Advancing Data Exchange Innovations in FAIR Data Spaces

Christoph Lange, Fraunhofer FIT, and project team

27.05.2024

FAIR Data Spaces

FAIR Data Spaces at a glance

- Vision: Development of a common cloud-based data space for industry and science
- Mission: Create and expand synergies between existing technologies and communities
- Financing: Funded by the Federal Ministry of Education and Research
- Runtime: May 2021 – December 2024
- Participation: 16 participating organizations

Project goals

FAIR* Data Spaces aims to build a shared cloud-based data space for economy and science by linking Gaia-X and NFDI

- Identify and leverage synergies of cooperation between the two initiatives
- Interweaving the content of the initiatives by clarifying legal and ethical issues and providing technical foundations
- Promoting a sovereign exchange of data between industry and science both nationally and in the EU in concrete applications and fields of work

*Guidelines findable, accessible, interoperable, reusable

Target communities

Gaia-X

Economy

- EU-initiated project, brings together stakeholders from industry, science and administration
- The goal is to create an open, transparent and secure federated data infrastructure
- Gaia-X architecture consists of a variety of individual platforms that follow a common standard

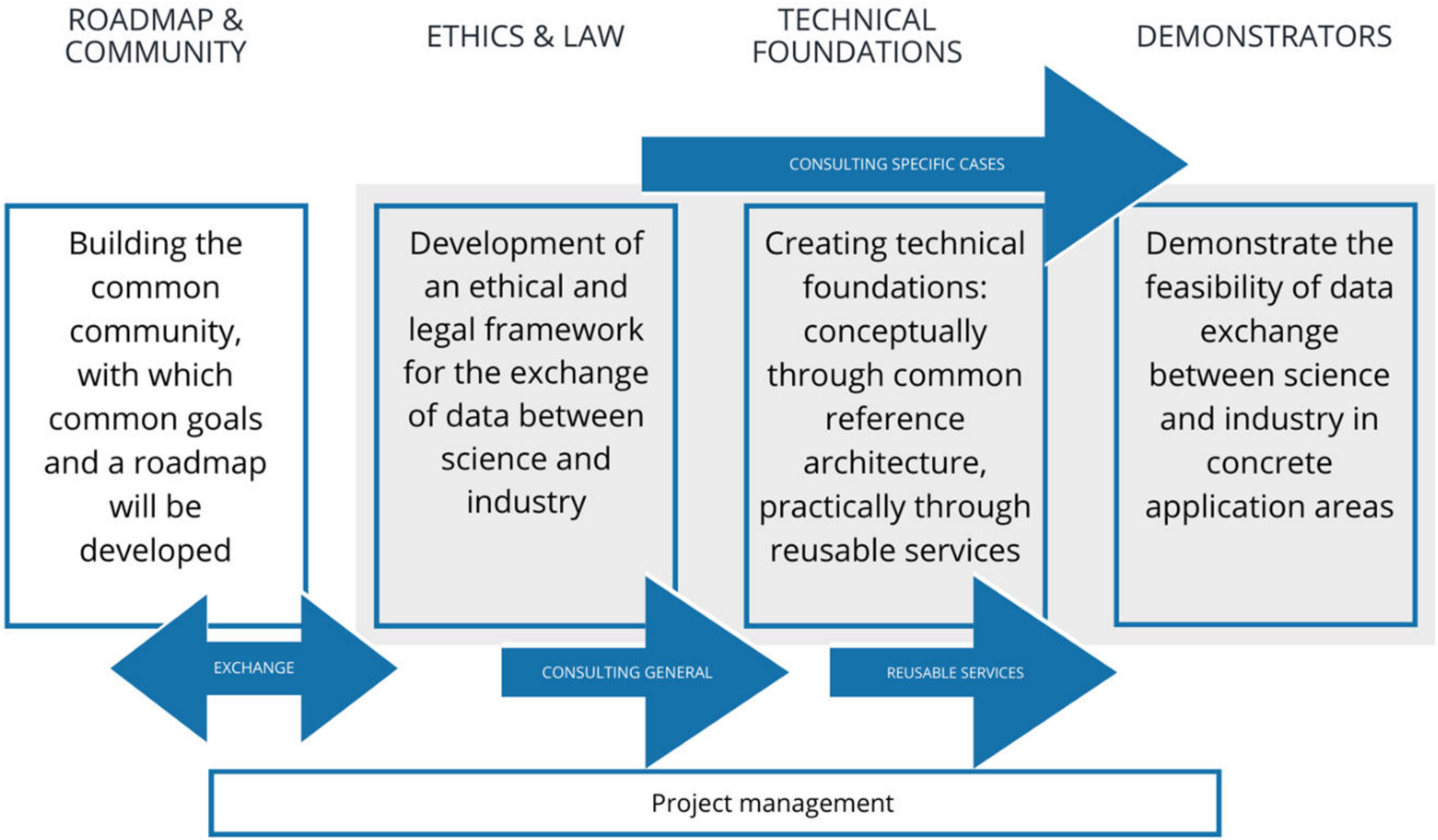
NFDI

Science

- purpose of the association is to promote science and research through a National Research Data Infrastructure, which establishes and further develops an overarching research data management in Germany and increases the efficiency of the entire German science and research system
- Making research data FAIR

In addition, linkage with EU data spaces and connection to EOSC

Overview task areas



ROADMAP & COMMUNITY

ETHICS & LAW

TECHNICAL FOUNDATIONS

DEMONSTRATORS

CONSULTING SPECIFIC CASES

Building the common community, with which common goals and a roadmap will be developed

Development of an ethical and legal framework for the exchange of data between science and industry

Creating technical foundations: conceptually through common reference architecture, practically through reusable services

Demonstrate the feasibility of data exchange between science and industry in concrete application areas

EXCHANGE

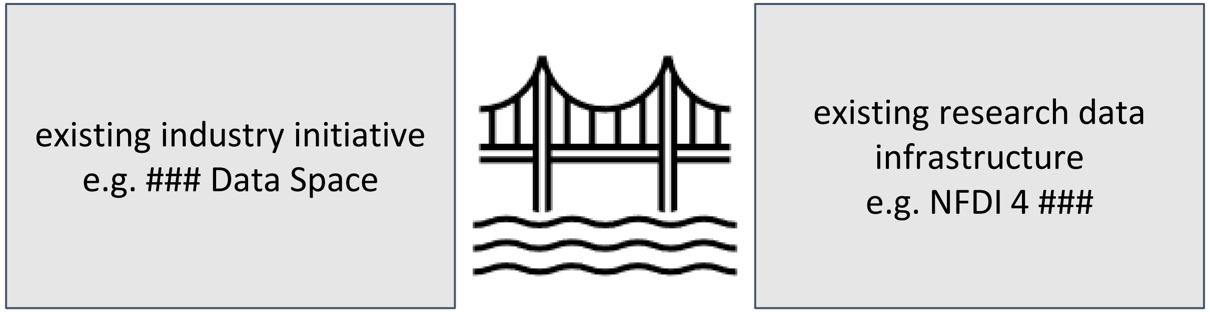
CONSULTING GENERAL

REUSABLE SERVICES

Project management

Objectives of the Open Call

- Have further demonstrators developed, beyond our three own ones (see below)
- Either extensions/additions to our demonstrators, or independent innovation
- Based on the same principles, standards and technology as FAIR Data Spaces
- Round 1 (call closed; June to November 2024): up to 3 contracts @ 60 k€
- Round 2: starting as soon as possible, ending in December 2024
  - ▪ same overall budget (180 k€)
  - ▪ additional requirement: bridging operational initiatives



existing industry initiative
e.g. ### Data Space

existing research data infrastructure
e.g. NFDI 4 ###

Open Call Round 1

Proposal evaluation criteria:

Interoperability with FAIR Data Spaces	15%
System architecture	15%
Development process	5%
Documentation	3%
Testing	15%
Security	7%
Innovation degree (evaluated by expert review board): ● Practical demonstration ● Industry↔research cross-benefit ● Going beyond state of the art (of FAIR Data Spaces and general) ● Awareness of ELSA	40%

Open Call Round 1

Requirements and Execution:

- Technology: common programming languages; container deployment; documented interfaces (e.g., OpenAPI); W3C semantic metadata (RDF etc.)
- Legal paperwork (not in the focus of this webinar), reference projects, team CVs
- 8-page concept paper:
  a. summary, architecture, frontend/backend implementation (addressing all evaluation criteria), innovation / added value.
  b. informative: time schedule, applicant profile, cost calculation
- Deliverable (at end of contract): open source code, documentation, website, video
- Schedule (~6 months):
  a. continuous implementation using repository and issue tracker
  b. regular status calls
  c. joint hackathon
  d. public presentation

Open Call Round 2

Discussion points for today – not formally implying any conditions that will apply to Round 2

- Who is here, representing what initiatives?
  a. ideally we'd have "consortia" formed from one company + and research organization (but not necessarily a member of an NFDI consortium or industry association)
     - However, we might also contract a single organization that provides strong evidence of support from "both sides" (e.g., letters of support)
     - It will not be sufficient to merely apply "industry-ready" *technology*, such as the EDC
  b. technical foundations on these slides are *representative*, so do build on what you have established in your existing initiatives, but there *must* be a "FAIR" interoperability layer.
     - as an alternative to Gaia-X & IDS specifically, you may refer to Data Spaces Blueprint
  c. Are other Fraunhofer institutes eligible – to be clarified! (Fraunhofer contracting "itself")
- Have you already reached out to "the other side"?
- What do you consider feasible within ~4 months (end of project minus formal/legal process)
- Any further questions you may have?
  a. First round was coordinated via e-Vergabe, second round most likely as well

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

Technical Foundations / Architecture (1)

Architecture with a heavy focus on modern cloud-native technologies

- Virtual machines

openstack [1]

- Storage

ceph [4]

Object storage (S3) [5]

- Containers

docker [2]

kubernetes [3]

- CI/CD

flux [6]

