FAIR technical infrastructures currently in development

Andreas Rauber
TU Wien

TU WIEN Informatics ifs

FAIR technical infrastructures currently in development (with quite some personal bias)

Andreas Rauber
TU Wien

TU WIEN Informatics ifs

Outline

- The ***F*** in FAIR
- FAIR ≠ Open

Findability

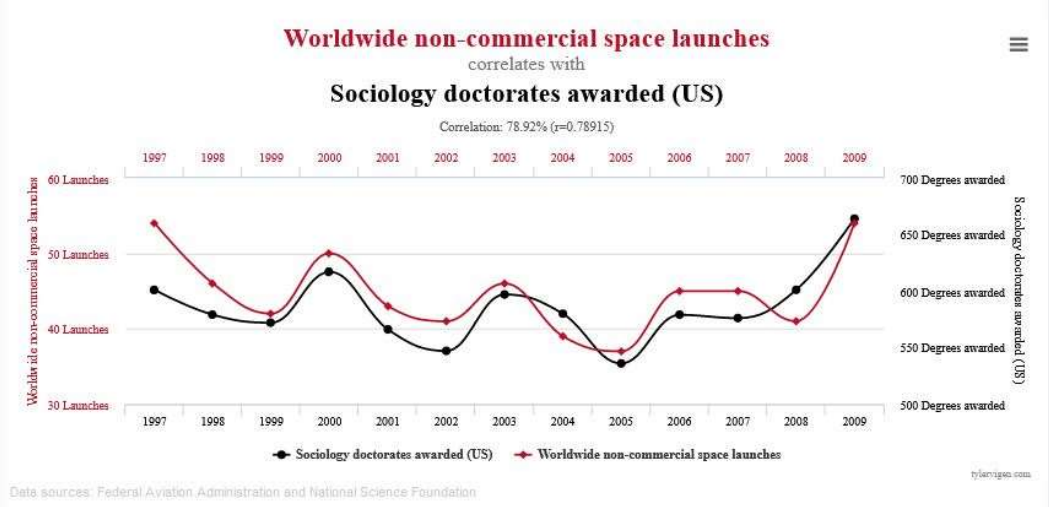
- Being able to find stuff (data)
- “Google for Data” – what do we need?
  - Description of the data
  - A full-text search engine (e.g. Elastic Search)
- Done?

Findability

- Being able to find stuff (data)
- “Google for Data” – what do we need?
  - Description of the data
  - A full-text search engine (e.g. Elastic Search)
- Done? – Well…
- FAIRness is predominantly for machines!
- Machines are not good in understanding full-text searches
- Results need to be machine-actionable
- Humans benefit from more structured approaches
- Especially when looking for interdisciplinary data
- Search for data is different…

Findability

- Example scenarios: Databases that have data on…
  - Temperature measurements
  - Temp./temperatuur/hitastig/lämpötila/θερμοκρασία/…
  - In the range of -40 to 45 Celsius; 18-22 millikelvin
  - Measured in locations above 66°34′N, along the Danube,…
  - That show a bi-modal distribution
  - That have a high correlation with the following dataset. (just for fun: https://www.tylervigen.com/spurious-correlations)
  - That has not been low-pass filtered…
  - That I may re-use…



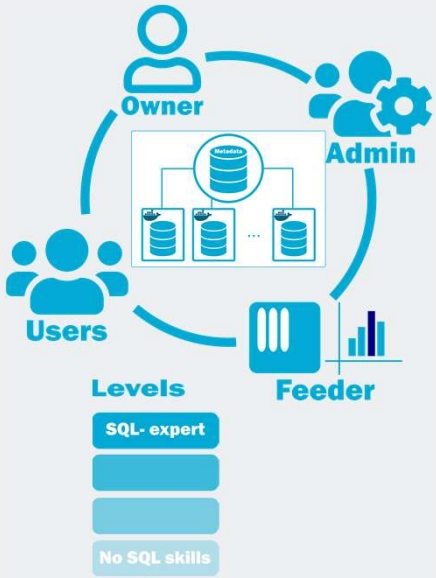
Findability

- Need comprehensive metadata on data
- In structured form
- Ideally generated automatically

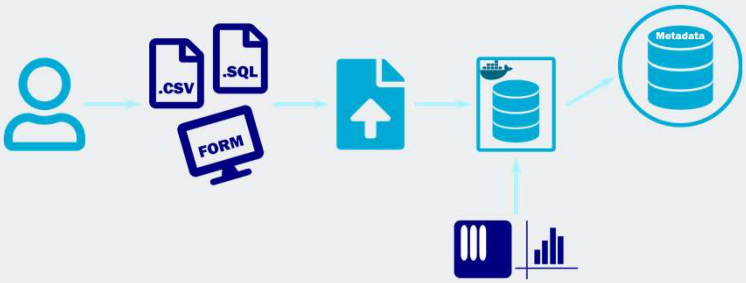
- Mapping to controlled vocabularies / ontologies
  - Attribute semantics
  - Measurement units
  - Conversion rules (primarily for ***I***, but also for ***F***)
- Assist user in mappings (semi-automatic)
- Statistical properties

