DataPLANT Services Design – Considerations Towards a Common NFDI Landscape

Dirk von Suchodoletz [1], Marcel Tschöpe [1], Jonathan Bauer [1], Kevin Schneider [2], Timo Mühlhaus [2]

[1] Computer Center, University of Freiburg, Germany;
[2] Computational Systems Biology, RPTU Kaiserslautern-Landau, Germany

Keywords: DataPLANT, cloud oriented service infrastructure, PLANT DataHUB, NFDI, data publication

1 Motivation

Many Research Data Management (RDM) services are not limited to a single scientific domain or facility. Since no individual research institution can fully support all aspects of RDM with the required depth and domain-specific requirements for every discipline, collaboration and shared services are both logical and necessary. This approach aligns with the broader objective of One NFDI[1] within the National Research Data Infrastructure (NFDI). The NFDI aims to create a cross-disciplinary RDM ecosystem that meets the specific needs of research groups and disciplines, while also enabling the (re)use of data across traditional boundaries as well as the deployment of AI approaches. In line with this vision, one of the core objectives and contributions of the DataPLANT consortium (Martins-Rodrigues et al. 2021) is to provide tools and services that can be shared with other consortia and deployed across flexible backend infrastructures. At the heart of these technical services is the PLANT DataHUB.

1 Refers to the still informal but increasingly widespread concept and ultimate goal of a cross-disciplinary and cross-institutional approach to joint undertakings in research data management.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23953 (CC BY-SA 4.0).

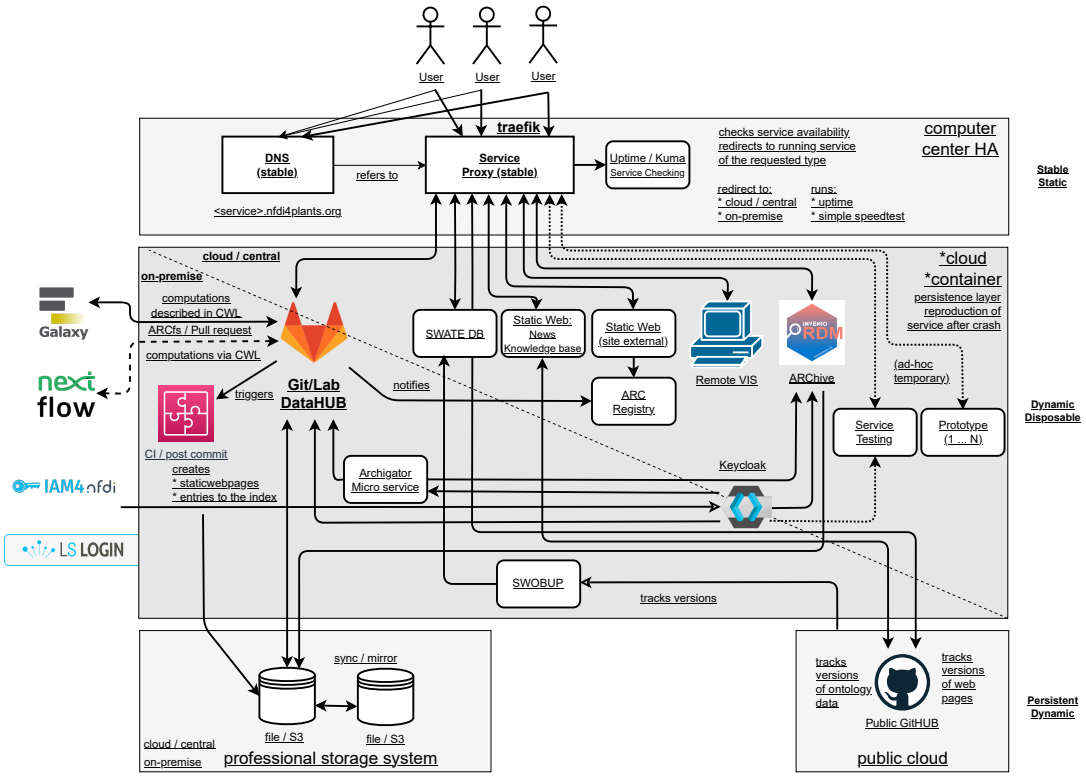

Figure 1: DataPLANT service infrastructure (Bauer, Tschöpe, and Suchodoletz 2025).

