1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.395

Published: 07 Sept. 2023

Establishing the Research Data Management Container in NFDIxCS

Firas Al Laban[1][https://orcid.org/0000-0001-8072-9384], Jan Bernoth[1][https://orcid.org/0000-0002-4127-0053], Michael Goedicke[2], Ulrike Lucke[1][https://orcid.org/0000-0003-4049-8088], Michael Striewe[2][https://orcid.org/0000-0001-8866-6971] , Philipp Wieder[3][https://orcid.org/0000-0002-6992-1866], and Ramin Yahyapour[3][https://orcid.org/0000-0002-9057-4395]

[1] Universität Potsdam, Germany

[2] University of Duisburg-Essen, Germany

[3] Gesellschaft für Wissenschaftliche Datenverarbeitung mbH, Göttingen, Germany

Keywords: Research Data Management Container, NFDIxCS, Roadmap, Research Software, Software Sustainability, FAIR

NFDIxCS[1] is a consortium within the family of NFDI[2], which defines and establishes a research data management (RDM) infrastructure for Computer Science (CS). Based on a broad community process the various types of research data, their metadata and quality criteria are agreed upon in the community. The resulting research data, along with all associated supplementary information and context such as software, metadata, and the corresponding execution environment, are provided as an integral part of the overall infrastructure to meet the FAIR principles [1].

One key aspect of this infrastructure, which encapsulates connected research artifacts into a package object format, is the Research Data Management Container (RDMC) [2]. A central question addressed by the RDMC is how software artifacts can be sustainably preserved in the scientific system for a long time. The citation of software and data reveals a tension between the interests of scientists and publishers [3]; however, after archiving the software and providing a traceable citation, it is still not guaranteed that the addressed findings can be executed. The goal pursued with the RDMC is to create a container that links data, software, and execution environments so that the artifact can be used in the future and without much effort for Research Software Engineers under previously determined access regulations.

[1] https://www.nfdixcs.org

[2] https://www.nfdi.de, https://www.dfg.de/nfdi

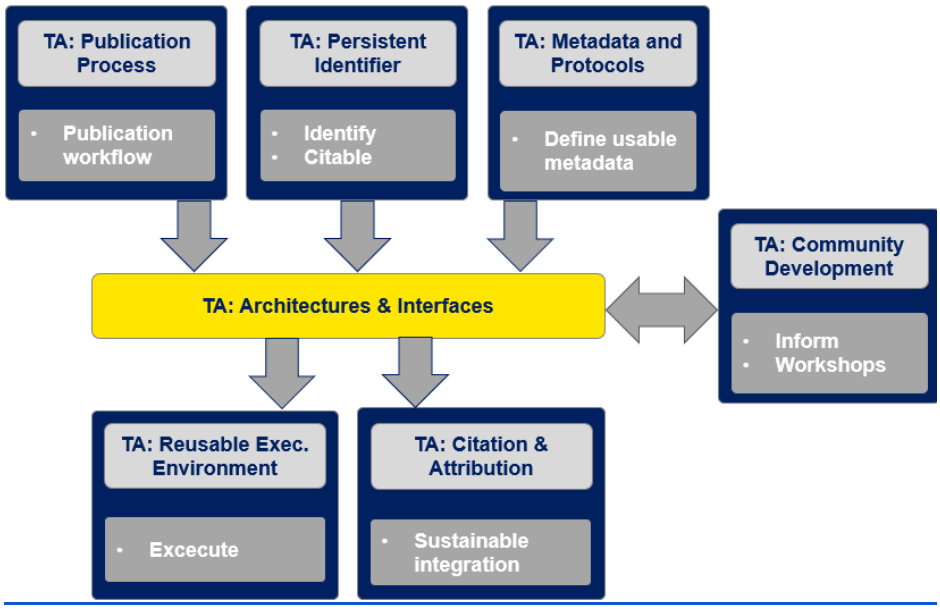

Figure 1. Role of the TA *Architectures & Interfaces* within the NFDIxCS Consortium.

The conception and development of the RDMC and its execution platform take place in the Task Area (TA) *Architectures & Interfaces,* which is closely linked to the other TAs (Figure 1).

Input from other TAs such as *Persistent Identifiers* and *Metadata and Protocols* is crucial for making containers findable and accessible in line with the FAIR principles. The TA *Publication Process* enables the RDMC to integrate automatic mechanisms in its workflow. Additionally, gathering requirements, conducting workshops for implementation, and community feedback are collected in the TA *Community Development* to integrate the user community into the RDMC development process. The deployment of the RDMC in an execution environment is addressed in the TA *Reusable Execution Environment*. The TA *Citation and Attribution* plays a vital role in integrating the RDMC into publication processes, ensuring proper credit and recognition for all involved parties.

The following roadmap (Figure 2) outlines stages for the development of the RDMC and its platform. Additionally, it demonstrates how external influences from the NFDIxCS consortium and beyond are considered.