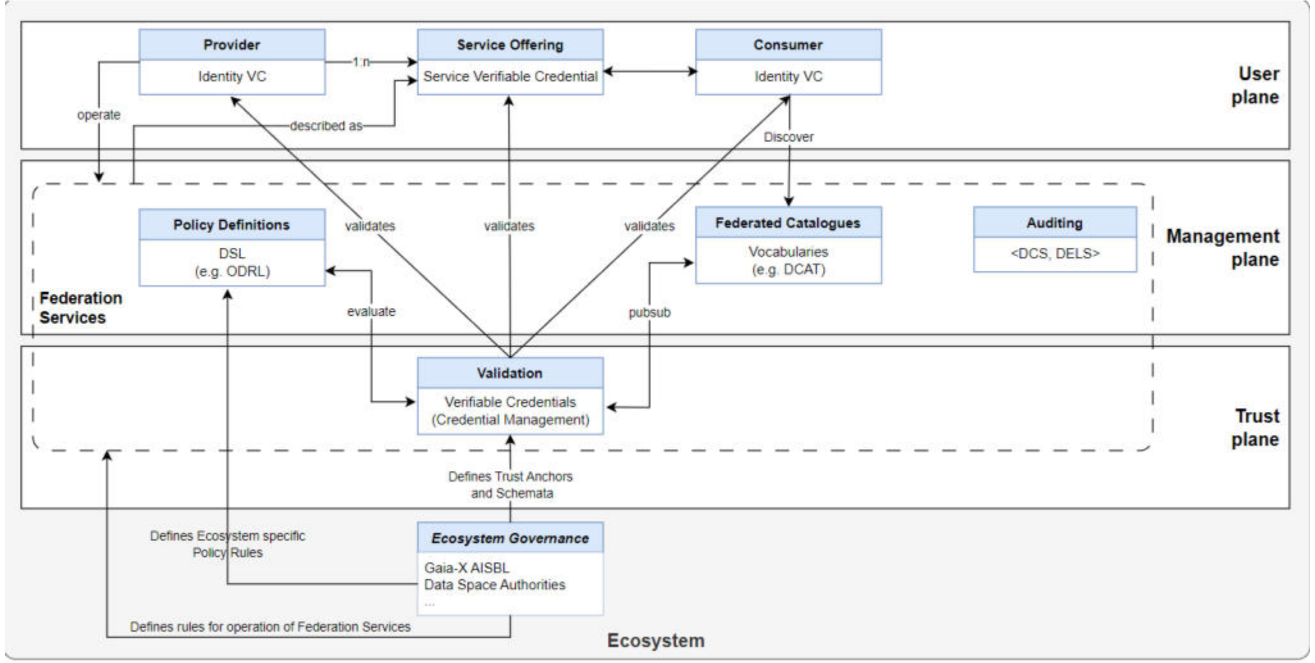
GitLab Runners

GitHub Actions

Powered by the de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE community cloud [7]

Technical Foundations / Architecture (2)

Gaia-X Architecture and Trust Framework (we are a “Gaia-X Qualified Project”)



User plane

Provider
Identity VC

Service Offering
Service Verifiable Credential

Consumer
Identity VC

1:n
operate
described as
Discover

Management plane

Federation Services

Policy Definitions
DSL
(e.g. ODRL)

Federated Catalogues
Vocabularies
(e.g. DCAT)

Auditing
<DCS, DELS>

validates
validates
validates
evaluate
pubsub

Trust plane

Validation
Verifiable Credentials
(Credential Management)

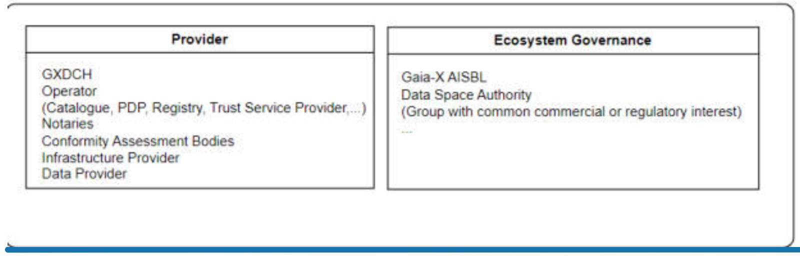
Defines Trust Anchors and Schemata

Defines Ecosystem specific Policy Rules

Ecosystem Governance
Gaia-X AISBL
Data Space Authorities
...

Defines rules for operation of Federation Services

Ecosystem

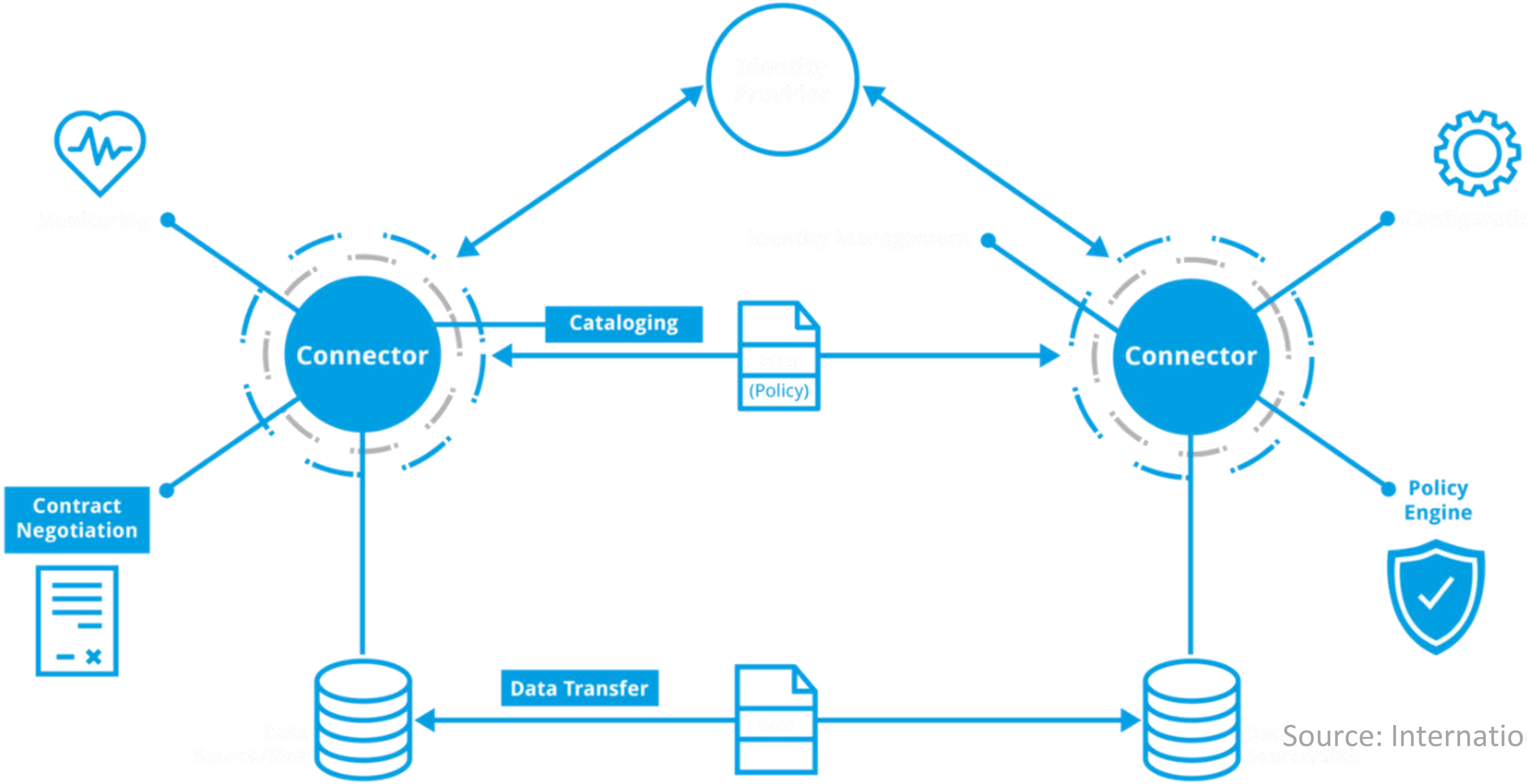


Provider	Ecosystem Governance
GXDCH Operator (Catalogue, PDP, Registry, Trust Service Provider,...) Notaries Conformity Assessment Bodies Infrastructure Provider Data Provider	Gaia-X AISBL Data Space Authority (Group with common commercial or regulatory interest) ...

Source: Gaia-X European Association for Data and Cloud AISBL

Technical Foundations / Architecture (3)

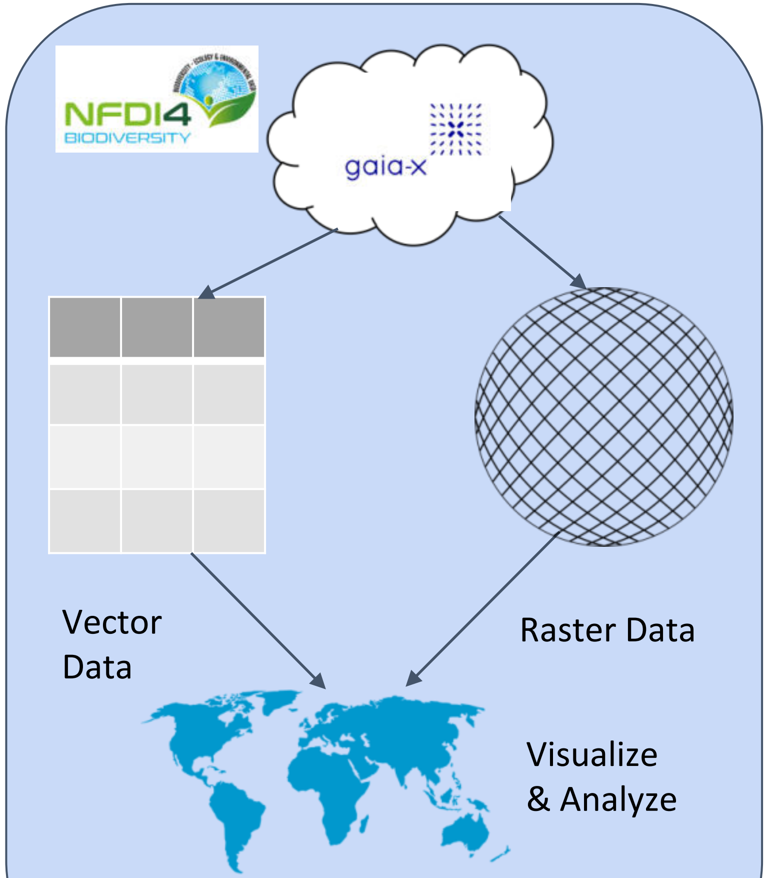
Sovereign Data Exchange via the Dataspace Protocol



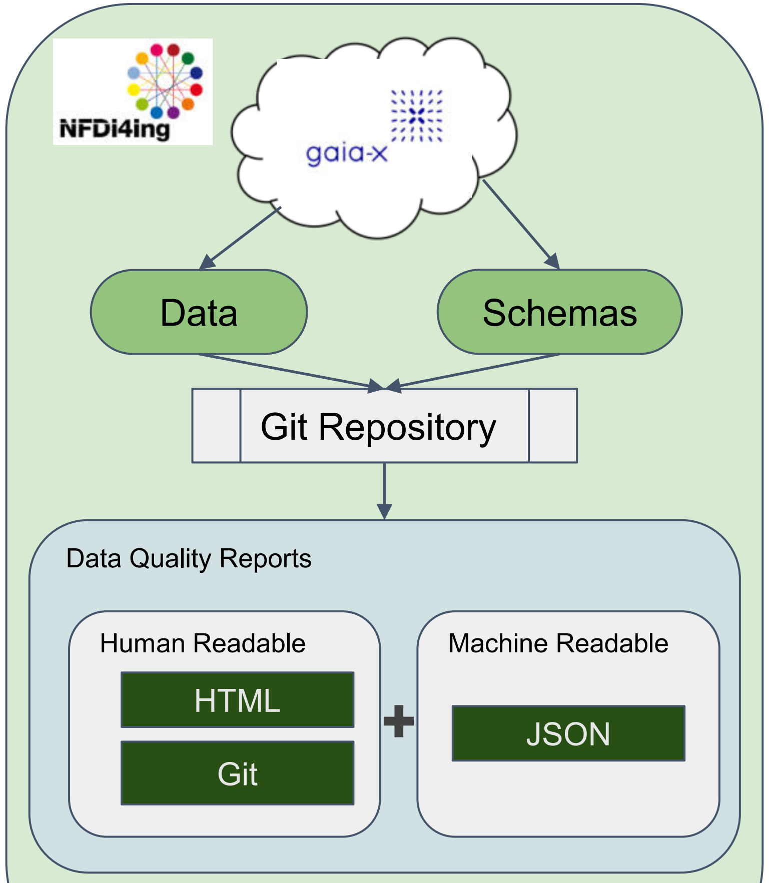
Source: International Data Spaces Association