In addition to core functions, the DataHUB offers Continuous Integration (CI) pipeline templates, automated quality assurance, and other interaction services to foster DataPLANT's open participation and contribution model (see Figure 1). This model encompasses all relevant assets, including Annotated Research Context (ARC)[2], raw data, metadata templates, ontologies, code, and workflow descriptions. While originally geared towards plant scientists, both the ARC concept and the use of software development-based approaches to RDM, along with the associated services, are now being adopted by other consortia (e.g. FAIRagro, BIOIMAGE) and researchers across a wide range of disciplines. The backend services are designed as flexible, cloud-based microservices, supporting both on-premises installations and future integration with a shared NFDI infrastructure.[3] The PLANT DataHUB thus provides a range of RDM workflows to support data scientists throughout different stages of the research data lifecycle, from development to the publication of results. The bottom-up, community-driven approach requires collaboration among multiple stakeholders. In building these services, we follow design principles that offer high-level guidance and promote the creation of sustainable and maintainable applications. At DataPLANT, tool development is driven by community needs and facilitated through direct communication between researchers and developers. This process follows

2 See https://arc-rdm.org/; *Visited on May 22, 2025,* and Weil et al. (2023) for further background on the ARC concept, implementation and usage.

3 Multicloud proposal to the NFDI Section "Common Infrastructure", https://doi.org/10.5281/zenodo.6510971; *Visited on May 22, 2025.*

an incremental and iterative approach, supported by community contributions and issue tracking via the project repository.[4]

2 Service landscape & Design considerations

Over the past four years, the DataPLANT team has developed a suite of tools and microservices to extend software frameworks like GitLab for the PLANT DataHUB (Suchodoletz et al. 2023) or InvenioRDM for the "ARChive"[5] to enhance the digital service ecosystem of plant scientists (Suchodoletz, Bauer, and Tschöpe 2023). These core components, which focus on data management, versioning, sharing and publishing, are designed as portable modules that can be integrated into a broader base infrastructure. Key design priorities include beside user-friendliness, security, re-deployability, and fast recovery in case of system failures or infrastructure disruptions.

All DataPLANT core components are executed in Docker, while the configuration of these services is stored on the project's public code repository to separate development from data management and for security reasons. We aim to an "Infrastructure-as-code" principle for reproducibility and transparency. As soon as changes are made to the configuration in the repository, they are propagated to the core components. This process builds upon CI jobs, which are provided on so-called runners. These recipes for building software are used to distribute the configuration. This further supports the idea of a code repository as a core component of the science gateway – both for managing infrastructure components and for the actual DataHUB holding the scientific data as ARCs.

The process of assessing quality parameters of an ARC is further referred to as validation of the ARC against a "validation package" (VP), which is an arbitrary set of validation cases defined by e.g. a PI or research project lead that the ARC must pass to qualify as valid in regard to the VP. While a basic set of VPs is implemented, they can be significantly and individually extended for each project to accommodate different use cases and can be automatically triggered through the built-in CI infrastructure. The entire CI pipeline is organized in such a way that allows contributors to develop a VP of their own, and thus extend the ARC integration with yet other systems.

To support scientists in publishing their work, the in-house development Archigator acts as a connector and authentication microservice between the DataHUB and ARChive as well as other publication platforms for data and results. The integration of other repository

4 Everyone can contribute to developments at the project's GitHUB space, https://github.com/nfdi4plants *(Visited on May 22, 2025).* Feature requests and improvements are discussed openly on that project space or during bi-weekly data steward meetings. When necessary, final decisions are made in cross–task area speaker meetings, with support from the Scientific Advisory Board.

5 DataPLANT themed version of InvenioRDM, https://archive.nfdi4plants.org *(visited on May 22, 2025),* for data publication.

backends and workflow orchestration systems, such as Galaxy and nf-core, can be easily generalized by building on the foundational design of the Archigator tool. Archigator then serves as a data provider for these backends, utilizing their individual APIs. A virtual filesystem based on PyFilesystem called ARCfs has been created to allow for the integration of filesystem based storage in e.g. Galaxy, so that ARCs can be seamlessly integrated into such platforms (Bauer, Tschöpe, Suchodoletz, et al. 2024). This will allow scientists to connect their data in a fully integrated way throughout the data lifecycle based on GitLab. Here, we envision and strongly promote GitLab as a DataHUB to serve as a future representation of a central platform in the NFDI context (Suchodoletz et al. 2024).

