

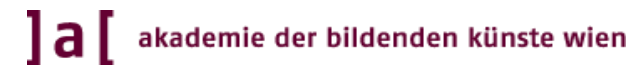


FAIR DATA  
AUSTRIA

# FDA-DBRepo

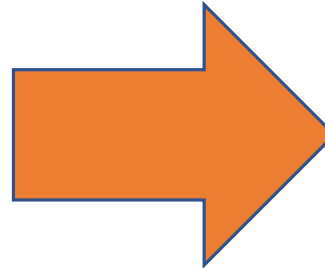
*A DATA PRESERVATION REPOSITORY SUPPORTING FAIR PRINCIPLES,  
DATA VERSIONING AND REPRODUCIBLE QUERIES*

1st of July 2021



## DBs in research: Common disadvantages

- Little attention to FAIRness
- Local DBs
  - Administration skills
  - No data versioning
  - No metadata
- Only populated and queried during project life-time
- Deposited post-project (DB-dumps) or exported in a file-based repositories



## Hence:

- High risk:  
*incompatibility; data is not machine-readable, understandable and reusable; hard to preserve, etc.*
- Lack of reproducibility
- Unusable DB dumps after the end of a project
- Does not work for:  
*live DBs, continuously growing DBs, etc.*

# FDA-DBrepo – VISION – 1

- Private cloud hosted repository
- Database administration is outsourced to repository infrastructure (and supported by experts, e.g. data stewards)
- Supports different levels of SQL-knowledge
- Metadata is generated and exposed  
→ FAIRness



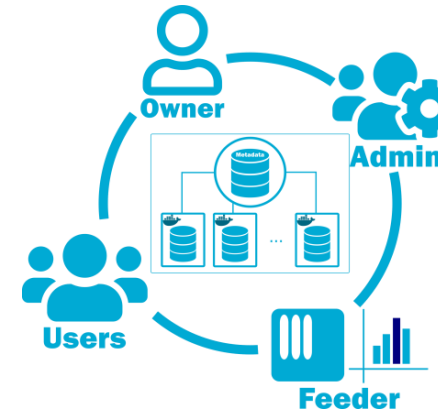
# FDA-DBrepo – VISION – 2

## Databases are:

- created directly in repository framework
- populated and used within repository
- searchable and reusable via metadata

## Data is:

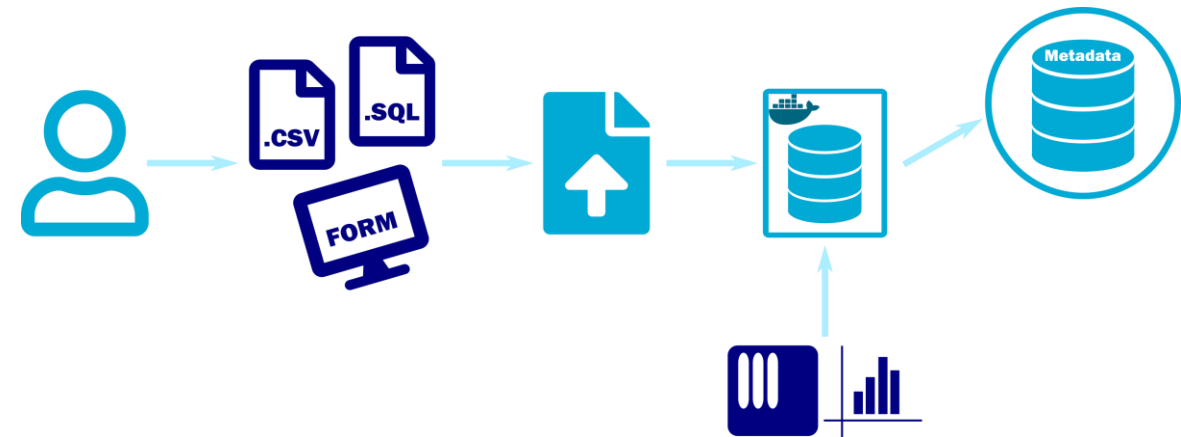
- reproducible due to versioning and time-stamping
- cite-able at fine granularity