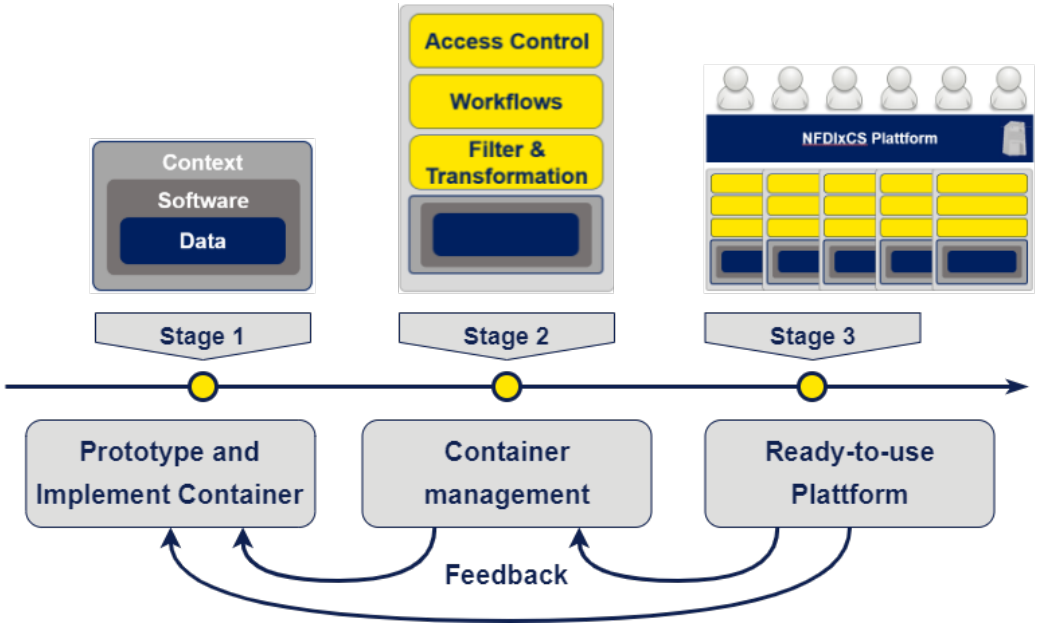

Figure 2. Implementing the RDMC.

Stage 1: Prototype and Implement Container

This stage is concerned with a prototype that realizes the basic requirements. The design and architecture of the RDMC prototype will be mostly agnostic regarding metadata formats, identifier standards, security mechanisms and similar. It will integrate the components that are needed to collect and organize data, and will provide the corresponding interfaces regardless of the data size or data type. One of the main characteristics of RDMC is to be flexible enough to deal with the content with no restricted rules and to use different metadata standards in parallel that may evolve within the NFDI community and beyond.

This stage will define data storage organization within the RDMC in such a way that metadata and related information about the execution environment is readable even if there is no access to the context. It will also support versioning of contents as well as of the related execution environments. Moreover, the RDMC will provide a mechanism to associate metadata and persistent identifiers both with the whole RDMC instance and specific data within that instance, e.g. for citation and attribution. The mechanism must be open to support different formats both for identifiers and metadata (such as Citation.cff [4] and DOI). This has also an international dimension, considering e.g. EOSC[3] for interoperable metadata.

Stage 2: Container management

Access control, workflows, and filter & transformation are specific features integrated into the RDMC. They control the access to and the processing of the data. This stage will focus on implementing modules that make these features usable based on an executable container platform like Docker[4]. Again, the challenge is to keep the definitions open and to not impose unnecessary restrictions, e.g. by making features only available within a specific platform implementation or a self-contained implementation for executable RDMCs.

One of the most important modules is the RDMC access layer that restricts access to the container's contents according to a user's permissions. It needs to allow for a fine-grained access control to make sure that permissions can be associated with any identifiable object

[3] https://eosc-portal.eu/

[4] https://www.docker.com/

within the RDMC, while metadata is universally accessible. Another module will implement support for workflows according to the use cases for RDMC, where the state of the workflow is encapsulated in the container. The latter is important to allow access rights to apply to specific phases of a workflow, such as access for reviewers during the publication process. Finally, filters and transformations as plugin modules may modify data access in specific use cases.

Stage 3: Ready-to-use Platform

By completing this stage, the RDMC will be established and be ready for the end users. The core activity in this stage is to create and launch a platform that allows to create, host, and access RDMCs. The main challenge is the creation of a (potentially federated) infrastructure, while most technical aspects of container management are already tackled in the previous stages. Technical management and monitoring features will also be considered in this stage to launch an evaluation process that collects evidence and factors regarding the best practices of the platform and RDMC in use. This ensures the sustainable development of the NFDIxCS infrastructure beyond the launching phase.

Conclusion

We have outlined the requirements, characteristics, and key concepts of the RDMC, which is the starting point of the NFDIxCS TA *Architectures & Interfaces* for the next 5 years. The RDMC encapsulates all related research data, contextual information including the software packages used, and the execution environment. This makes it much easier to realize the FAIR principles and the reproducibility of results for research data in Computer Science. We also envisage applications outside CS and will collaborate closely with other NFDI consortia.

Disclaimer

Parts of this text could be generated or rephrased by ChatGPT, DeepL Write, LanguageTool, and Google Docs spell checking, but were carefully checked and revised by the authors.

References

1. C. Engelhardt, How to be FAIR with your data. Göttingen: Göttingen University Press, 2022. DOI: https://doi.org/10.17875/gup2022-1915.
2. M. Goedicke and U. Lucke, “Research Data Management in Computer Science - NFDIxCS Approach,” 2022. In: Demmler, D., Krupka, D. & Federrath, H. (Hrsg.), INFORMATIK 2022. Gesellschaft für Informatik, Bonn. (S. 1317-1328). DOI: 10.18420/inf2022_112.
3. M. Harrison, “Open Science, Publications and Code”, In: OPEN SCIENCE EUROPEAN CONFERENCE, Proceedings of the Paris Open Science European Conference: OSEC 2022. Paris, Marseille, France: OpenEdition Press, 2022, ISBN: 9791036545627, pp. 183-187. DOI: 10.4000/books.oep.15829.
4. S. Druskat, J. H. Spaaks, N. Chue Hong, R. Haines, and J. Baker, “Citation File Format (CFF) - Specifications,” 2021. DOI: 10.5281/zenodo.4813122