NFDI Data Integration

Description of the topic and possible resulting services:

Data integration provides users, tools, and applications with unified views on data and metadata originating from different sources. There is no question that data integration is one of the essential building blocks in every NFDI consortium and in the NFDI as an integrated infrastructure. However, while physical data integration has been known from data warehousing for more than 30 years, NFDI comes with specific requirements, making data integration a challenging task. The reason for this is that NFDI behaves more like a so-called data lake where heterogeneous data sets with various kinds of schemata, partly with schema or without any schema, are supposed to be managed within a common cloud-based storage infrastructure. In addition, some data sets with high-volume, e.g. molecular data and satellite data, or with privacy concerns, e.g., medical data, are not physically available instead just virtually via dedicated interfaces. Thus, data integration requires, in addition, a mediator-based approach to support federated architectures, where data remains in the different sources and integration only takes place when access is made.

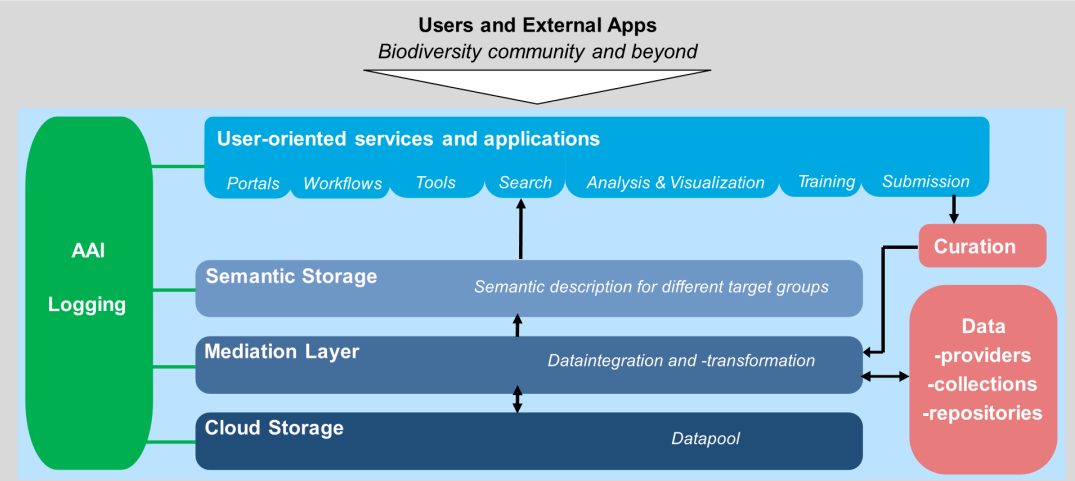

Figure 1: Initial architecture of the RDC with its three core layers cloud storage, mediation layer, and semantic storage.

In agreement with the basic architecture of RDC (Research Data Commons) and the FAIR data principles, data integration in NFDI follows the observation that different users and applications require different schemata depending on the specific context. Thus, data integration comprises the following services.

1. It is important to ingest the many different data sources of external providers into the multi-cloud storage either physically or virtually.

2. A powerful and comprehensive management of schemas and mappings between them is needed as a basis for data integration services.
3. Targeted and appropriate services for discovering data in the cloud storage will be needed. The foundation of such discovery services are indexes on metadata or link structures among similar data objects or user profiles.
4. Services are required for transforming the technical schemata of the cloud storage via the mediation layer of RDC into semantic schemata most suitable for users and applications.
5. It is essential to check and document data and metadata quality in the various phases of the data integration processes like data ingestion and schema transformation.

Data integration relates to all cross-cutting aspects like versioning, AAI, and logging that need to be considered when offering data integration services. For example, a comprehensive versioning of data, metadata, and transformations is a prerequisite for a provenance service within the integration process.

NFDI Section(s) and their Working Group(s) responsible for negotiation/development/evaluation of this topic:

orchestration:

- Section 1: Common Infrastructures (CI) → WG Data Integration

related Sections and WGs:

- Section 2: Metadata, Terminology, Provenance, Search & Harvesting
- in Section CI: Research Software Engineering, Research Knowledge Graphs, Federated Multicloud, Long Term Archival (and Access)

Potential Basic Service related to this topic - Research Data Integration Commons:

All NFDI consortia have in one form or another the problem of dealing with heterogeneous data and metadata from different sources and their integration. Specific solutions to this problem exist in some cases, and solutions are being worked on or sought in other consortia. Supporting these efforts and making them coherent and synergetic will be crucial

for the success of the NFDI. An overarching, well-orchestrated and coordinated approach will be important to efficiently find solutions for the individual consortia while advancing the integration of data and metadata across consortia. This requires a precise analysis of the requirements, which differ in detail, followed by work on an appropriate architecture and the implementation and introduction of corresponding services.

The current data integration approaches in different consortia are not yet homogenized and standardized . There are limitations with respect to functionality and scalability. Furthermore, a unified solution would foster consortia overarching data and service integration. To achieve acceptance, robustness, security and scalability established data integration technologies will be selected, customized and extended to the requirements of NFDI (one technologie candidate for data integration workflows would e. g. be Apache Spark). At the same time important targets will be ease of use, documentation and the adoption of durable approaches.