## Levels

SQL- expert

No SQL skills



# FDA-DBRepo – SET UP

- Each database is encapsulated in a Docker container  
→ flexibility, scalability
- Microservice architecture  
→ modularity, scalability
- Services for different tasks:
  - Analyze service (e.g. to propose data types and the primary key column when uploading data via CSV)
  - Query service
  - Container service
  - etc.
- RESTful API interface

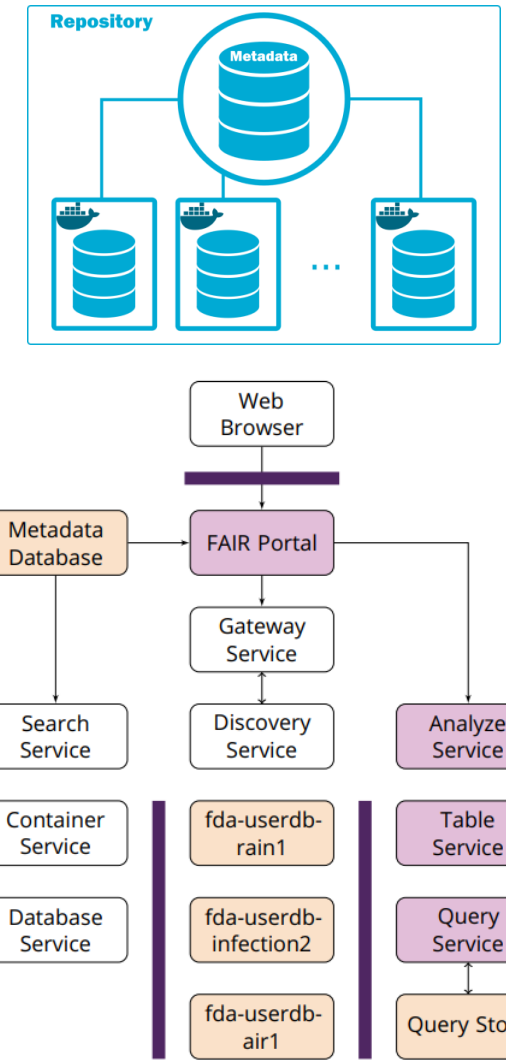


Figure 2: Overview architecture of the FDA-DBRepo infrastructure with three exemplary user databases protected by a firewall barrier around them.

# FDA-DBrepo – ANALYZE SERVICE

- Maps the metadata about databases and tables to controlled vocabulary
- Forwards the metadata to the metadataDB
- Makes suggestions on datatypes for columns within a CSV file
- Makes primary key recommendations

