FAIR Data Spaces

FAIR Data Spaces - Building a Common Cloud-Based Data Space for Industry and Science by Linking Gaia-X and NFDI

Zeyd Boukhers (Fraunhofer Institute for Applied Information Technology), Daniela Mockler (Nationale Forschungsdateninfrastruktur (NFDI) e.V.)

CoRDI (Conference on Research Data Infrastructure)

Poster abstract

1. FAIR Data Spaces Demonstrators

1.1 NFDI4Biodiversity and Gaia-X

In this demonstrator, we have implemented a new data connector to the Aruna Object Storage [1] with a use case of linking public data and industrial data. The implementation includes easy cloud deployment using Docker and Kubernetes, connections to Copernicus and data from NFDI4BioDiversity [2], user data upload and annotation, as well as data visualization and manipulation functionalities. The system enables users to combine and analyze different datasets, resulting in enriched layers of information.

This implementation has been evaluated in a series of workshops that were dedicated to collect feedback from users, allowing them to use common Python tools. Each user was provided with a unique view and allowed to use their own data in conjunction with their own story. User feedback indicates that the demonstrator is versatile and intuitive due to its visual data representation, the recording of exploratory workflows, and the incorporation of Python functionality. Furthermore, connecting new data sources from the cloud and deployment proved to be straightforward. The next step is to include the demonstrator in the Gaia-X Federated Catalogue.

1.2 NFDI4Ing and Gaia-X

This demonstrator relies on the GitLab installation of RWTH Aachen, offered as a Community-SaaS. Users are authenticated via the existing DFN-AAI federation [3], granting access to resources managed directly by GitLab, especially Git repositories and custom workflows. Decentralized "Task Runners" use the workflow steps to combine and automatically apply research data management tasks such as loading referenced datasets, quality controls, analyses, or publications to the data. In this scenario, GitLab orchestrates the custom workflows on the decentralized infrastructure, currently provided by the users themselves.

The demonstrator investigates two private-public-cloud-scale-out scenarios based on this setup: scaling computing resources via "Task Runners" and scaling storage resources via referenced datasets. It processes data from fluid systems engineering and traffic systems.

Workflow steps are technically referenced as Docker images, with orchestration handled by the "GitLab Runner." In the targeted scale-out scenario, the "GitLab Runner" utilizes a cloud-based Kubernetes instance, where both the "GitLab Runner" and individual workflow steps are instantiated as containers. The "GitLab Runner" automatically transfers smaller datasets from the Git repository, while larger datasets must be stored, referenced, and processed in an object store. As next steps, we plan to 1) instantiate the cloud infrastructures in the Open Telekom Cloud and deNBI, 2) examine the scaling properties of the workflow images in the Kubernetes cluster, and 3) create a catalog of tested workflow images for use in data-driven workflows.

1.3 NFDI4Health and Gaia-X

In this demonstrator, we developed a use case of distributed analytics on Malaria disease, as one of the life-threatening diseases in many regions. Using a public dataset detailing Malaria cases and deaths from 2000-2017, we preprocessed the data, categorizing countries based on the WHO-region attribute. Three subsets of data emerged: Eastern Mediterranean and Africa, Americas and Europe, and Southeast Asia and the Western Pacific. Each subset is then stored in the so-called Personal Health Train Station: Leipzig, Cologne and cloud-De.NBI. Using our developed tool PADME [4], which is an implementation of Personal Health Train [5], we analysed the data across these

stations, providing users with detailed regional statistics. This infrastructure installed in the De.NBI cloud is based on Gaia-X and registered in the Gaia-X registry via Self-Descriptions. As a next step, we will consider sophisticated analytic methods based on common use cases.

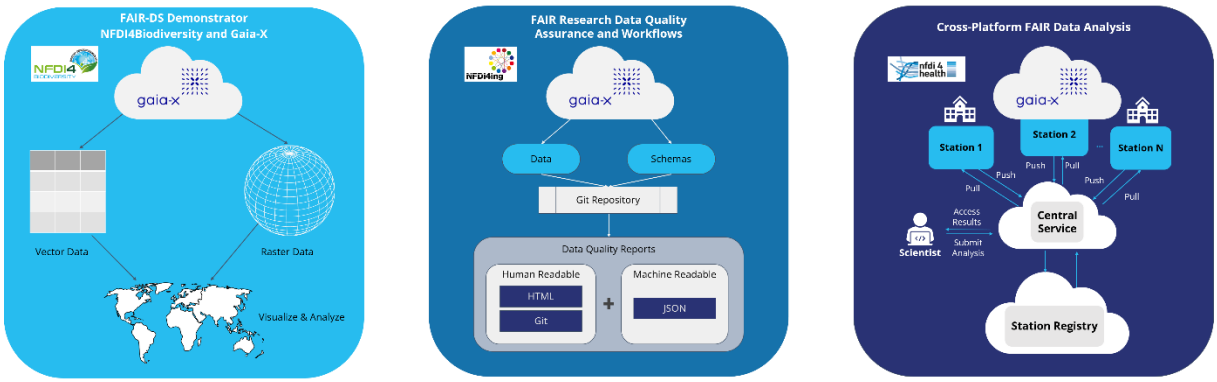

Figure 1. Schematic of the three demonstrators of FAIR Data Spaces.

2. Conclusion

Two years into the FAIR Data Spaces project, first building blocks for dataspaces have been set up and their interaction demonstrated using the three demonstrators. We will present the project achievements and discuss the next steps in expanding the demonstrators.

Funding

The FAIR Data Spaces project is funded by the Federal Ministry of Education and Research (BMBF).

Acknowledgement

We thank the 16 participating institutions in FAIR Data Spaces for their contributions and for the possibility to present the current project results at CoRDI 2023.

References

1. “ArunaStorage”. URL https://github.com/ArunaStorage (06.07.2023)
2. “NFDI4BioDiversity”. URL https://www.nfdi4biodiversity.org/en/ (06.07.2023)
3. “DFN-AAI”. URL: https://www.aai.dfn.de/index.en.html (06.07.2023)
4. “Platform for Analytics and Distributed Machine Learning for Enterprises” (PADME). URL: https://padme-analytics.de/ (06.07.2023)
5. “The Personal Health Train”. URL https://pht.health-ri.nl/ (06.07.2023)

https://www.nfdi.de/fair-data-spaces/

#FAIRDataSpaces @FAIRDataSpaces

https://www.nfdi.de/fair-data-spaces-newsletter/