FAIR Data Spaces Demonstrators

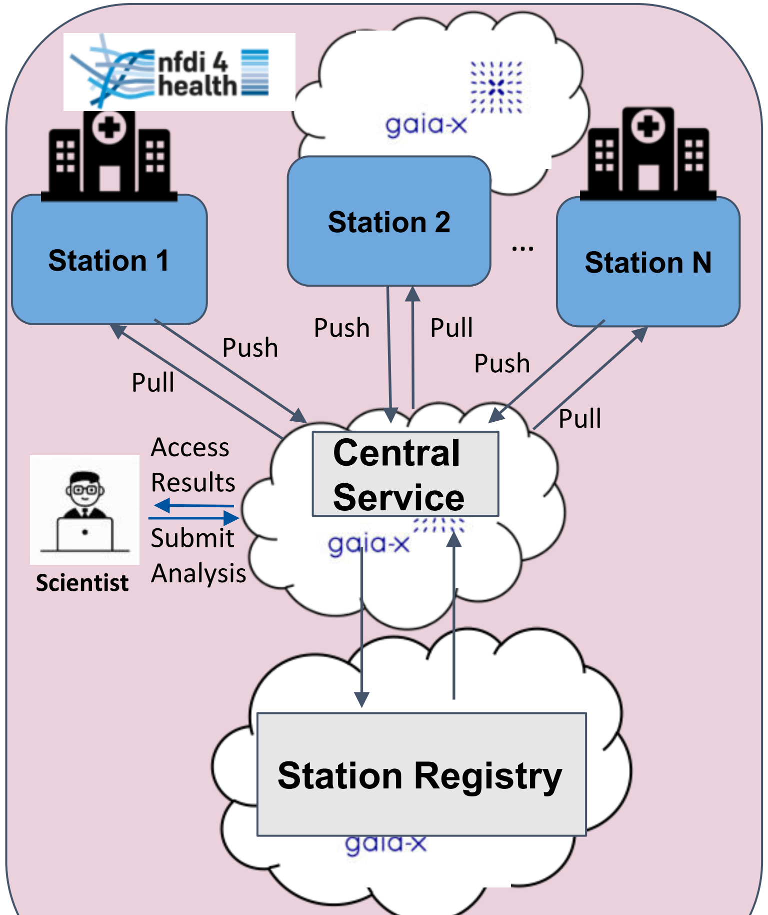
https://github.com/FAIR-DS4NFDI/FAIR-DSWiki/wiki



FAIR-DS Demonstrator
NFDI4Biodiversity and Gaia-X



FAIR Research Data Quality
Assurance and Workflows



Cross-Platform FAIR Data Analysis

Demonstrator 4.1: NFDI4Biodiversity

Nikolaus Glombiewski, Bernhard Seeger (Philipps-Universität Marburg)

NFDI4Biodiversity Demonstrator Overview

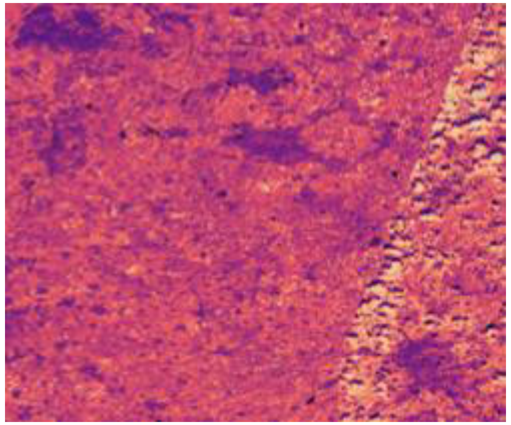
- Application:
  - Spatio-Temporal Data Analysis
  - Heterogeneous Data Sources
  - Rust, Python, Angular
  - Docker, OpenIdConnect (Keycloak)

- Connection to Research Data Infrastructure:
  - Part of NFDI4Biodiversity Research Data Commons
  - The Visualization, Analysis and Transformation (VAT) System powered by Geo Engine

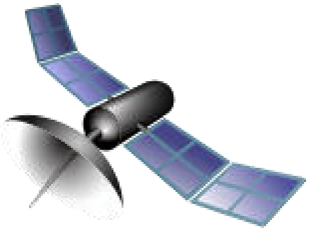
- Connection to Gaia-X:
  - Service Offering in a Federated Catalogue

Spatio-Temporal Data Analysis

Species	Coordinates	Date
Bird	48.856614, 2.352221	10.05.1977
Cat	41.8933203, 12.4829321	18.04.1980
Elephant	52.517037, 13.38886	21.10.2015

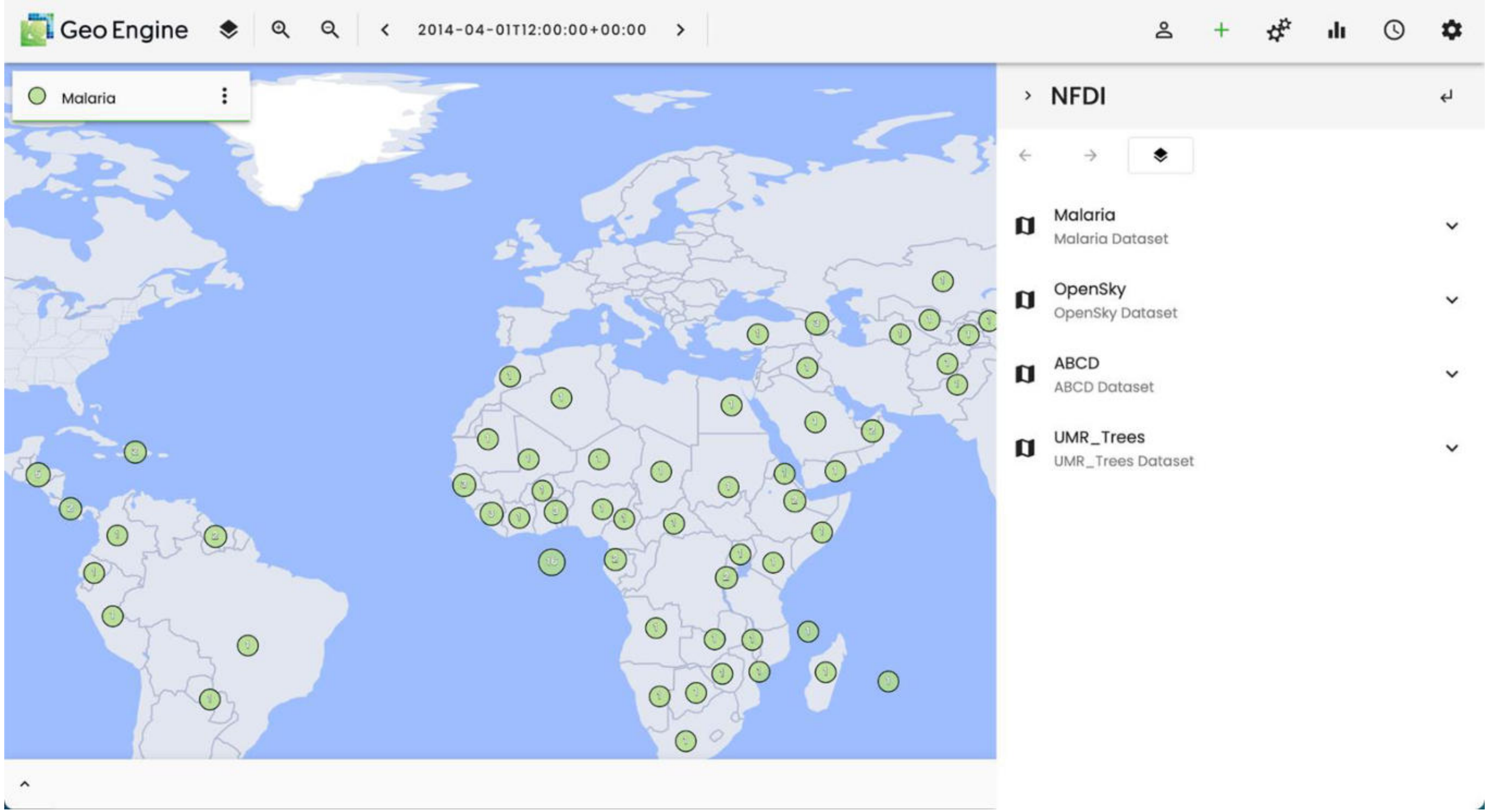


Example 1:
Vector Data
(e.g. manual recordings)



Example 2:
Raster Data
(e.g. images from satellites)