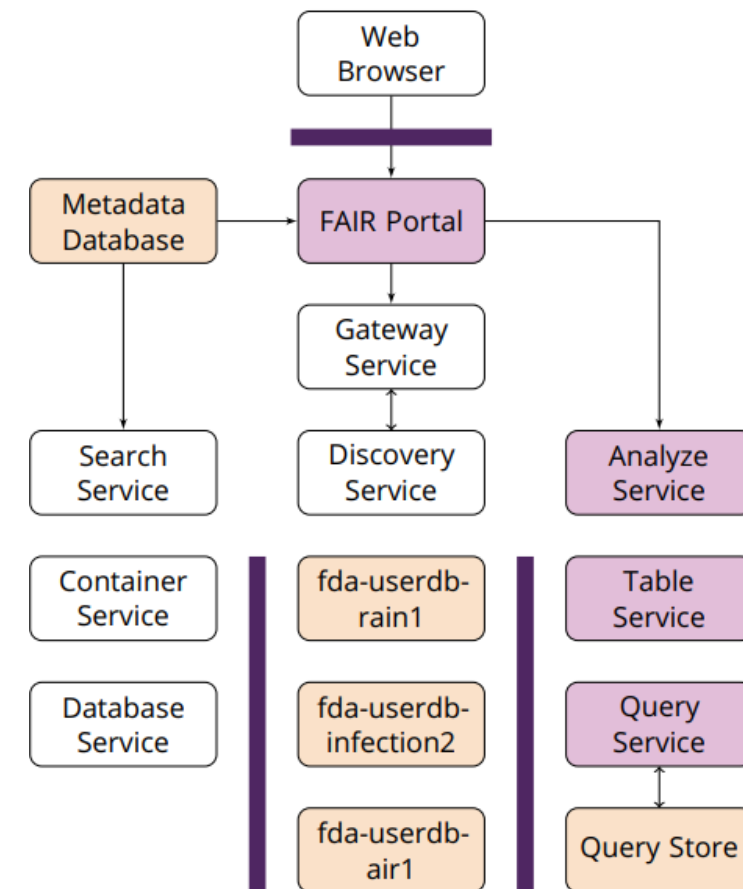


Figure 2: Overview architecture of the FDA-DBrepo infrastructure with three exemplary user databases protected by a firewall barrier around them.

# FDA-DBRepo – QUERY STORE & EXECUTION

- Each query saved in the Query Store for re-execution and citation
- Queries are normalized
- Each query is assigned a persistent identifier (e.g. DOI)
- Re-execution allows subsets to be identified and cited
- Hashing ensures correctness of results

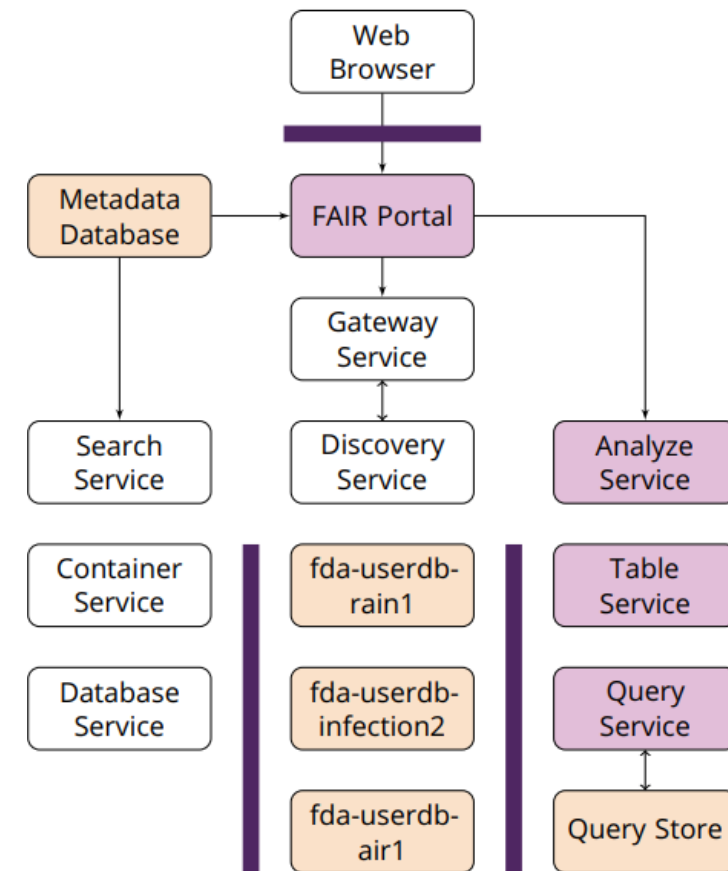


Figure 2: Overview architecture of the FDA-DBRepo infrastructure with three exemplary user databases protected by a firewall barrier around them.

