1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.355

Published: 07 Sept. 2023

Towards a Research Data Commons in the German National Research Data Infrastructure NFDI : Vision, Governance, Architecture

Michael Diepenbroek[1][https://orcid.org/0000-0003-3096-6829], Ivaylo Kostadinov[1][https://orcid.org/0000-0003-4476-6764], Bernhard Seeger[2][https://orcid.org/0000-0002-9362-153X], Frank Oliver Glöckner[3,4][https://orcid.org/0000-0001-8528-9023], Marius Dieckmann[5][https://orcid.org/0000-0001-5130-546X], Alexander Goesmann[5][https://orcid.org/0000-0002-7086-2568], Barbara Ebert[1][https://orcid.org/0000-0003-3328-6693], Sonja Schimmler[6,7][https://orcid.org/0000-0002-8786-7250], York Sure-Vetter[8][https://orcid.org/0000-0002-4522-1099]

[1] GFBio - German Federation for Biological Data, Germany

[2] University of Marburg, Germany

[3] MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany

[4] AWI - Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany

[5] Justus Liebig University Giessen, Germany

[6] Fraunhofer FOKUS, Germany

[7] Technical University of Berlin, Germany

[8] NFDI - National Research Data Infrastructure & KIT - Karlsruhe Institute of Technology, Germany

Abstract.

The concept of a "Research Data Commons" (RDC) established itself as an infrastructure ecosystem for science based on open standards and federated resources to facilitate the sharing of research data and services. The consortia of the German National Research Data Infrastructure (NFDI) have identified the collaborative provisioning of resources and services to be of key importance for a functioning and efficient RDC and are leveraging different corresponding measures to establish a sustainable concept in line with international developments.

Keywords: Research Data Commons, Cloud Infrastructure, NFDI, Germany

The concept of a "Research Data Commons" (RDC) established itself as an infrastructure ecosystem for science based on open standards and federated resources to facilitate the sharing of research data and services. The “commons” principle is based on the belief that resources are best used when they are managed in a collaborative and participatory manner [1].

The German National Research Data Infrastructure (NFDI) [2], with 27 consortia [3], identified cross-cutting topics [4] and initiated a number of sections [5] covering and fostering these topics including an RDC.

Internationally, the concept of a "Research Data Commons" has become increasingly important for the development of science infrastructures. Examples include the Australian

Research Data Commons (ARDC) [6], the US National Cancer Institute Research Data Commons (NCI RDC) [7], and, as an important part of the European research and innovation strategy, the European Open Science Cloud (EOSC) [8]. The Global Open Research Commons Interest Group (GORC IG) [9] of the Research Data Alliance (RDA), finally, assembles existing initiatives and heads at more convergence and networking between the various developments.

All of these initiatives aim to provide a unified, open and trusted digital infrastructure for the storage, sharing and reuse of research data. This should enable researchers to integrate data from different sources and thus gain efficiency and new scientific knowledge. In EOSC in particular, structures have already been formed to deal with comparable issues of harmonization of data infrastructures, including implementation projects and working groups that negotiate agreements on relevant fundamental topics (e.g. EOSC Interoperability Framework [10]). The NFDI relates to this central European infrastructure project on several levels: as an overall organization, and via both the consortia and the sections. It is of utmost importance for the development of the NFDI RDC to follow the EOSC principles [11] and to incorporate relevant results from the EOSC technical standardization groups.

The RDC concept is also in line with the goals of the International Data Spaces Association (IDSA) [12], which promotes a virtual data space leveraging existing standards and technologies, as well as governance models. In addition, IDSA supports data sovereignty as a crucial design aspect and proposes a corresponding Reference Architecture Model. The related project FAIR Data Spaces (FAIR DS) [13] aims at the practical concretization of a reference architecture and provides various demonstrators in close cooperation with NFDI. Several NFDI consortia contribute to FAIR DS.

Further, with the beginning of 2023 cross-cutting topics are supported by a new consortium, Base4NFDI [14]. The consortium aims at organizing and evaluating the development of basic services, available at scale, by well-structured workflows embedded into the NFDI governance. Base4NFDI is supposed to support infrastructural supply for potentially all consortia, whereby competing developments and incompatible solutions should be avoided.