In the case of Archigator and the publication of the scientists' results in i.e. ARChive, these CI processes help to assess the presence of mandatory publication-related metadata, quantify and publish the results. In addition to the core components, the in-house developments Swobup and SWATE (Mühlhaus et al. 2022), implemented as part of DataPLANT, support plant scientists in the provision and management of ontologies. These have also been developed with the principles of reusability and rapid recovery in mind. For reproducibility and provenance tracking all ontology data are published on the GitHub project space to engage the plant community. This data is retrieved by Swobup from the public repositories and integrated into the DataPLANT infrastructure as soon as changes are made to the public repositories and also when the Swobup and SWATE services are restored.[6]

All these tools and (micro-)services were programmed with on-premise operation in mind. Several communities within DataPLANT expressed the need to deploy their individual instances of the services (in particular the DataHUB). Partner institutions[7] have started to deploy their own DataHUB instances. Most of the core components have been developed as Docker containers or integrated directly into custom Docker images. We provide Docker Compose templates to configure the images for the various infrastructures available on site. The requirements for such on-premises deployments are: a virtualisation or container platform, a fast scratch space, and high-capacity storage such as NFS shares or, ideally, object storage. Although the basic requirements do not vary between institutions, the setups depend on the locally available infrastructures. The DataPLANT infrastructure team provides technical guidance to help local administrators create the required setup in terms of performance and data safety. In general, we encourage the use of a dynamic-yet-disposable model using cloud-based infrastructures (Bauer, Tschöpe, Schnürle, et al. 2024).

6 All tools, services, container descriptions are available from the project's public GitHUB referenced above.

7 At the beginning of 2025: Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Heinrich Heine Universität Düsseldorf, Eberhard-Karls Universität Tübingen, Universität zu Köln (planned).

3 Service Integration

The "glue that holds it all together" is a reverse proxy: a single entry point gateway running on a redundant virtual machine infrastructure (Bauer, Tschöpe, and Suchodoletz 2025). The proxy automatically manages Transport Layer Security certificates for new services, greatly increasing the flexibility of the infrastructure by allowing rapid changes to back-end services without having to change the static host names of user-facing services. Using the Traefik,[8] DNS names can be automatically created or changed on the fly, enabling a very dynamic structure. In addition, several monitoring services built into the proxy detect service outages, while logging services track performance data, access patterns and accounting information. The data are transferred from the proxy so that no additional disk space is used on the proxy to keep the setup simple to separate the individual components and also to take safety aspects into account. We actively participate in various working groups of the NFDI section "Common Infrastructure", primarily focusing on Multicloud, Overall Architecture, nfdi.software, Data Management Planning, Terminology Services, Identity and Access Management, and Research Software Engineering. We have also used past NFDI conferences to promote the service integration agenda. Through these efforts, we aim to encourage broader adoption by other communities and integration into the future NFDI service landscape.

The user management and accounting is realised in the DataPLANT infrastructure with Keycloak: a powerful open-source tool that acts as an AAI proxy between web-based services and either internal Keycloak accounts or external identity providers (IdP). Many external IdPs are relevant in the NFDI context, like Life Sciences AAI, ORCID and in particular the upcoming NFDI-AAI. While still in development, the NFDI-AAI aims to provide a NFDI-wide infrastructure to foster inter-consortia activities. The first steps to integrate DataPLANT Keycloak as a community AAI in the so-called NFDI-Infrastructure-Proxy were done as part of an incubator project.[9]

The further development of a conceptual Authentication, Authorization and Accounting (AAA) framework with mutual assurance (as part of IAM4nfdi) is a crucial step toward a sustainable services landscape. An effective accounting system requires coordination between the involved AAI, which provides appropriate attributes for attribution of resource reservation, and the offered services. This involves defining standardized service and infrastructure offerings that can be linked to their respective resources consumed on the provider side. For the DataHUB service relevant parameters might include storage volume and redundancy level, type of storage (e.g., object, block, or file storage), and expected performance parameters. For computing resources e.g. for the intended workflows, further parameters and limits are needed, such as CPU hours for HPC or cloud flavors with specific configurations and defined runtimes.

8 Widely used implementation of a reverse proxy, https://traefik.io; *Visited on May 22, 2025.*

9 Overview on projects within IAM4nfdi completed https://incubators.nfdi-aai.de/; *Visited on May 22, 2025.*

To enable future cost compensation, information about the infrastructure users and their institutions must be properly recorded and maintained. Additionally, it is useful to provide ongoing accounting information for monitoring purposes to both users and their funding entities. It would also be beneficial to offer forecasts on expected usage and related costs (Leendertse and Suchodoletz 2020). AAA concepts and structures need to be further developed to facilitate inter-institutional exchange of RDM services.

4 Outlook

The infrastructure stack has been in production for over four years across multiple locations. The volume of stored data is continuously increasing, CI services are being expanded, and the integration of computational workflows is being enhanced. Over the coming years, the focus will be on maintaining and improving the existing service landscape, securing broad community support, and ensuring sustainable long-term operation and development beyond the initial funding.