Geo Engine: Adding Data Sources



NFDI

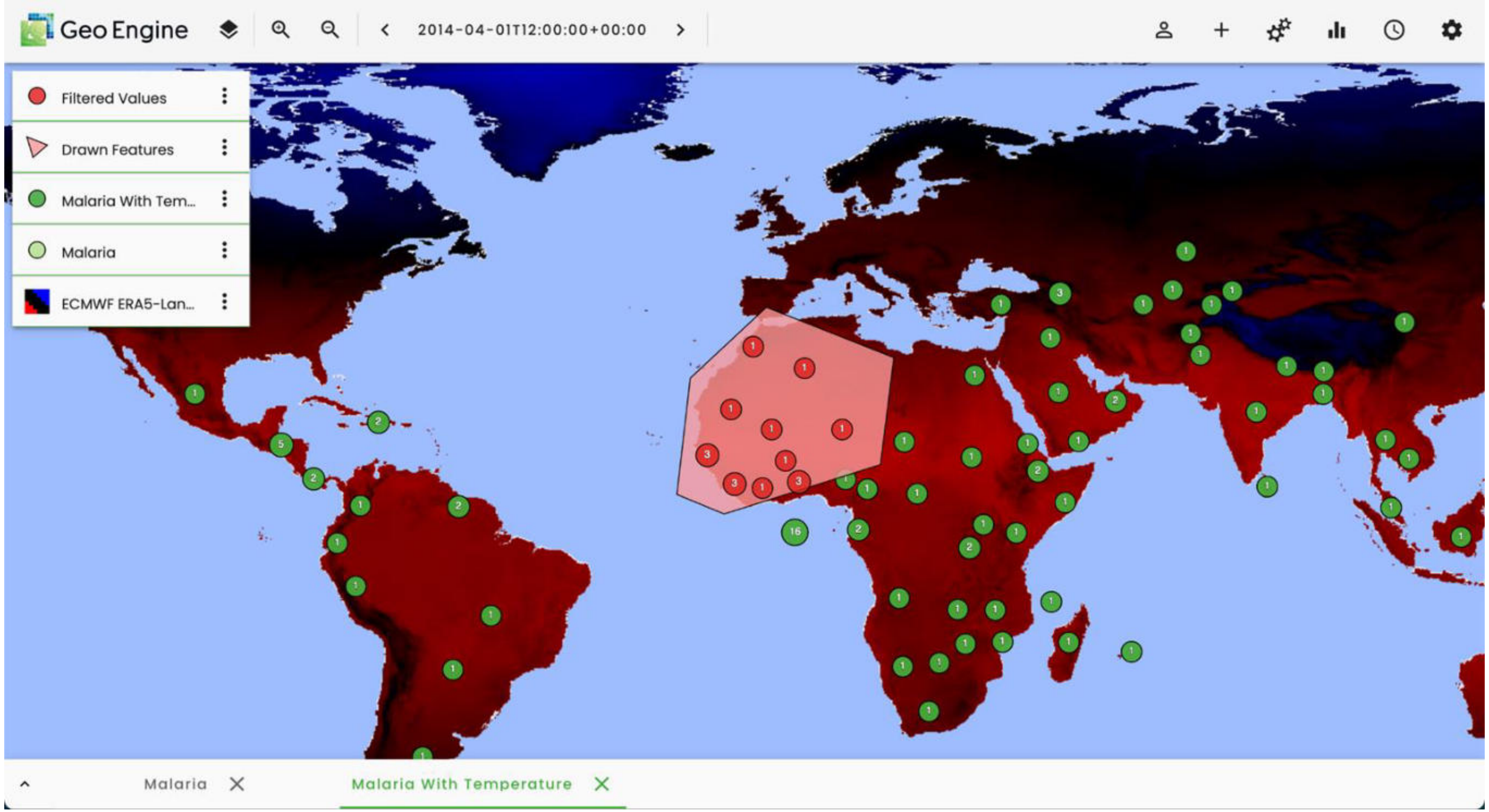
Malaria
Malaria Dataset

OpenSky
OpenSky Dataset

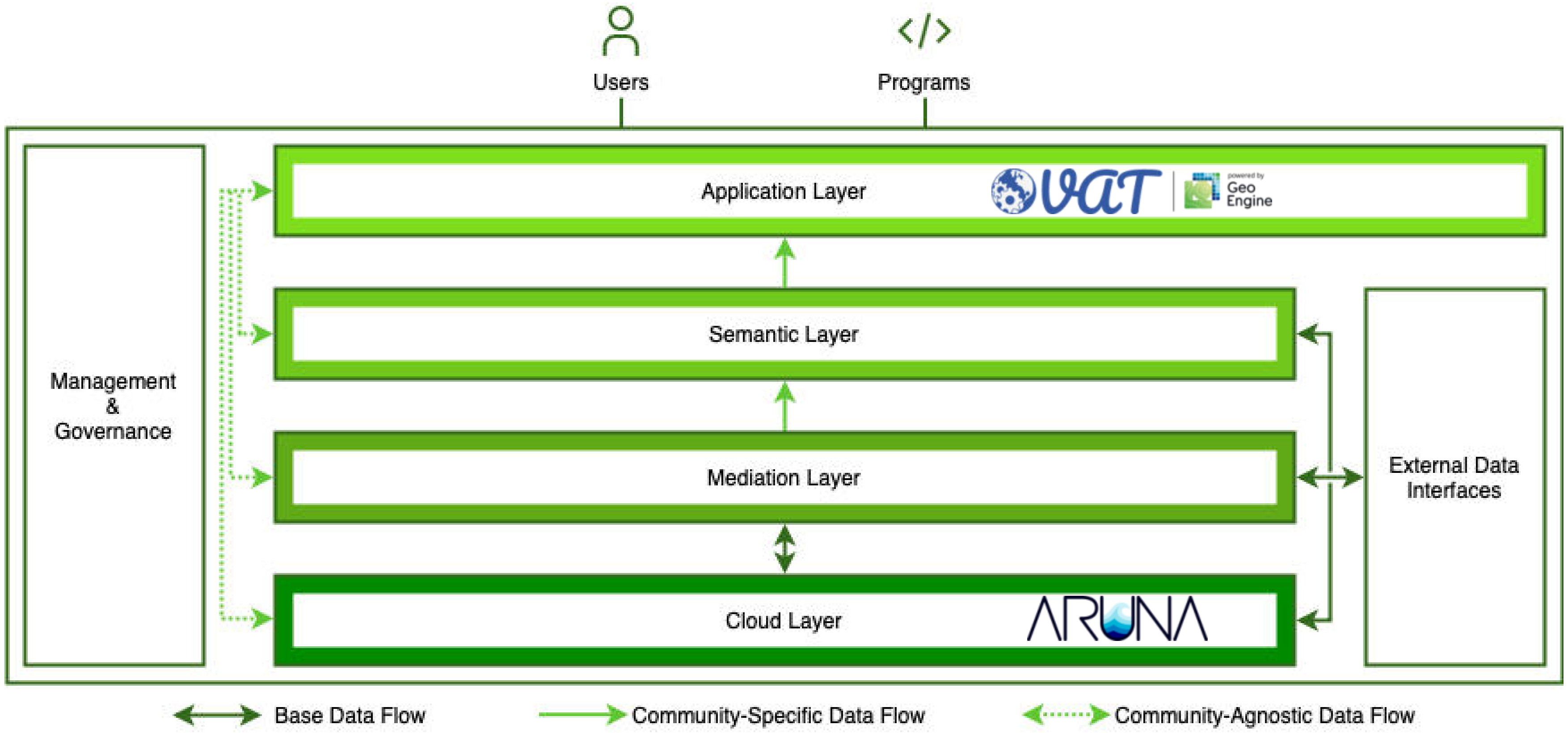
ABCD
ABCD Dataset

UMR_Trees
UMR_Trees Dataset

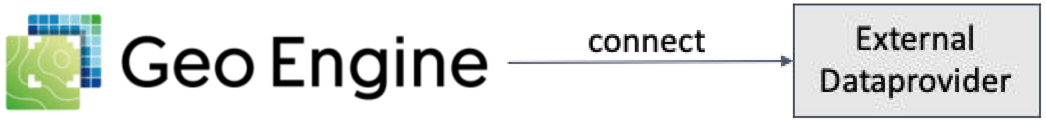
Geo Engine: Analysis



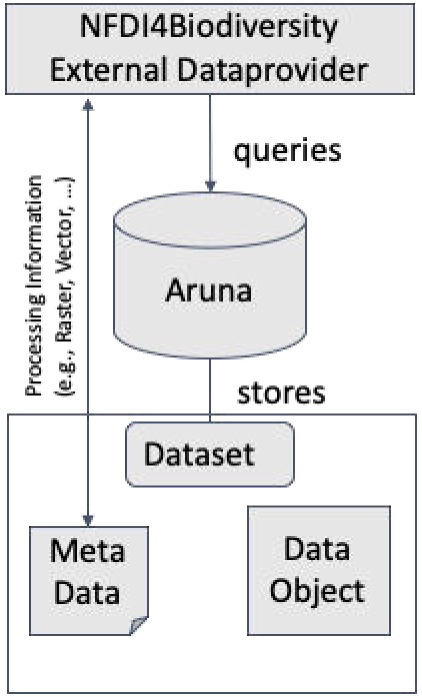
NFDI4Biodiversity: Research Data Commons



Connecting Layers with Geo Engine



- Application for Spatio-Temporal Data
- External Data Providers:
  - Standardized Protocols for Spatio-Temporal Data
  - Custom Data Exchange when necessary
- Also in “Mediation Layer” for offering FAIR Datasets

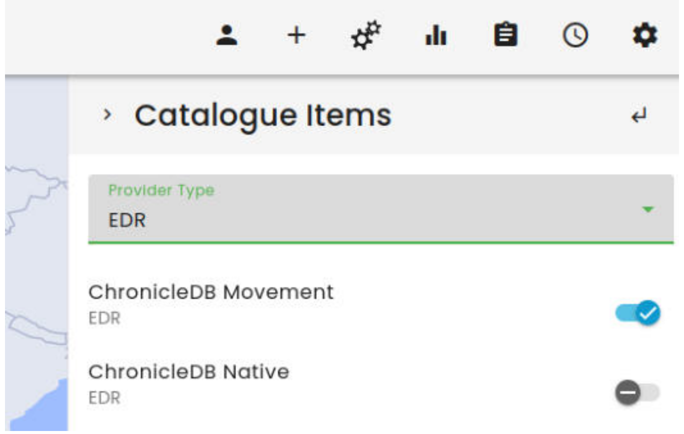

NFDI4Biodiversity External Dataprovider
queries
Processing Information (e.g., Raster, Vector, ...)
Aruna
stores
Dataset
Meta Data
Data Object

Geo Engine and Gaia-X Federated Catalogue

- Technological Basis:
  Gaia-X compliant catalogue developed by Eclipse XFSC (Cross Federation Services Components) using Spring, OpenIdConnect, PostgreSQL, Neo4j

- GeoEngine Self-Description:
  - A Service Offering in JSON-LD format
  - Verifiable Credentials: Set of claims or attributes, digitally signed by a trusted entity
    Who? What? Where? Which standard?

- Adding Data from supported dcat:DataService
  - OGC Environmental Data Retrieval (EDR)
  - In principal: Aruna Object Storage



Catalogue Items

Provider Type: EDR

ChronicleDB Movement
EDR

ChronicleDB Native
EDR

Demonstrator 4.2:

Data Quality Assurance and Workflows

Jonathan Hartman, RWTH Aachen University

Data Quality Assurance and Workflows

Goals

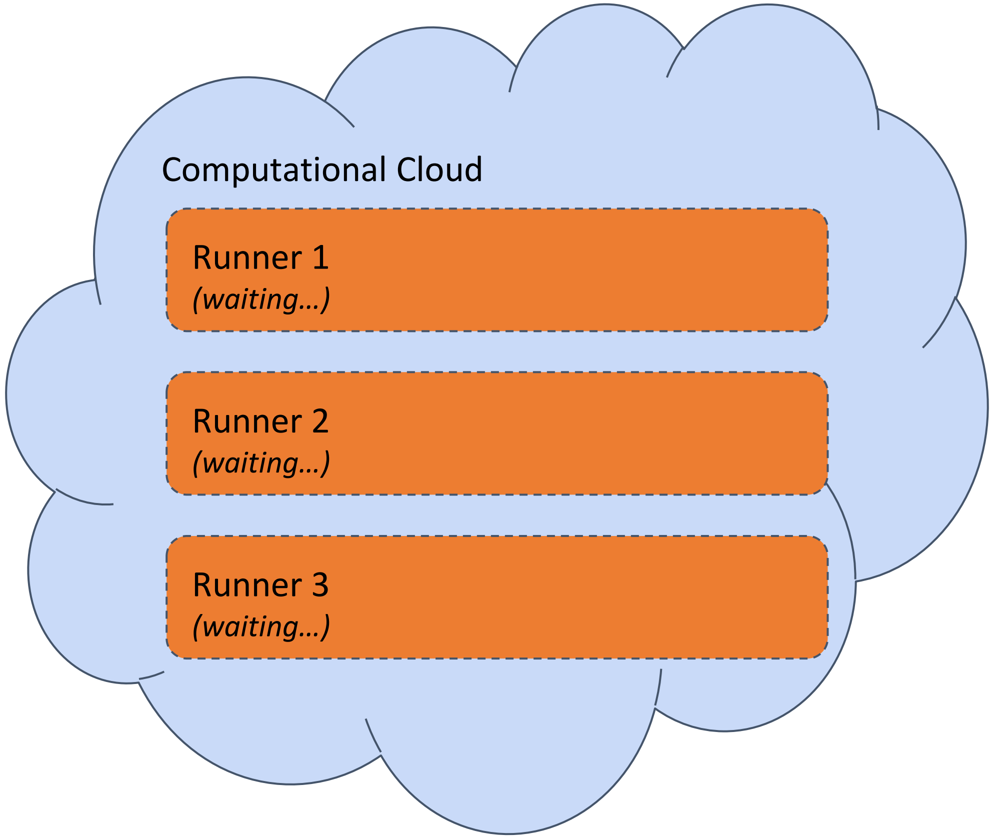
- Build off of existing Infrastructure
  - git.rwth-aachen.de
  - Open Telekom Cloud (OTC)
- Provide an example of Automated Analysis / Data QA
  - the demonstrator
- Provide a framework for hosting / sharing Workflows