The respective architecture and the services have to be, and will be, put forward in close cooperation with the NFDI consortia, the activities in the Section (Meta)data, terminologies, provenance, the other WGs in the Section Common Infrastructure and international initiatives. To assure continuous evaluation and review of needs an external advisory board will be established and agile development approaches will be applied.

Potential partners with existing expertise (list the potential partner institutions, their consorti(a) and their roles):

The following list comprises technology experts as well as application domain experts. It is intended only as a first starting point which will be extended:

Consortium	Expert	Expertise
NFDI4Biodiversity	Bernhard Seeger (Uni Marburg) Birgitta König-Ries (Uni Jena)	cs
Text+	Andreas Henrich (Uni Bamberg)	cs
NFDI4Earth	Lars Bernard (TU Dresden) Claus Weiland (Senckenberg) Jan Bumberger (UFZ)	domain
DataPLANT	Dirk von Suchodoletz (Uni Freiburg) Timo Mühlhaus (TU Kaiserslautern)	cs
NFDI4Ing, NFDI-MatWerk	Marius Politze (RWTH Aachen) Rainer Stotzka (Karlsruhe Institute of Technology)	cs

NFDI4Objects	Matthias Renz (Uni Kiel)	cs
NFDI4Culture	Christian Bracht (Uni Marburg) Harald Sack (FIZ Karlsruhe) Lambert Heller (TIB Hannover)	domain
NFDI4DataScience	Sonja Schimmler (Fraunhofer FOKUS & TU Berlin) Stefan Dietze (GESIS & HHU Düsseldorf) Peter Mutschke (GESIS)	cs & domain
NFDI4Health	Toralf Kirsten (Uni Leipzig) Johannes Darms (ZB MED)	cs & domain
NFDI4Microbiota	Alexander Goesmann (JLU Gießen) Manja Marz (Uni Jena)	domain
KonsortSWD	Peter Mutschke (GESIS)	domain
PUNCH4NFDI	Jörn Künzemöller (U Bielefeld) Christoph Wissing (DESY)	domain
Additional Members without a participation of a consortia	Dirk Riehle (Uni Erlangen) Thorsten Papenbrock (Uni Marburg)	cs

Description of the needs addressed by this potential service on NFDI consortia:

Due to the intended availability of powerful data integration services, the benefit for users will be twofold. Data integration and tool connectivity will be facilitated for individual NFDI consortia. The use of an NFDI-wide solution here allows for expanded functionality with reduced effort. In addition, overarching data integration will be supported. Data of different consortia will be available in the data model most suitable for the underlying application purpose. It is the foundation for unlocking the potential of a new type of data-driven research across disciplines and leads to a substantial gain in scientific knowledge.

State of the art for this potential service:

Data integration is one of the most severe problems that prevent the reuse of data. There are various approaches to data integration like mediator-based integration in a federated system and active integration, e.g. data warehousing and data cubes. Existing tools and frameworks range from ETL tools over workflow approaches (such as Apache Spark) to simple solutions such as OpenRefine or XSLT-Scripts. However, there is no one solution that will meet all requirements, and all solutions require customizations and enhancements to meet the needs of heterogeneous research data. All the more important is a targeted selection and adaptation as well as optimized service offerings based on this.

Describe the overall strategy for the possible service with regard to the following stages:

1) Service initialisation strategy:

The overall strategy to establish integration services follows an agile approach. This strategy allows us to attract NFDI consortia as customers at a very early stage and to reevaluate services. First, we start a close cooperation with three to five consortia to dive into deeper discussions on the specific requirements and define user stories that we will use throughout the project to demonstrate success at the earliest possible stage.

2) Service integration strategy (if applicable):

Based on first-hand experiences with data integration problems, we start a more systematic evaluation of methods and tools suitable for establishing data integration services within NFDI. As both physical and virtual sources need to be integrated, we check whether available tools and services are ready-to-use for this setting and the diverse data and metadata types in the NFDI, and if not, how to adapt them and which need to be developed?

3) Ramping strategy for service operation (if applicable):

Based on the results of our agile development process in the first phases, we design a service architecture for data integration most suitable for NFDI. Our architecture will be set as the standard for data integration services that are broadly used in NFDI. The data integration services focus on automation and domain-agnostic integration techniques, which sets it apart from many existing approaches.

Address possible challenges and risks:

The challenge in NFDI is to establish data integration services that are considered useful for the many different consortia. Thus, there is a risk that our project is too ambitious or requires more time than the project of NFDI.

List other NFDI basic or subject specific topics or services with which this service will interact:

Data integration services provide the vertical connection among the technical layer and the user layer in the RDC. Thus, it is closely connected to many other services, e.g., multi cloud services, metadata services, AAI to name a few.

Preliminary List of References

- Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier.

- Dong, X. L., & Srivastava, D. (2015). Big data integration. *Synthesis Lectures on Data Management*, *7*(1), 1-198.

- Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys (CSUR)*, *53*(2), 1-42.

- Hai, R., Quix, C., & Jarke, M. (2021). Data lake concept and systems: a survey. *arXiv preprint arXiv:2106.09592*.
  https://arxiv.org/pdf/2106.09592.pdf

- Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97-120.
  https://arxiv.org/pdf/2107.11152

- Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986-1989.
  http://www.vldb.org/pvldb/vol12/p1986-nargesian.pdf

- Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016, April). From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web* (pp. 1419-1428).

- Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85