dimensional model for Cahn-Hilliard equation [5] that simulates the phase-separation of a binary A-B alloy. Moreover, the developed computational workflow framework adheres to FAIR principles [6], such that abstracted components are Findable, Accessible, Interoperable, and Reusable. In addition, going forward we plan to provide our RDM tool through electronic lab notebooks (ELNs) with minimal adjustments and user-friendly guidelines.

Lastly, we believe that the CSE workflow description presented here as a part of the MaRDI consortium serves a scientific tool for research data management in numerical mathematics.

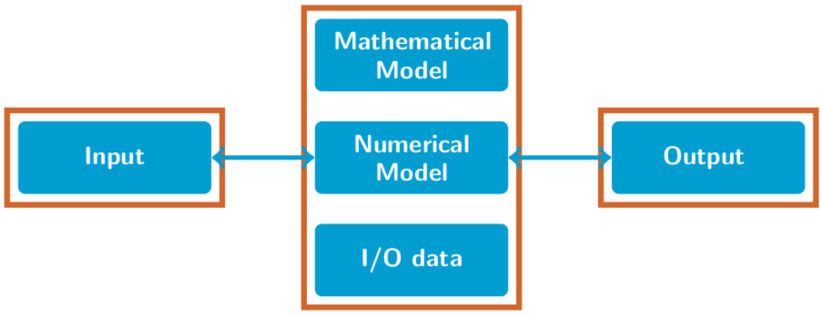

Figure 1. MaRDIFlow: A CSE workflow framework for documentation and integration of FAIR computational experiments

Data availability statement

Results presented in this work are apart of an ongoing investigation, however a working prototype of our workflow framework is available and documented at https://zenodo.org/record/78635

Competing interests

The authors declare that they have no competing interests.

Funding

Authors are supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative"

References

[1] D. Talia, "Workflow systems for science: Concepts and tools," *International Scholarly Research Notices*, vol. 2013, 2013.

[2] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, *et al.*, "FAIR Computational Workflows," *Data Intelligence*, vol. 2, no. 1-2, pp. 108–121, 2020.

[3] M. Wolf, J. Logan, K. Mehta, *et al.*, "Reusability first: Toward FAIR workflows," in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, IEEE, 2021, pp. 444–455.

[4] MaRDI. "Mathematic research data initiative." (2021), [Online]. Available: https://www.mardi4nfdi.de.

[5] J. W. Cahn and J. E. Hilliard, "Free energy of a nonuniform system. I: Interfacial free energy," *The Journal of chemical physics*, vol. 28, no. 2, pp. 258–267, 1958.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.