Data Quality Assurance and Workflows

Infrastructure

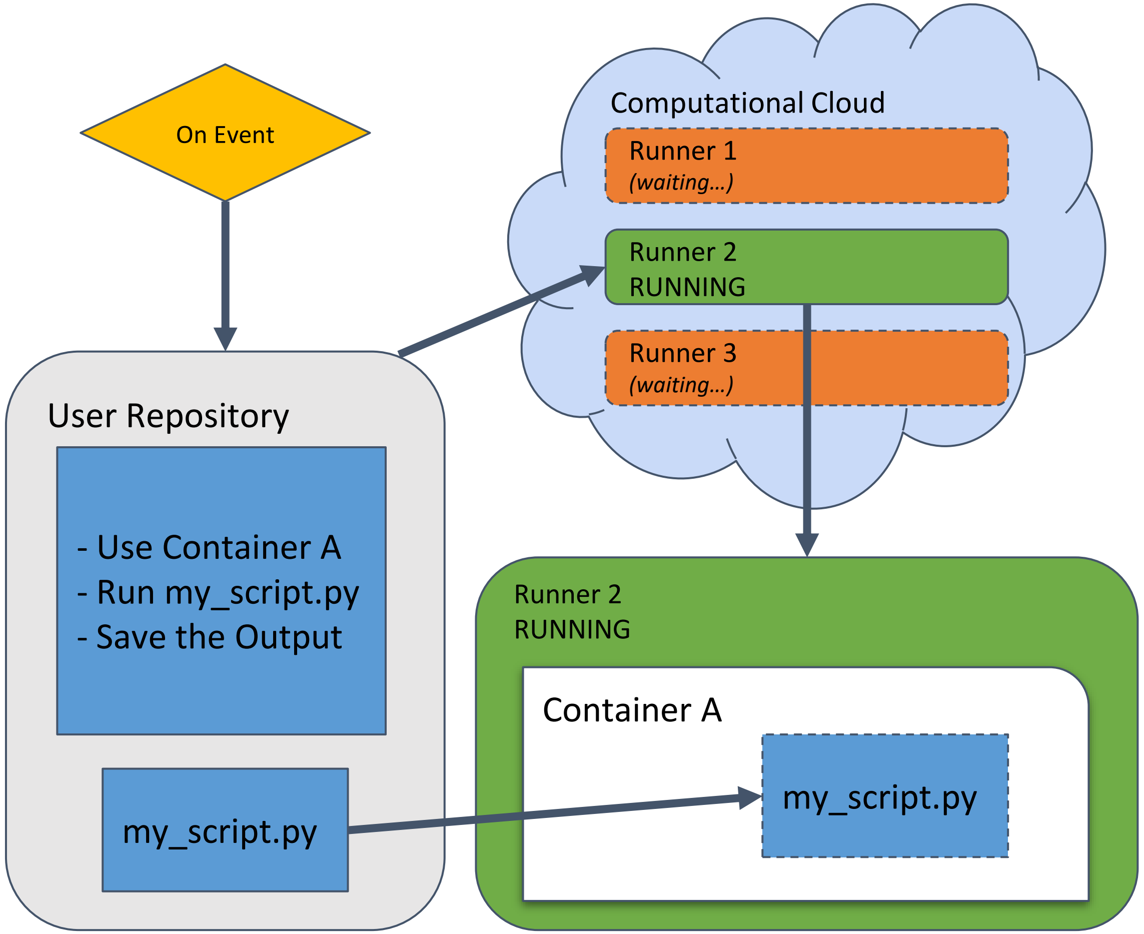
GitLab & Runners

- Lightweight agents controlled by CI/CD Scripts from Repositories
- Scalable, based on the workload
- Isolated, each runner context is run in its own container
- Customizable, capable of loading a huge variety of containers
- Multiple runners can be assigned to a project / group.



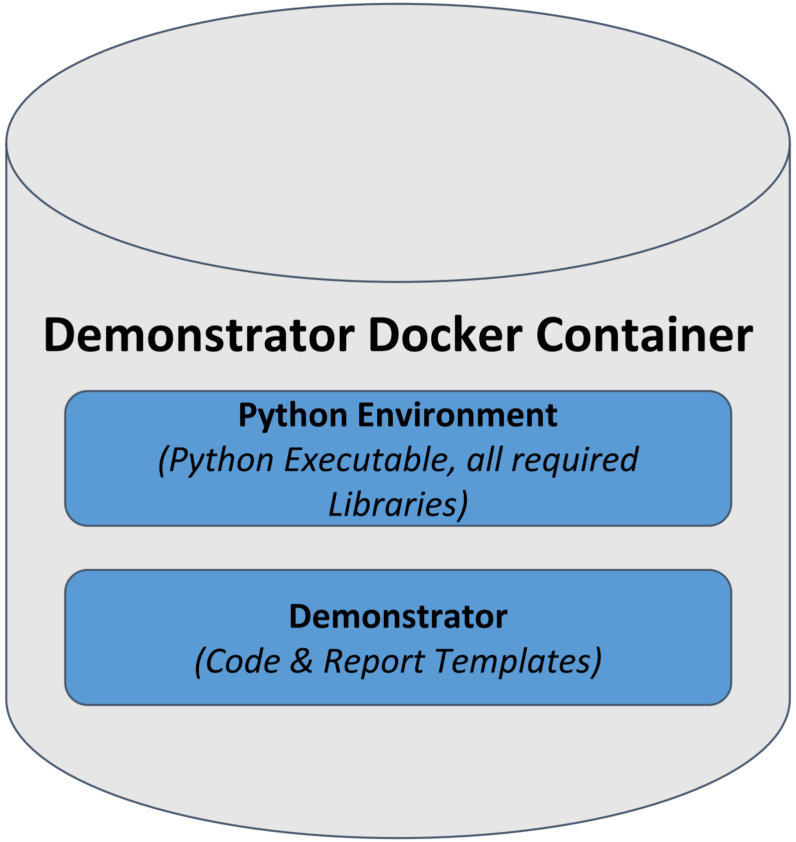
Data Quality Assurance and Workflows

1. A Repository is triggered by some event *(Commits, Merge requests, Scheduled, Hooks)*
2. An assigned runner picks up the job
3. The appropriate Container is loaded
4. Any scripting steps can be executed in the created environment



FAIR Data Spaces

Data Quality Assurance and Workflows


Demonstrator Docker Container
Python Environment
(Python Executable, all required Libraries)
Demonstrator
(Code & Report Templates)

Demonstrator

- written in Python
- based on the Frictionless standard & library
- Available as a “pippable” library

