1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.414



Published: 07 Sept. 2023

# DataPLANT Cloud Oriented Service Infrastructure

## Open for Integration and Adaptation

Dirk von Suchodoletz[1][https://orcid.org/0000-0002-4382-5104], Jonathan Bauer[2][https://orcid.org/0000-0002-5624-2055], and Marcel Tschöpe[3][https://orcid.org/0000-0002-3731-7664]

[1] Computer Center, University of Freiburg, Germany

**Abstract.** A core objective of the DataPLANT consortium is to provide a science gateway as a set of flexible cloud-based (micro) services. The setup aims at both on-premises installations and future integration into a shared NFDI infrastructure. We will present the DataPLANT DataHUB, which provides various RDM workflows to support research data scientists at different stages of the data lifecycle - from development to publication of the results obtained.

**Keywords:** FAIR Data sharing, Service Infrastructure, RDM Platform, cloud deployment, GitLab, InvenioRDM

## 1. Introduction

The aim of this presentation in the **Enabling RDM** track is to exchange ideas with other NFDI consortia on the services required for RDM and the principles for service development and deployment. These considerations can be used as input for joint infrastructure development, e.g. in the context of the NFDI Common Infrastructure section or Base4NFDI. Over the past two and a half years, the DataPLANT team has developed a set of software and system components that provide services to the basic plant research community [1,2]. The set of tools and microservices that have been developed and evolved to date have focused on extending the existing digital landscape of the typical plant scientist. The core services focus on data management, versioning, sharing and publishing. All services are centered around cloud deployable modules.

The development of applications and tools to support community-driven research data management requires the involvement of several parties. During the development of the services, we agreed on design principles to provide a high-level guidance and a set of criteria for creating desirable and maintainable applications. In DataPLANT, tool development is always motivated by community requirements, conveyed by researchers, e.g., through data stewards, to developers [3]. Developments in DataPLANT follow an incremental and iterative approach, ensuring commitment and alignment of expectations of all stakeholders. Another aim of the service development is to enable both central and local installation of services without divergent implementations. Thus, we hope to encourage adoption by other communities and integration into a future NFDI service infrastructure.

## 2. Basic Infrastructure

The basic infrastructure is the "Persistent/Dynamic" layer of Figure 1, which consists of the required storage and compute resources, which are provided by the de.NBI cloud in our setting. Optimally, the hosting infrastructure for virtual machines (VMs) or containers allows for automation, e.g., to redeploy a service after a failure or update.

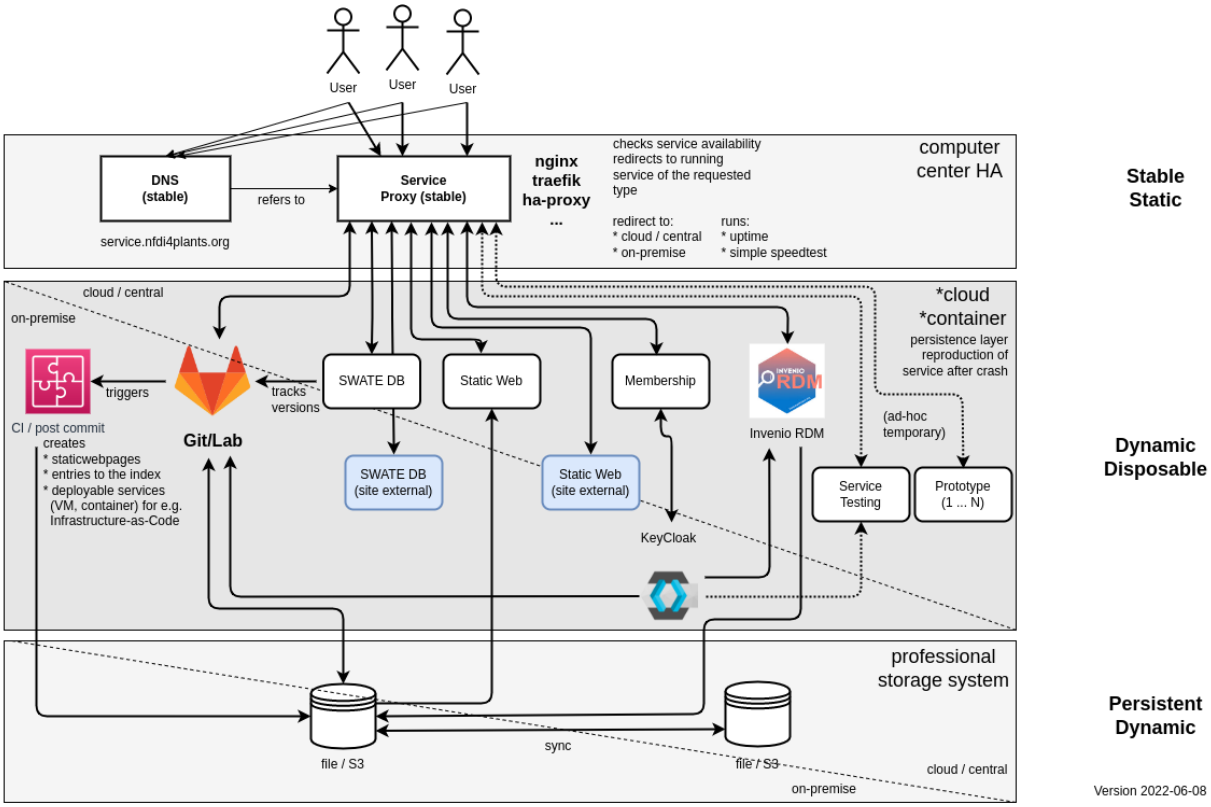


**Figure 1.** The DataPLANT services infrastructure building blocks.

Storage resources are used to store both the necessary service configurations and user data. Depending on the service provided locally, storage resources may be provided in the form of traditional network file systems such as NFS or SMB, or as object storage. The storage service implements the necessary redundancy to provide the required level of security for user data.