The NFDI consortia bring with them a variety of heterogeneous and distributed information infrastructures that, today, are networked only to a limited extent. To establish cross-domain basic services, harmonization within the community of users and providers is indispensable. In doing so, the NFDI builds on decades of experience of its member institutions with the provision of central services. Often, these services are well-tested in single domains, but not harmonized across domains. A central requirement for the NFDI sections is to identify components for common use and to develop proposals on how these can be technically structured and implemented, and which prerequisites are necessary for networking. The sections thereby follow the guiding principle of the NFDI as a whole: existing data and services should be reused as much as possible and intelligently connected.

A specific goal is the NFDI RDC, fostered by the section Common Infrastructures [15]. The current RDC concept is essentially based on a multi-cloud infrastructure that can be used by the consortia and their partners to pool and share data, software, compute resources, and services. Hybrid solutions and edge components are not excluded. The RDC concept envisages various shared cloud services like an application layer with access to high-performance computing, collaborative workspaces, a federated framework for (meta)data integration, persistent identifier, terminology services, data management planning and long-term archiving services as well as a software marketplace (Fig. 1).

Security of data and services is seen as a further important topic. The RDC will be equipped with a common authentication and authorization infrastructure (AAI) and certification services (zero trust). Particularly noteworthy are current works focusing on the development of RDC components performed by the consortia NFDI4Biodiversity [16], NFDI4Microbiota [17],

NFDI4DataScience [18], and NFDI4Earth [19]. The first two pursue an RDC approach as a central concept for their infrastructure development strategy, in particular a distributed core storage infrastructure [20]. In addition, NFDI4Biodiversity is working on effective ways to integrate data and metadata from diverse data providers.

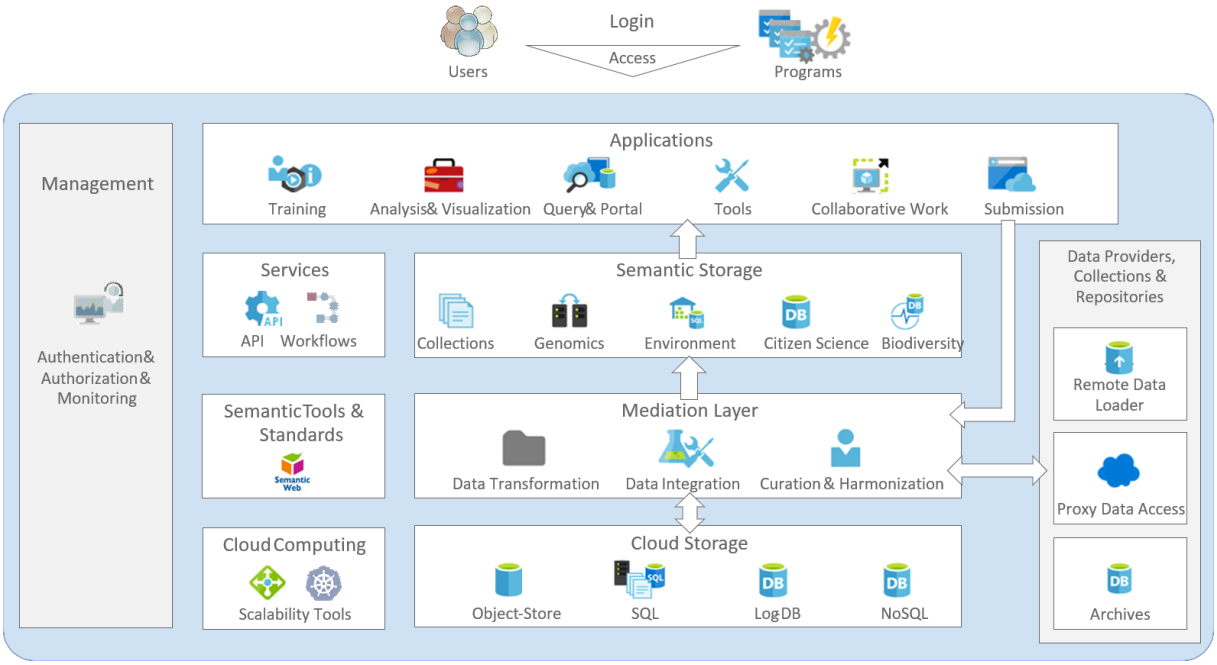

Figure 1. Overview of the RDC architecture, envisioned by the NFDI4Biodiversity and NFDI4Microbiota consortia. Realistically, a final architecture for the entire NFDI will amount to a networking of RDCs resembling a data mesh [21].

Outlook

The NFDI RDC is a holistic infrastructure concept for managing scientific data and services. The RDC implies a federated governance as a foundation of a participative bridge between data and service providers and consumers. It is expected to have a significant effect on the openness and quality of data and suggests a substantial increase in efficiency in the use of existing data. Overall, this provides strong incentives for potential users of NFDI offerings.