We will further evaluate and implement basic services provided through Base4NFDI when they become available to complement our service stack. Our infrastructure provisioning approach is modular, allowing other communities to reuse and recombine created services and tools. Individual modules can be replaced by future joint services of groups of consortia or the whole NFDI, such as for data publication. A layered model is used to separate tools and services from the underlying storage and compute infrastructure (Bauer, Tschöpe, and Suchodoletz 2025). This fits into the proposed service stack for the NFDI EOSC EU node.[10]

The service stack, either partially or completely, is made available to users beyond DataPLANT's core participants. All components are open-source and accessible from public repositories to promote collaboration and innovation. In the envisioned second funding phase of DataPLANT, a provider circle is planned to be set up to ensure continuous development and sustainability of services. Efforts will be made to integrate more closely with services from other consortia, state initiatives and international developments on research data management, enabling overarching RDM services.

Acknowledgements

We acknowledge the support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1).

10 See, https://www.nfdi.de/nfdi-is-part-of-eoscs-build-up-phase; *Visited on May 22, 2025.*

Authorship Contributions

Dirk von Suchodoletz is the main author of this paper and coordinator of this project. Marcel Tschöpe and Jonathan Bauer significantly contributed to the initial design of the Archigator microservice, which interfaces with the respective APIs of the adapted GitLab and InvenioRDM packages, as well as to the implementation of the CI pipeline. They were also involved in the AAI (Authentication and Authorization Infrastructure) for DataPLANT, including the integration of Keycloak and IAM4NFDI. Kevin Schneider and Timo Mühlhaus are among the primary authors of the ARC specification, contributing to the definition and refinement of user requirements for the DataHUB, the data publication workflow, and the validation process. All authors jointly designed and developed the DataPLANT service landscape, with additional contributions from other team members over the past four years.

Conflict of Interest

There are no conflicts of interest.

Bibliography

Bauer, Jonathan, Marcel Tschöpe, Paul Chr. Schnürle, Julian Weidhase, Christoph Garth, and Timo Mühlhaus. 2024. *Cloud based flexible service infrastructure stack for the NFDI.* Structure graphics for this short paper. Visited on July 7, 2025. https://events.gwdg.de/event/658/contributions/2388/.

Bauer, Jonathan, Marcel Tschöpe, and Dirk von Suchodoletz. 2025. *DataPLANT services design – Considerations towards a common NFDI landscape.* Figure of the structure of DataPLANT service integration. https://doi.org/10.11588/heidok.00036418.

Bauer, Jonathan, Marcel Tschöpe, Dirk von Suchodoletz, Cristina Martins-Rodrigues, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, and Louisa Perelo. 2024. "From DataPLANT's DataHUB to DataPUB(lication)". In *Proceedings of the 15th International Workshop on Science Gateways (IWSG2023),* edited by Jens Krüger and Sandra Gesing. Publication by the 15th International Workshop on Science Gateways (IWSG2023), 13-15 June in Tübingen. https://doi.org/10.15496/publikation-100323.

Leendertse, Jan, and Dirk von Suchodoletz. 2020. “Kosten und Aufwände von Forschungsdatenmanagement”. *Bausteine Forschungsdatenmanagement,* number 1 (1): 1–7. https://doi.org/10.17192/bfdm.2020.1.8246. https://bausteine-fdm.de/article/view/8246.

Martins-Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger, and Björn Usadel. 2021. “DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung”. *Bausteine Forschungsdatenmanagement,* number 2 (2): 46–56. https://doi.org/10.17192/bfdm.2021.2.8335.

Mühlhaus, Timo, Dominik Brillhaus, Marcel Tschöpe, Oliver Maus, Björn Grüning, Christoph Garth, Cristina Martins Rodrigues, and Dirk Von Suchodoletz. 2022. “DataPLANT – Tools and Services to structure the Data Jungle for fundamental plant researchers”. In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 132–145. heiBOOKS. https://doi.org/10.11588/heibooks.979.c13724.

Suchodoletz, Dirk von, Jonathan Bauer, and Marcel Tschöpe. 2023. “DataPLANT Cloud Oriented Service Infrastructure”. In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.414.

Suchodoletz, Dirk von, Dominik Brilhaus, Marcel Tschöpe, and Jonathan Bauer. 2023. *GitLab as a tool for Research Data Management.* Zenodo, 1. Conference on Research Data Infrastructure, Karlsruhe. https://doi.org/10.5281/zenodo.10021180.

Suchodoletz, Dirk von, Timo Mühlhaus, Christoph Garth, and Björn Usadel. 2024. *Software repository for the NFDI.* Structure graphics for this short paper. Visited on July 7, 2025. https://events.gwdg.de/event/658/contributions/2386/.

Weil, Heinrich Lukas, Kevin Schneider, Marcel Tschöpe, Jonathan Bauer, Oliver Maus, Kevin Frey, Dominik Brilhaus, et al. 2023. “PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research”. *The Plant Journal,* 1–15.