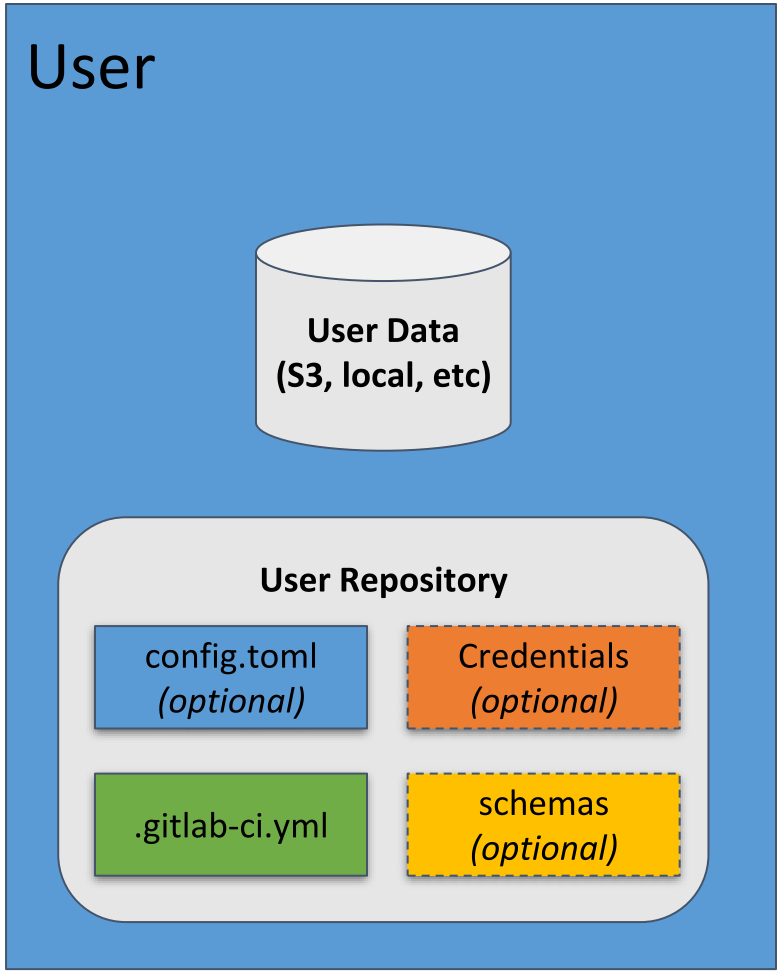
Provided as a Docker Container

- Hosted on git.rwth-aachen.de

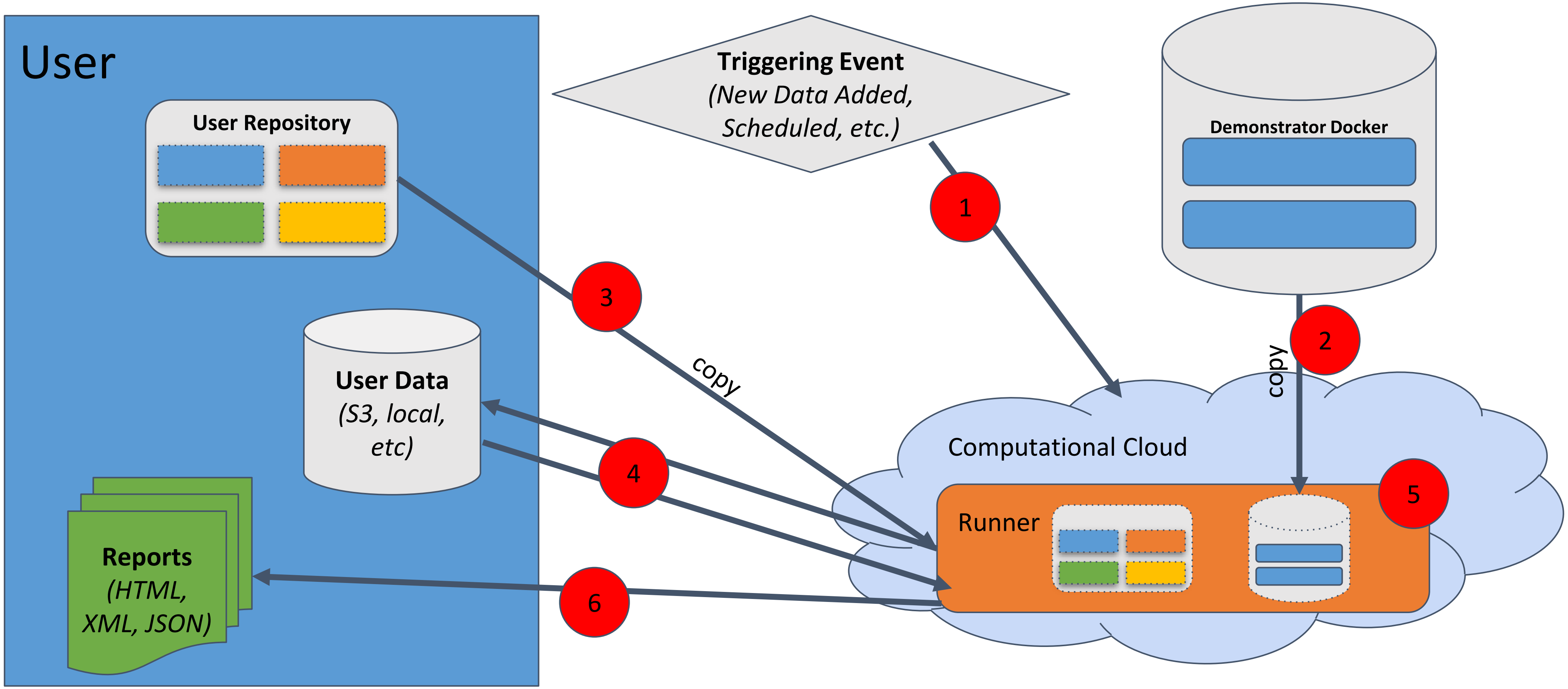
Data Quality Assurance and Workflows

Maintained by the User:

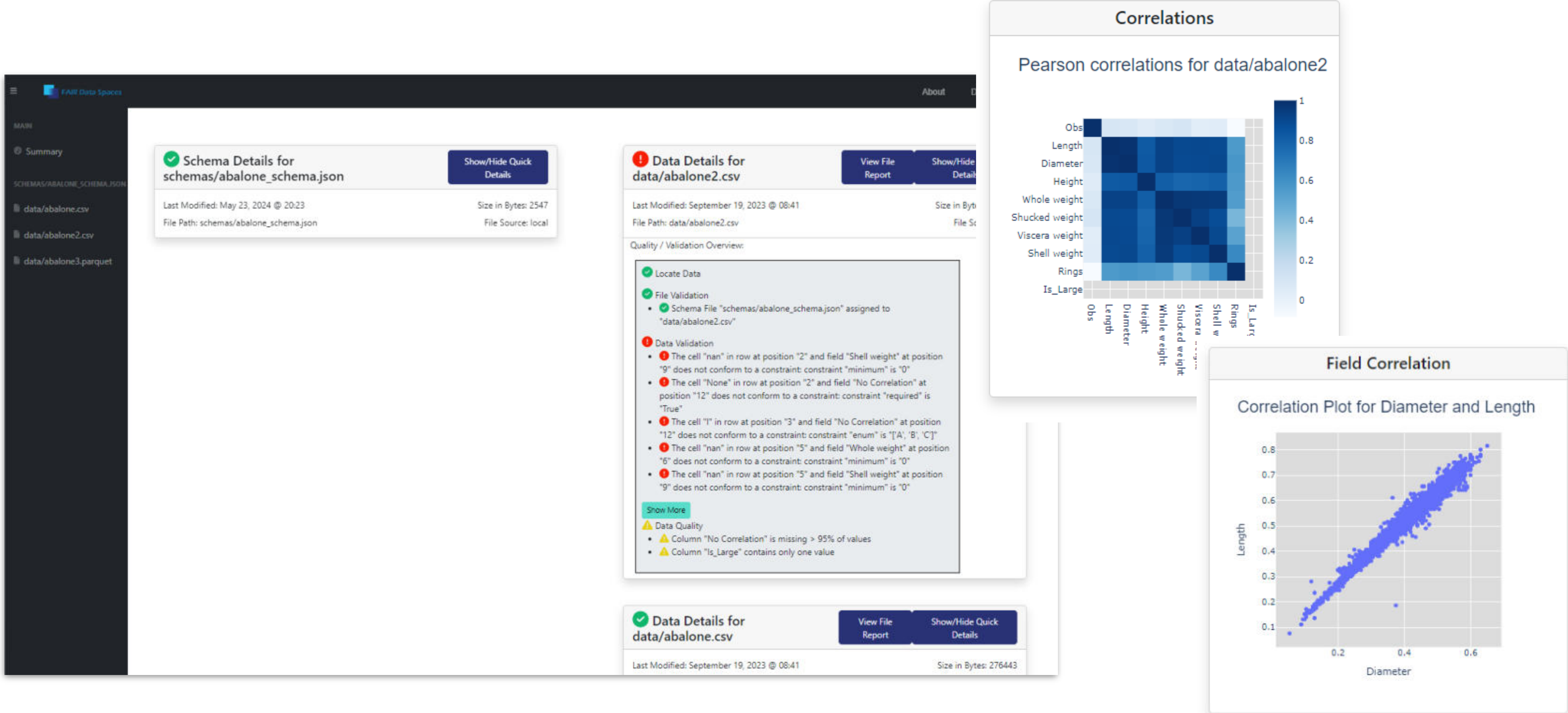
- Data to be Analyzed
  - Tabular
- A GitLab repository
  - CI/CD Script
  - Config file *(optional)*
  - Access Credentials to the data *(optional)*
  - Data Schemas *(optional, can also be stored with the data)*



Data Quality Assurance and Workflows



Data Quality Assurance and Workflows



✓ Schema Details for schemas/abalone_schema.json

Last Modified: May 23, 2024 @ 20:23 Size in Bytes: 2547

File Path: schemas/abalone_schema.json File Source: local

! Data Details for data/abalone2.csv

Last Modified: September 19, 2023 @ 08:41 Size in Byt

File Path: data/abalone2.csv File Sc

Quality / Validation Overview:

✓ Locate Data

✓ File Validation

- ✓ Schema File "schemas/abalone_schema.json" assigned to "data/abalone2.csv"

! Data Validation

- ! The cell "nan" in row at position "2" and field "Shell weight" at position "9" does not conform to a constraint: constraint "minimum" is "0"
- ! The cell "None" in row at position "2" and field "No Correlation" at position "12" does not conform to a constraint: constraint "required" is "True"
- ! The cell "I" in row at position "3" and field "No Correlation" at position "12" does not conform to a constraint: constraint "enum" is "['A', 'B', 'C']"
- ! The cell "nan" in row at position "5" and field "Whole weight" at position "6" does not conform to a constraint: constraint "minimum" is "0"
- ! The cell "nan" in row at position "5" and field "Shell weight" at position "9" does not conform to a constraint: constraint "minimum" is "0"

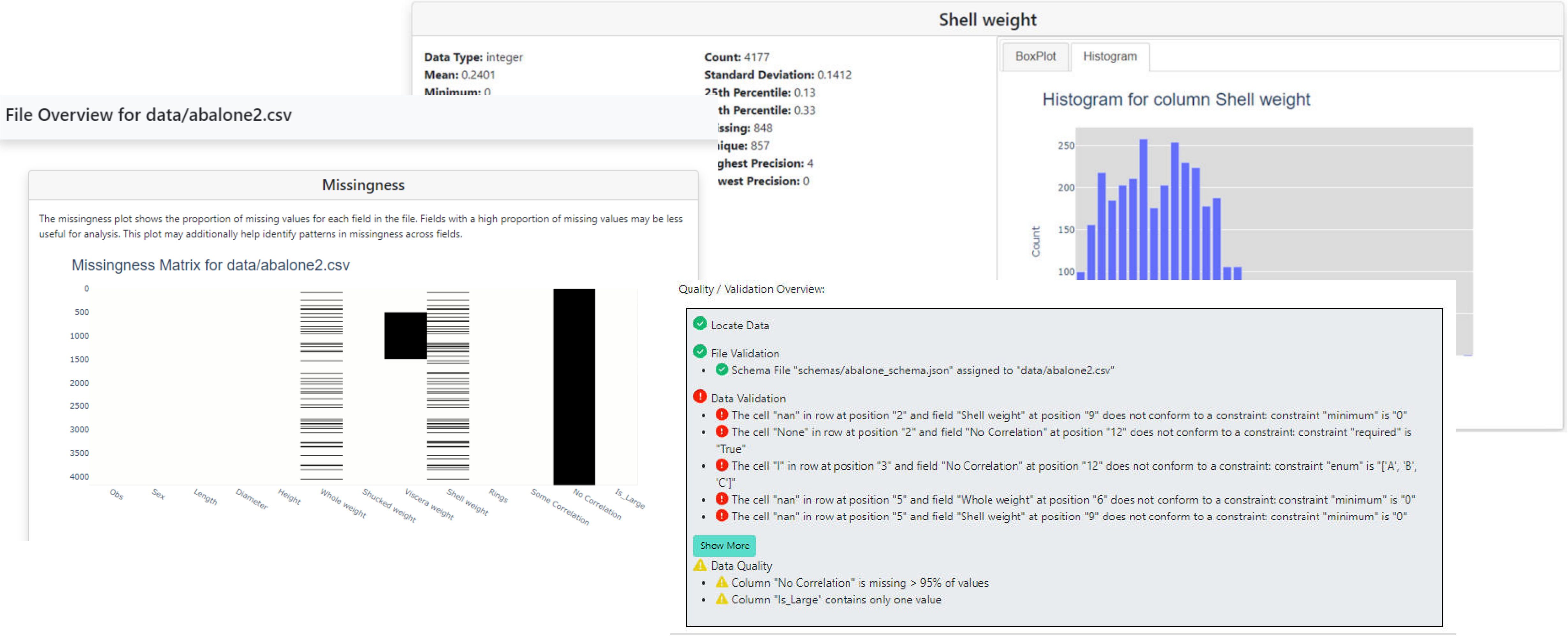
⚠ Data Quality

- ⚠ Column "No Correlation" is missing > 95% of values
- ⚠ Column "Is_Large" contains only one value

✓ Data Details for data/abalone.csv

Last Modified: September 19, 2023 @ 08:41 Size in Bytes: 276443

Data Quality Assurance and Workflows



Shell weight

Data Type: integer
Mean: 0.2401
Minimum: 0

Count: 4177
Standard Deviation: 0.1412
25th Percentile: 0.13
th Percentile: 0.33
issing: 848
ique: 857
ghest Precision: 4
west Precision: 0

File Overview for data/abalone2.csv

Missingness

The missingness plot shows the proportion of missing values for each field in the file. Fields with a high proportion of missing values may be less useful for analysis. This plot may additionally help identify patterns in missingness across fields.