An authentication instance is required to authenticate users and the services behind them. The DataPLANT user management builds on existing AAIs. Well established services such as the Life Sciences AAI and ORCID can be combined with local authentication within the central DataPLANT authentication service. The infrastructure is based on KeyCloak, which supports modern authentication protocols such as OpenID Connect and SAML, enabling the integration of multiple AAIs and identity brokering. Providing AAI identity management that can easily connect to GitLab and other services using either protocol simplifies user management. Connecting multiple AAIs through KeyCloak allows our community to use their existing accounts, such as the Life Sciences AAI, their home institution or ORCID. We can assign different roles depending on the source of the account or specific attributes. Permissions can be derived from these roles to differentiate between users. These range from privileged users with full access to the data and the ability to create archives/publications, to users who only have a reporting function and/or read-only access to raw data.

The DataPLANT microservices are packaged as Docker containers running in the "dynamic/disposable" layer in Figure 1. This allows for a flexible deployment strategy, as the containers can be run both on VMs and/or directly in a container runtime environment such as Kubernetes or OpenShift. Depending on the local requirements for on-premises services, only a subset of the DataPLANT services will be considered.

The core helper services in DataPLANT consist of Traefik as a proxy service, Keycloak as an identity and access management framework for user management, and Kuma as a service monitoring framework. The entry proxy for DataPLANT services performs a number of functions, providing a stable point for users to connect to regardless of where individual services are actually geographically located and running. Another intended feature is that the proxy has the ability to trigger the (re)deployment (to a suitable infrastructure) if it is not available. The proxy can be used to redirect to other network destinations (outside the site itself) depending on the source institution (so you can transparently implement partially on premise). It is the established connection point that makes it easy to test new services. For these tasks it would be desirable to be able to configure the proxy on-the-fly via an API (should have a simple reload mechanism). Host helper services such as speedtest or uptime monitor.

## 3. DataHUB of combined GitLab and InvenioRDM

The DataPLANT DataHUB as a science gateway is primarily based on the open-source framework GitLab. It provides an entry point to various services, starting with a versioned, generated web page and additional modules for community interaction. Key features of the DataHUB are versioning, the ability to work in groups, support for multiple contributions, and easy-to-use access management. The DataHUB platform is where the DataPLANT Annotated Research Contexts (ARCs) [4] evolve to a certain state. While these can be tagged or shared, GitLab is not intended to provide long-term access or citation. In order to fulfil these essential FAIR criteria, we have implemented a data publication service that complements the DataHUB. It is implemented using the turnkey repository framework InvenioRDM, supported by a large international community of research institutions and led by CERN in Switzerland. InvenioRDM was chosen over other frameworks due to the modern microservice-oriented design, the open-source nature of the project and the large community involved in the project. For our central DataHUB installation, we provide a separate Docker image with slightly different modifications. This image also includes the InvenioRDM publishing mechanism, which allows the user to publish an ARC directly to InvenioRDM from the DataHUB and automatically receive a DOI for the publication. To achieve this, we use GitLab's event hooking mechanism and a modified GitLab Auto DevOps pipeline. We prefer this approach to project-specific CI/CD templates so that users do not have to set them up themselves.

The goal of sustainability of the research data dictated our decision of underlying frameworks and technologies used in the DataPLANT infrastructure. The choice of a git-based solution for managing ARCs was driven by the widespread use of the protocol as well as predictable long-term adoption. While InvenioRDM has its own data model for publication, the community is already working on exporting datasets in platform-independent and well-established data packaging concepts such as OCFL and RO-Crates to enable potential future migrations.

## Acknowledgement

We acknowledge the support of DataPLANT, funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442077441. We thank the Ministry of Science, Research and Education in Baden-Württemberg for their support of the BioDATEN Science Data Center which provided the necessary means for the InvenioRDM workflow integration.

## References


1. D. von Suchodoletz, T. Mühlhaus, J. Krüger, B. Usadel, and C. Martins Rodrigues “DataPLANT – ein NFDI-Konsortium der Pflanzengrundlagenforschung,” 2021. [Online]. Available: https://bausteine-fdm.de/article/view/8335
2. T. Mühlhaus, D. Brilhaus, M. Tschöpe, O. Maus, B. Grüning, C. Garth, C. Martins Rodrigues, and D. von Suchodoletz, “DataPLANT – tools and services to structure the data jungle for fundamental plant researchers,” in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 132–145. [Online]. Available: https://doi.org/10.11588/heibooks.979.c13724
3. D. von Suchodoletz, T. Mühlhaus, D. Brilhaus, H. Jabeen, B. Usadel, J. Krüger, H. Gauza and C. Martins Rodrigues, “Data Stewards as ambassadors between the NFDI and the community” in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 366–373. [Online]. Available: https://doi.org/10.11588/heibooks.979.c13750
4. C. Garth, J. Lukasczyk, T. Mühlhaus, B. Venn, J. Krüger, K. Glogowski, C. Martins Rodrigues, and D. von Suchodoletz, “Immutable yet evolving: Arcs for permanent sharing in the research data-time continuum,” in E-Science-Tage 2021: Share Your Research Data, V. Heuveline and N. Bisheh, Eds. Heidelberg: heiBOOKS, 2022, pp. 366–373. [Online]. Available: https://doi.org/10.11588/heibooks.979.c13751