The first major challenge is embedding a national RDC into the national and international scientific services landscape, especially with regard to RDC developments in other countries. Within NFDI, in particular in the light of the basic services developments, intensive discussions and negotiations are needed to identify communalities and eventually agree on a common RDC vision and architecture. In view of the prerequisites in the various consortia and general IT developments, an agile development process is expected.

A further challenge is to organize the necessary capabilities and capacities, in particular for building up federated RDC components. Important aspects to address in this respect are compliance and sustainability of supplied services.

Data availability statement

The submission is not based on data.

Underlying and related material

None.

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was written in the context of the work of the association German National Research Data Infrastructure (NFDI), an initiative of the Joint Science Minister Conference of the Federal Republic of Germany and the 16 federal states. The work of the NFDI consortia is funded by the German Research Foundation DFG: NFDI4Biodiversity (grant number 442032008), NFDI4Microbiota (grant number 460129525), NFDI4DataScience (grant number 460234259), NFDI4Earth (grant number 460036893).

Acknowledgement

We thank the dedicated staff and partners who help to shape and support the structures and results presented here.

References

1. Grossman, R.L. Ten lessons for data sharing with a data commons. Sci Data 10, 120 (2023). https://doi.org/10.1038/s41597-023-02029-x
2. Nationale Forschungsdaten Infrastruktur (NFDI). https://www.nfdi.de/?lang=en (2023-04-24)
3. Nationale Forschungsdaten Infrastruktur (NFDI) - Consortia. https://www.nfdi.de/consortia/?lang=en (2023-04-24)
4. Glöckner, Frank Oliver, Diepenbroek, Michael, Felden, Janine, Overmann, Jörg, Bonn, Aletta, Gemeinholzer, Birgit, Güntsch, Anton, König-Ries, Birgitta, Seeger, Bernhard, Pollex-Krüger, Annette, Fluck, Juliane, Pigeot, Iris, Kirsten Toralf, Mühlhaus, Timo, Wolf, Christof, Heinrich, Uwe, Steinbeck, Christoph, Koepler, Oliver, Stegle, Oliver, Weimann, Joachim; Schörner-Sadenius, Thomas; Gutt, Christian; Stahl, Florian; Wagemann, Kurt; Schrade, Torsten; Schmitt, Robert; Eberl, Chris; Gauterin, Frank; Schultz, Martin; Bernard, Lars. (2019). Berlin Declaration on NFDI Cross-Cutting Topics (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3457213
5. Nationale Forschungsdaten Infrastruktur (NFDI) - Sections. https://www.nfdi.de/sections/?lang=en (2023-04-24)
6. Barker, M., Wilkinson, R. & Treloar, A. The Australian Research Data Commons. Data science journal 18 (2019). https://doi.org/10.5334/dsj-2019-044
7. NCDI Cancer Research Data Commons. https://datascience.cancer.gov/data-commons (2023-04-24)
8. European Open Science Cloud (EOSC). https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud (2023-04-24)
9. Global Open Research Commons IG. https://www.rd-alliance.org/groups/global-open-research-commons-ig (2023-04-24)

10. European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., et al., EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, https://doi.org/10.2777/620649
11. EOSC Declaration (2017). https://eosc-portal.eu/sites/default/files/eosc_declaration.pdf (2023-04-24)
12. International Data Spaces Association (IDSA). https://internationaldataspaces.org/ (2023-04-24)
13. FAIR Data Spaces. https://www.nfdi.de/fair-data-spaces/?lang=en (2023-04-24)
14. Basic Services for NFDI. https://base4nfdi.de/ (2023-04-24)
15. Nationale Forschungsdaten Infrastruktur (NFDI) - Section Common Infrastructures. https://www.nfdi.de/section-infra/?lang=en (2023-04-24)
16. NFDI4Biodiversity. https://www.nfdi4biodiversity.org/de/ (2023-04-24)
17. NFDI4Microbiota. https://nfdi4microbiota.de/ (2023-04-24)
18. NFDI4Datascience. https://www.nfdi4datascience.de/ (2023-04-24)
19. NFDI4Earth. https://www.nfdi4earth.de/ (2023-04-24)
20. ARUNA. https://aruna-storage.org (2023-04-25)
21. Dehghani Z (2020) Data Mesh Principles and Logical Architecture - https://martinfowler.com/articles/data-mesh-principles.html (2023-04-24)