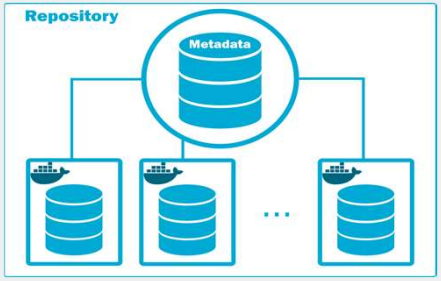
- Provenance
- Licenses (primarily for ***R***, but also for ***F***)

Findability for FDA DB-Repo: Vision



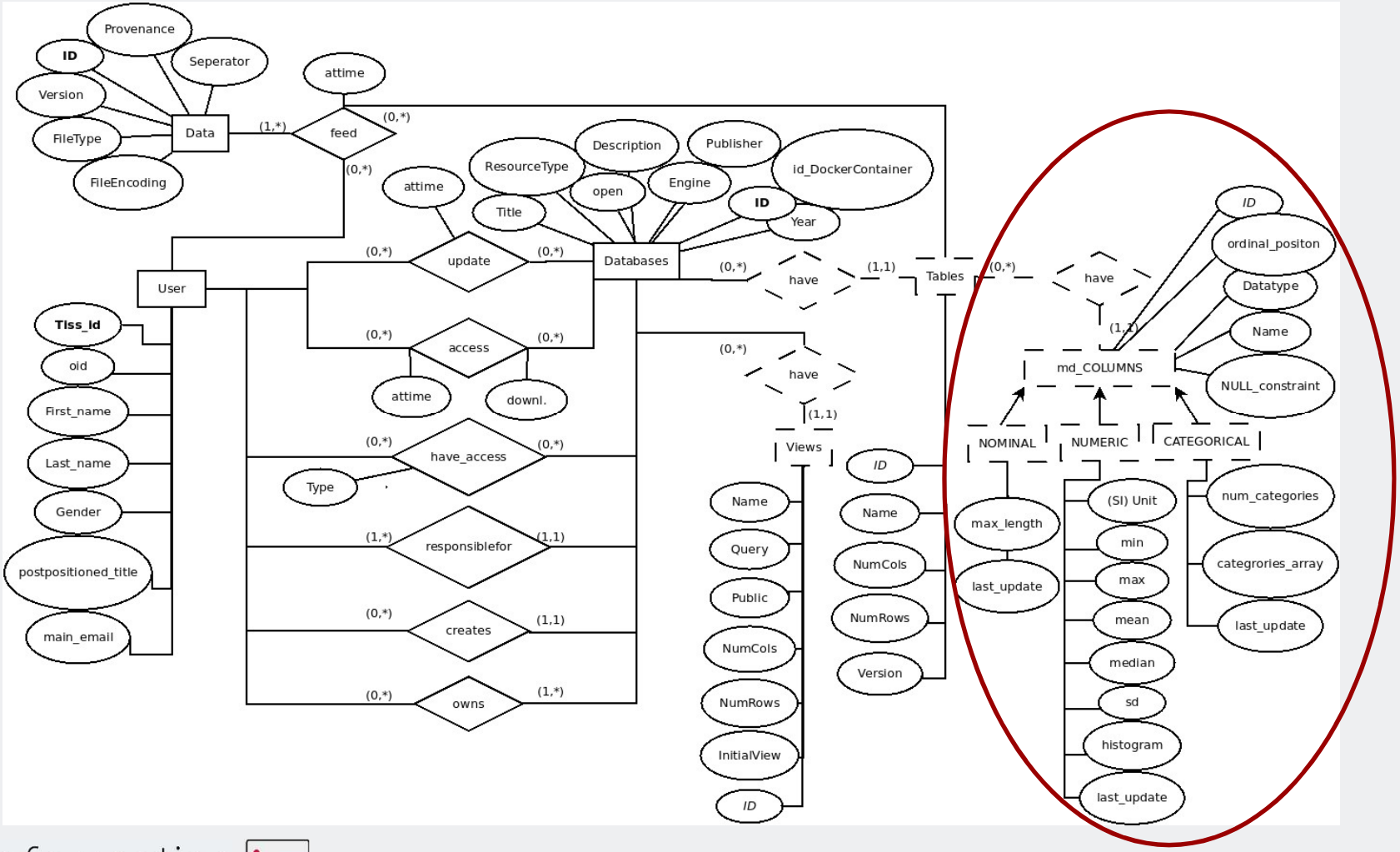
- