Research Data Commons (RDC)

ⓘ This page and its associated sub-pages describe the architecture and existing services of the cloud-based information platform, the so-called Research Data Commons, of NFDI4Biodiversity. This page gives a broad overview of the RDC, while the subpages provide more detail and describe specific services that are already available, in development or still in the coneceptual phase. Please note that this is a living document that will continue to evolve. The development of the RDC is a long-term goal pursued together with other NFDI consortia and has an impact on several Base4NFDI projects that are at a very early stage or in the planning stage.

[Overview] [Architecture] [Connectivity among Layers and Components] [Detailed Information on Tools and Services]

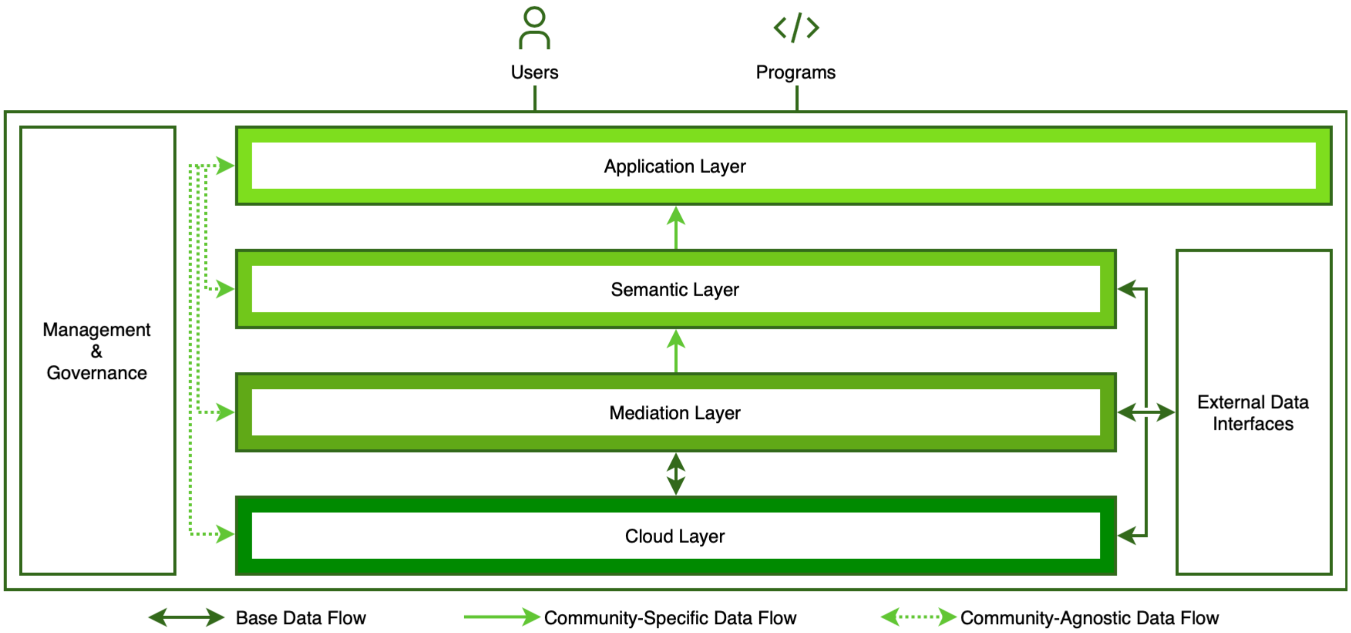
Overview

The Research Data Commons (RDC) is conceptualised as an expandable, cloud-based research infrastructure that provides scientists, data providers, and data consumers with powerful tools for creating FAIR data products and facilitates the exchange of data and services in a collaborative manner, both within the German National Research Data Infrastructure (NFDI) and beyond. In NFDI4Biodiversity, RDC development specifically serves to empower users from the biodiversity domain to reuse heterogeneous data sources, correlate them and conduct complex analyses to extract new research insight. In agreement with the FAIR principles, we work to make the offered data products and services computer-actionable, i.e., services will provide a FAIR application program interface (API). Based on an initial design of an architecture, RDC is developed incrementally, with the initial architecture becoming more specific and the first services of a reference implementation available. Selected components may be developed in cooperation with other NFDI consortia, and services will be continuously growing in number as the NFDI4Biodiversity project progresses.

Architecture

A brief overview of the RDC architecture is outlined in the attached figure. In order to manage the complexity of the RDC, we decided to organize the software architecture in layers and software components that interact with each other via well-defined interfaces. There are a total of four layers entitled:

- Cloud Layer
- Mediation Layer
- Semantic Layer
- Application Layer
- Management and Governance
- External Data Interfaces



The lower layers contain more technical functionality whereas the upper layers are primarily designed to end-users with domain knowledge. Each of the three lower layers consists of a few different technical components.