# FDA-DBrepo – BENEFITS

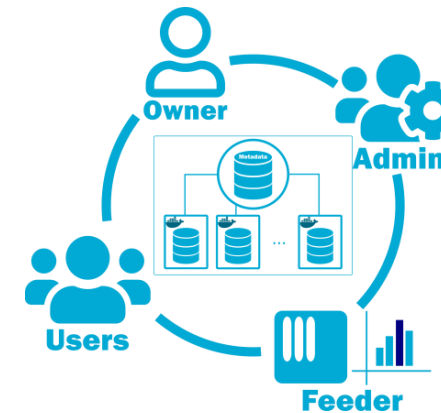
Scalability

Flexibility

Dynamic

Usability

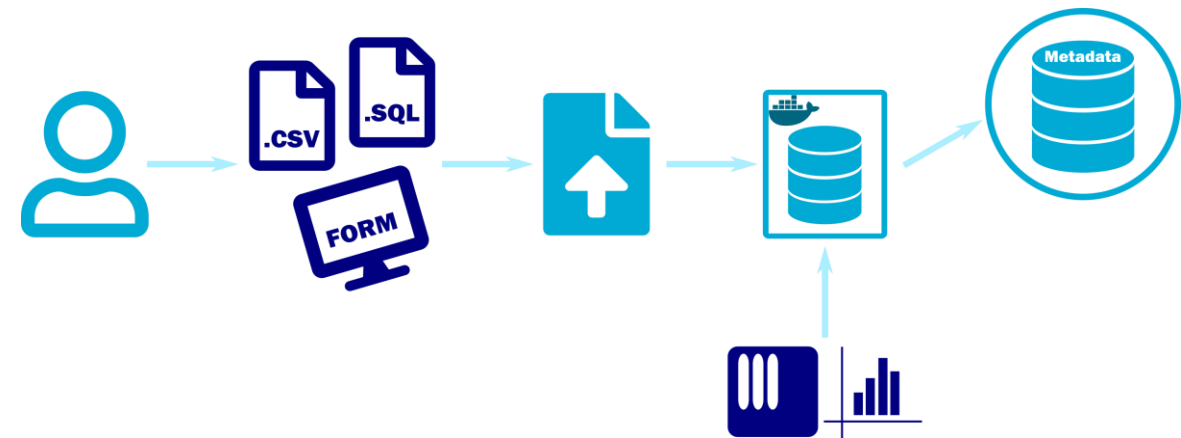
Separation of concerns

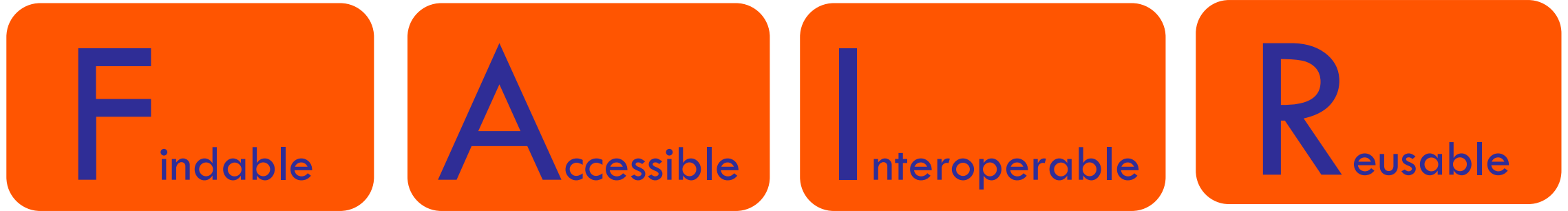


## Levels

SQL- expert

No SQL skills

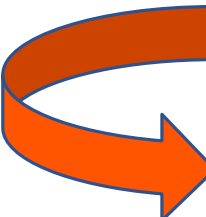




- The metaDB contains:
  - *Information about:*
    - *The data (e.g.: creator, publisher, resource type, etc.)*
    - *The database scheme (e.g.: database name, table names, attribute names)*
  - *Measurement units of numerical columns*
  - *Basic statistical information about the data*
- Data is mapped onto controlled vocabularies
- User interface available for querying the DB (SQL, faceted browsing)
- The DB can be queried via REST interface
- Dynamic data citation (RDA WGDC)
- Updates are versioned and time-stamped