Quality / Validation Overview:

✓ Locate Data

✓ File Validation

- ✓ Schema File "schemas/abalone_schema.json" assigned to "data/abalone2.csv"

! Data Validation

- ! The cell "nan" in row at position "2" and field "Shell weight" at position "9" does not conform to a constraint: constraint "minimum" is "0"
- ! The cell "None" in row at position "2" and field "No Correlation" at position "12" does not conform to a constraint: constraint "required" is "True"
- ! The cell "I" in row at position "3" and field "No Correlation" at position "12" does not conform to a constraint: constraint "enum" is "['A', 'B', 'C']"
- ! The cell "nan" in row at position "5" and field "Whole weight" at position "6" does not conform to a constraint: constraint "minimum" is "0"
- ! The cell "nan" in row at position "5" and field "Shell weight" at position "9" does not conform to a constraint: constraint "minimum" is "0"

⚠ Data Quality

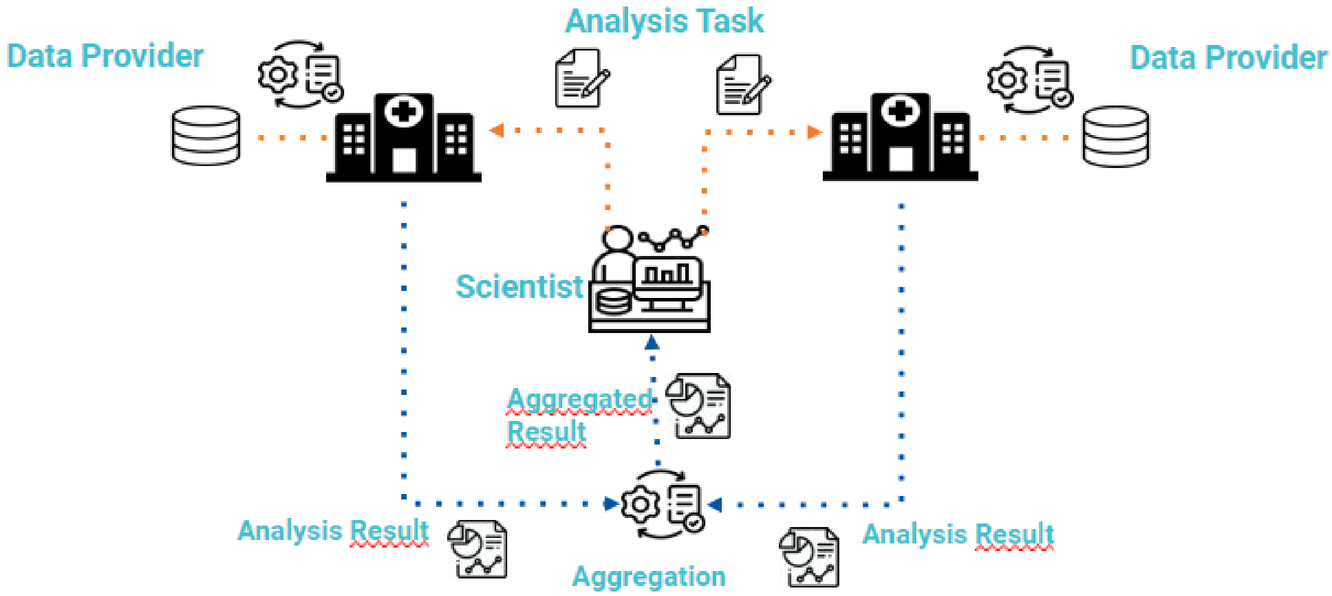
- ⚠ Column "No Correlation" is missing > 95% of values
- ⚠ Column "Is_Large" contains only one value

Demonstrator 4.3:
Cross-Platform FAIR Data Analysis PADME PHT

Yeliz Ucer Yediel, Muhammad Hamza Akhdar (Fraunhofer FIT),
Macedo Maia, Toralf Kirsten (University of Leipzig),
Mehrshad Jaberansary, Oya Beyan (University of Cologne)

Cross-Platform FAIR Data Analysis PADME PHT

- Idea: “Bring the algorithms to the data” by using Distributed Analytics (DA)
- Benefits:
  - The data remains in the control of the data providers
  - Research can leverage otherwise inaccessible data
  - The results are made more robust by incorporating a variety of datasets.
- Provides ecosystem from the first idea to the analysis results
  - Central Components: Playground, Train Creator, Train Store House, Train Requester,
  - Client Software: PHT Station



PADME in a Nutshell

https://padme-analytics.de/ , https://docs.padme-analytics.de/

- Implementation of the PHT/FL concepts by using FAIR standards
- Result of a collaboration between four research institutes

- Based on containerization technologies (www.docker.com), deployed on Kubernetes env.

- Benefits:
  - Operating system agnostic
  - Data source and data structure agnostic
  - Programming-language agnostic

PHT PADME and EDC Integration

Use Case 1: PHT as a Data Provider

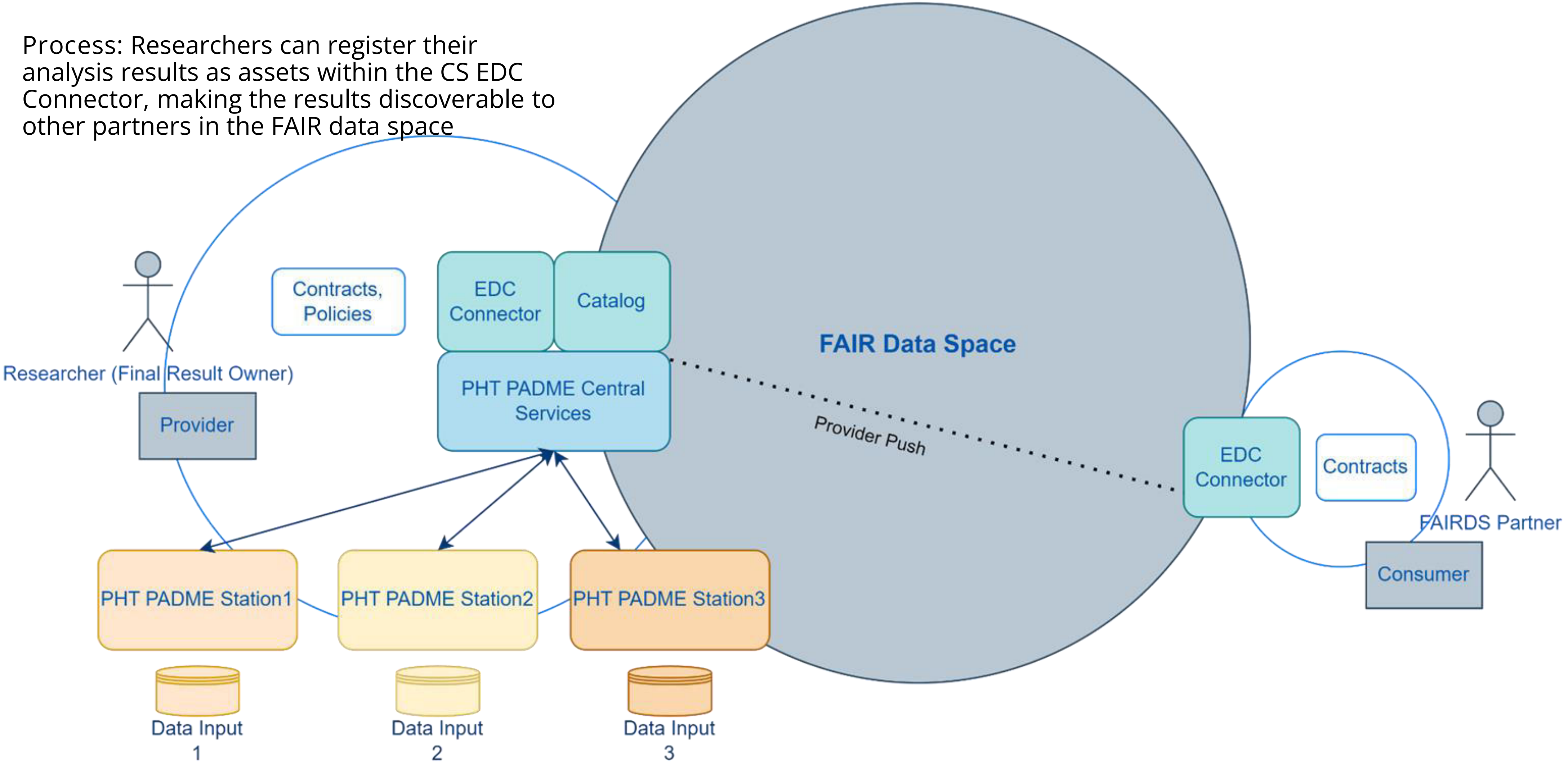
- Objective: Enable sharing of analysis results using EDC connector
- EDC Data Transfer Mode : Provider Push
- Process: Researchers can register their analysis results as assets within the CS EDC Connector, making the results discoverable to other partners in the FAIR data space