The **Cloud Layer**is the technical backbone based on a multi-cloud infrastructure including for example the de.NBI cloud and GDWG. These clould providers offer scalable functionality for distributed computing as well as cloud storage with near infinite resources such that users are empowered to run compute-intensive jobs or analyze very large data sets in a user-friendly way. In addition, there are cloud services like the Aruna Object Storage (AOS) for managing data in a unified model.

The **Mediation Layer** provides a self-service collection of community-agnostic tools for creating FAIR data products. A community-specific community, for example, a team that works together in one of the Use Case projects of NFDI4Biodiversity, is responsible for a specific data product and also responsible for its development process. A community unterstands the purpose the data product under its responsibility is generally used for, e.g. specific analyses and annotations. In particular, a data product uses common terminology of a certain discipline or subdiscipline. The mediation layer offers tools for data product developers to manage metadata, transform data from a technical data model into a semantic model, and improve data quality. Workflows serve to describe the steps of creating the data product from the source data sets, and thus, it also documents the provenance of a data product.

The **Semantic Layer** provides the community-specific data products that are created and maintained in the Mediation Layer using the technical datasets accessible from the Cloud Layer and other data from interfaces to external data providers. The data products comply with the FAIR principles and are computer-actionable, i.e. a computer system is able to find and access them and understand the corresponding schema. Data products are either physically available or built on demand when the product is accessed. In addition, self-service collection of community-agnostic tools are provided, such as Jupyter notebooks, which enable domain experts to create, deploy, and maintain community-specific applications on top of these data products.

The **Application Layer** consists of concrete applications and services developed for end users. These services can be community-agnostic, such as a search tool for datasets, or community-specific, such as a data portal for dragonflies or other species of interest. Community-specific applications are built on top of the data products in the semantic layer, while community-agnostic applications can access data from different layers. For example, the search tool requires access to data from the Cloud Layer and the Semantic Layer.

In addition to these four layers, there are two other essential elements in the architecture. The first one **Management & Governance** features tools and policies to manage rules and access rights for the resources offered in the four horizontal layers, including user management and monitoring of usage of the technical resources. The second, called **External Data Interfaces**, features a collection of interfaces for accessing external data sets. Obviously, RDC requires connectivity to established large data providers without the need to manage copies of their data in the Cloud Layer.

Connectivity among Layers and Components

The arrows within the architectural sketch illustrates the data flows between the layers and components. In order to unterstand the arrows from the Application Layer, it is first important to know the difference between **community-specific applications** and **community-agnostic applications**.

A *community-specific application* operates on data products offered in the Semantic Layer. In general, these data products are either physically available in the Semantic Layer or are generated on demand whenever a user requires access. In the first case, the Semantic Layer manages these available data produts using a storage service of the Cloud Layer. In the second case, the data product has to be created on-the-fly in the Mediation Layer, e.g., a harmonized composition of external data sets and data sets from the Cloud Layer of RDC. Once the data product is (partially) created, it will be cached for efficiency reasons in the Semantic Layer. The green arrrows in the middle of the architectural sketch illustrates this data flow.

The dataflows of a *community-agnostic application* are not as strict and are free to access an abritrary data source offered in any RDC layer. For example, the search tool for data sets generally requires access to all the three layers below the Application Layer. In particular, it requires access to the data discovery index maintained in the Cloud Layer. The dotted arrows on the left hand side of our diagram features the dataflows from community-agnostic applications.

In addition to these four layers, there are two other components. Management & Governance provides cross-cutting functionality related to each of the four layers, such as rules for participation, tools for monitoring resource usage in the RDC, and basic access control services. External Data Interfaces is the component for accessing external metadata and data sets. It is directly connected to the Mediation Layer, where data sets are transformed into a community-specific data model. In addition, GFBio Data Submission, a community-agnostic application, supports the transfer of data sets to the GFBio archives. Thus, the data flow of this component is bidirectional.

Finally, we want to emphasize that RDC strives for offering its functionality via an API such that other computer systems are able to access metadata and data without any user interaction. Moreover, RDC also provides applications for user interaction like a web-based data portal.

Detailed Information on Tools and Services

- BIIGLE
- BiodivPortal
- GFBio Search (Search and Harvesting-Infrastructure)
- Cloud Layer
  - Aruna Object Storage
- Management & Governance
  - Authentication Services: Life Science Login
  - Service Monitoring: Scorpion