# FDA-DBrepo – DATA VERSIONING

- Tracks changes in the DB
- Creation/deletion timestamp in each table
- Update is handled with new entry in table



Id	Firstname	Lastname	Created	Deleted
1	max	mustermann	2021-05-05 12:00:00	2021-05-06 16:00:00
2	Martin	Weise	2021-05-05 12:00:01	Null
1	Max	Mustermann	2021-05-06 16:00:00	Null
3	Moritz	Staudinger	2021-05-06 16:00:00	Null
4	Cornelia	Michlits	2021-05-06 16:00:00	Null

# TEAM



Andreas Rauber



Cornelia Michlits



Martin Weise



Moritz Staudinger



Raman Ganguly

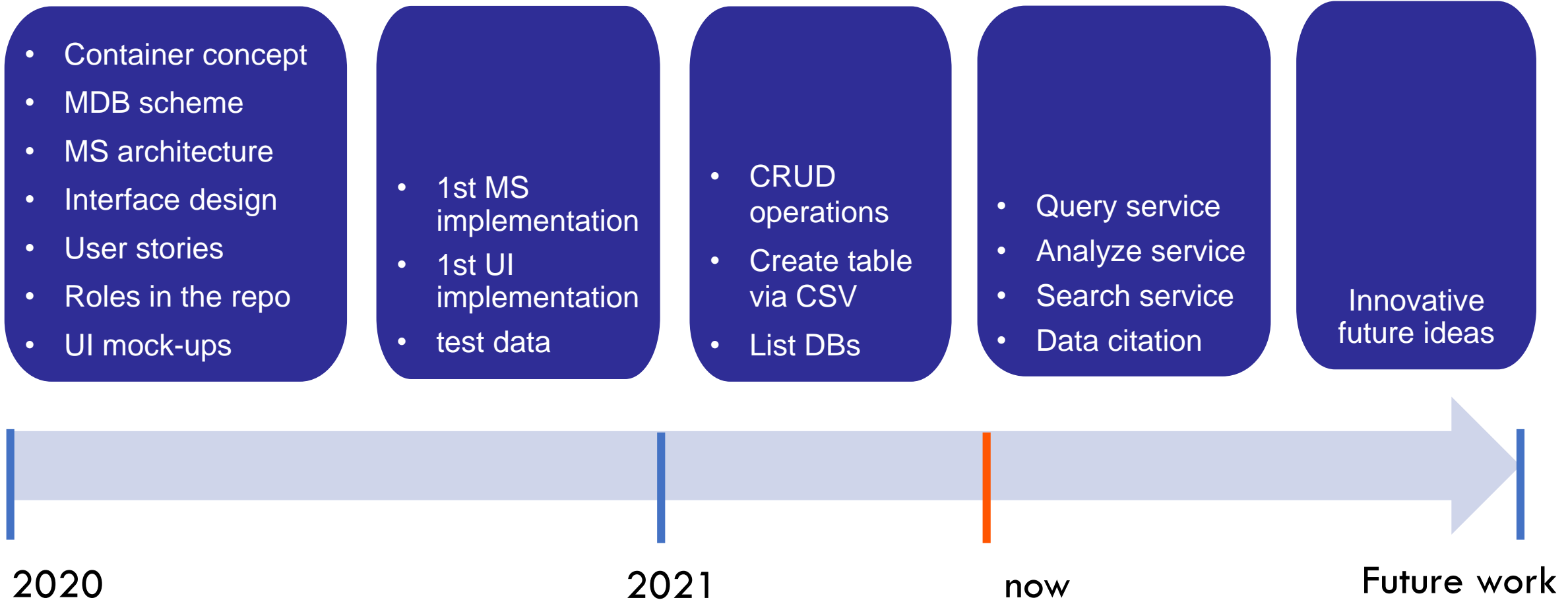


Kirill Stytsenko



Eva Gergely

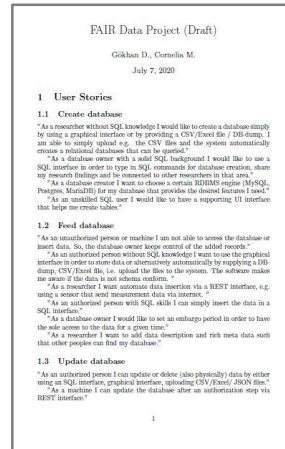
# TIMELINE



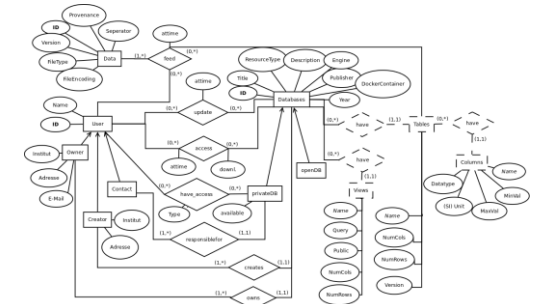
## Levels

SQL- expert

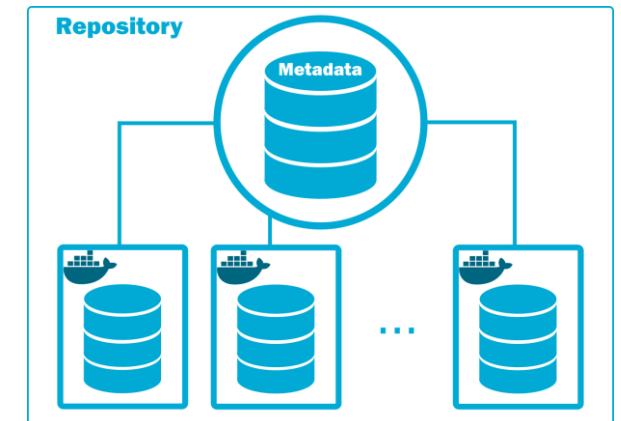
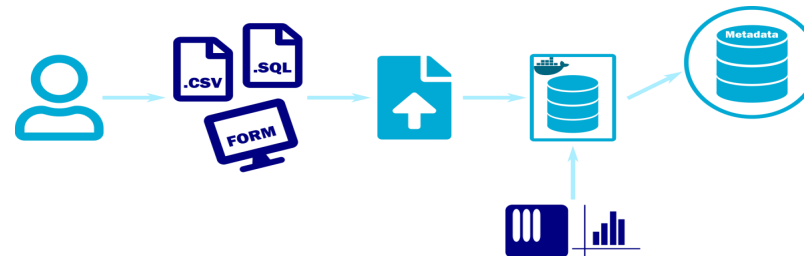
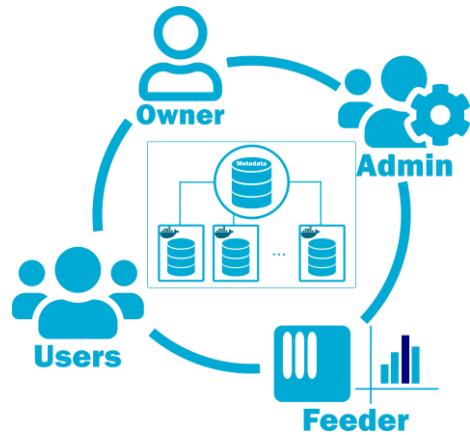
No SQL skills



FAIR DATA AUSTRIA



# THANK YOU!



## KONTAKT

<https://github.com/fair-data-austria/dbrepo>

[rauber@ifs.tuwien.ac.at](mailto:rauber@ifs.tuwien.ac.at)

[raman.ganguly@univie.ac.at](mailto:raman.ganguly@univie.ac.at)