Use case 2: PHT as a Data Consumer

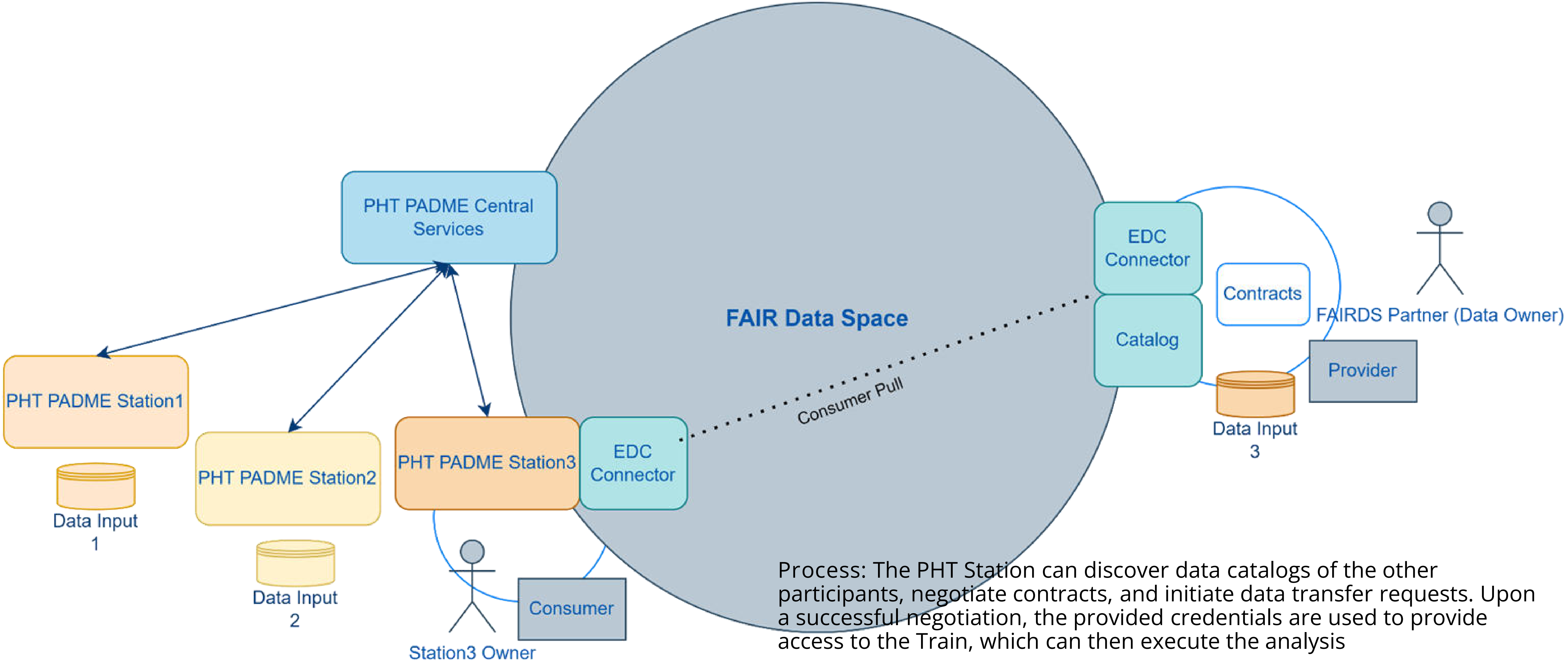
- Objective: Enable the PHT Station to consume data from other providers within the FAIR Data Space
- EDC Data Transfer Mode : Consumer Pull
- Process: The PHT Station can discover data catalogs of the other participants, negotiate contracts, and initiate data transfer requests. Upon a successful negotiation, the provided credentials are used to provide access to the Train, which can then execute the analysis

EDC Integration into CS - Provider Push Scenario

Process: Researchers can register their analysis results as assets within the CS EDC Connector, making the results discoverable to other partners in the FAIR data space



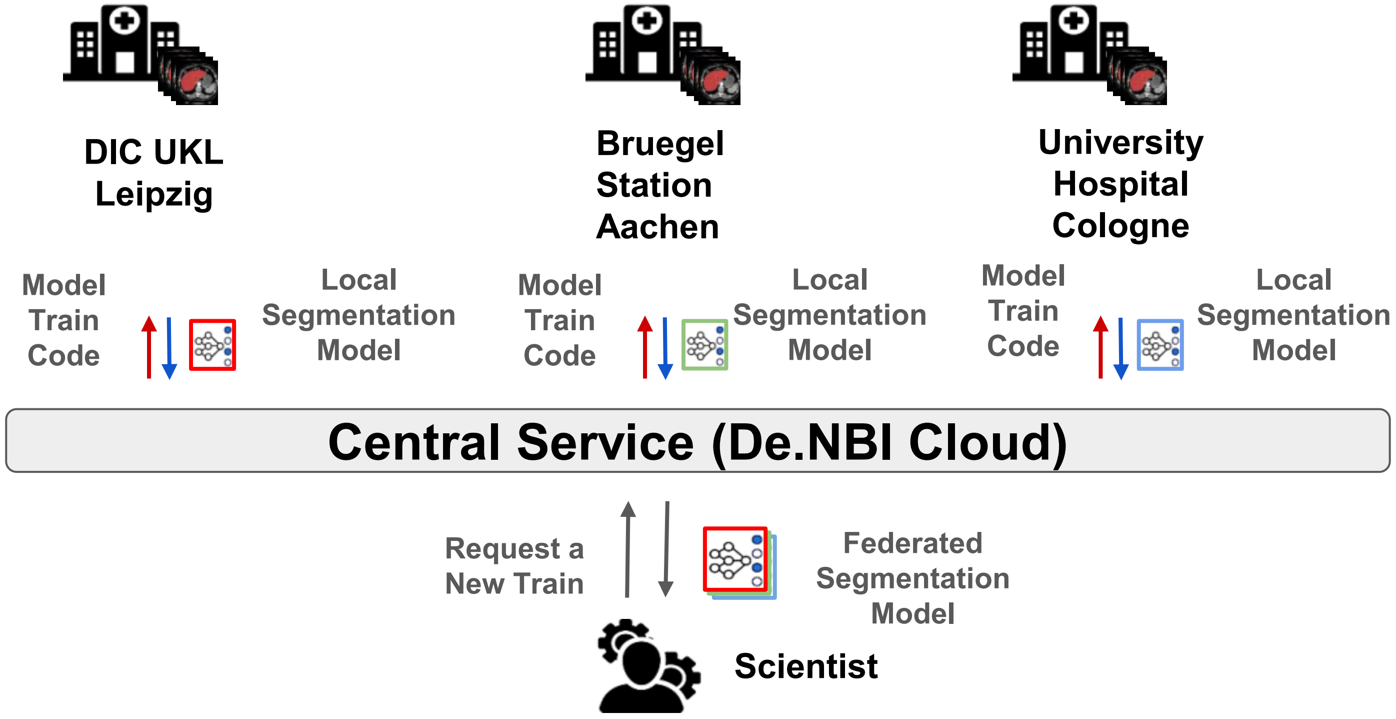
EDC Integration into Station
Consumer Pull Scenario



Process: The PHT Station can discover data catalogs of the other participants, negotiate contracts, and initiate data transfer requests. Upon a successful negotiation, the provided credentials are used to provide access to the Train, which can then execute the analysis

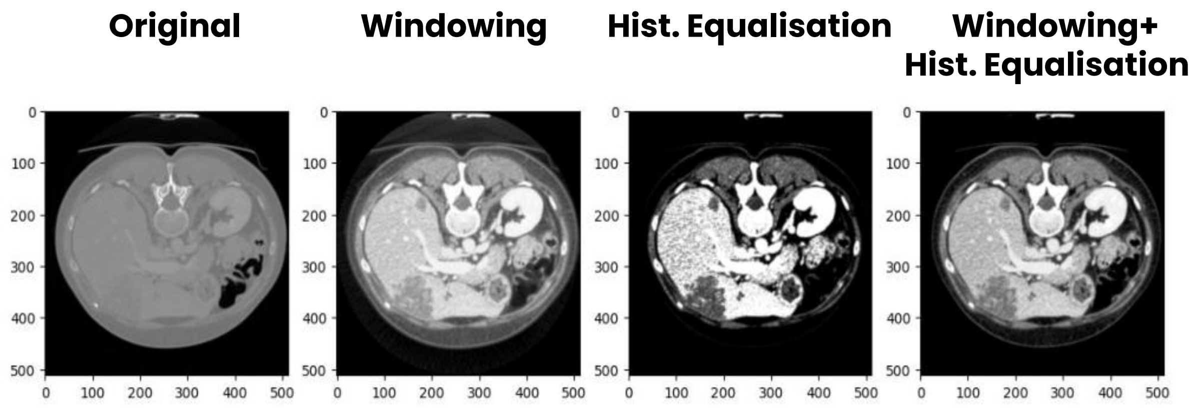
Federated Learning over Multiple PADME PHT Stations

- Federated learning involves training a central model using data distributed across multiple Stations (Client/Provider) and Central Service (Server/Consumer)
- Local models are trained on each PHT Station
- Each local model are sending to Central Service
- The application of a aggregation function over each local model weights determines a federated learning
- The federated model are sending back to each Station and retrained in the next round

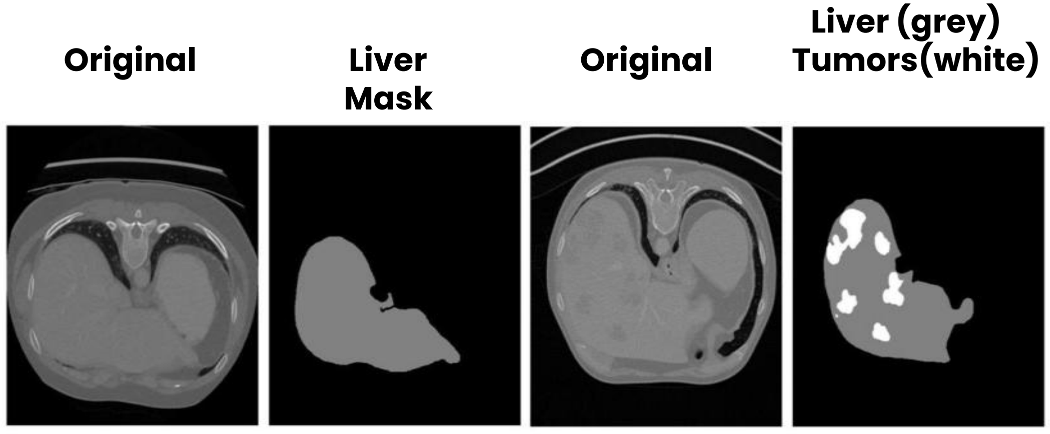


Use case: Liver Tumours Segmentation

- Problem Statement
  - Based on The American Cancer Society's:
    - About 41,630 new liver cancer cases in US were diagnosed in 2023
    - About 29,840 people have died of these cancers
  - AI-based approaches helps to early detect tumours
  - However, the data can be distributed in different sources (e.g., hospitals)
  - Data access depends on distinct rules or regulations from each data provider
- Possible Solution
  - Explore Computed Tomography (CT) scans for image segmentation
  - Federated learning models over data from multiple data providers
- Liver CT Scan Data for Segmentation:
  - The CT Liver dataset consists of 3D NIFTI images or 2D DICOM scans
  - Segmentation masks are the labels
  - Segmentation models for medical scans:
    - UNET
    - nn-UNET
    - Dense-UNET



Original **Windowing** **Hist. Equalisation** **Windowing+ Hist. Equalisation**



Original **Liver Mask** **Original** **Liver (grey) Tumors(white)**

No cancer **With cancer**

All project participants

Atos

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EMBL

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Fraunhofer

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

nfdi
Nationale
Forschungsdaten
Infrastruktur

Philipps Universität Marburg

RWTH AACHEN
UNIVERSITY

TIB
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

UNIVERSITÄTS
KLINIKUM
HEIDELBERG

UNIVERSITÄT
LEIPZIG

Universität
zu Köln

WWU
MÜNSTER

ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Thank you for your interest!

Contact: Christoph Lange & Zeyd Boukhers, christoph.lange-bever@fit.fraunhofer.de, zeyd.boukhers@fit.fraunhofer.de

use subject “Open Call”

Stay in touch:

🌐 www.nfdi.de/fair-data-spaces

@FAIRDataSpaces

#FAIRDataSpaces

Community Newsletter: https://www.nfdi.de/newsletter-abo/

Wiki: https://github.com/FAIR-DS4NFDI/FAIR-DSWiki/wiki

FAIR Data Spaces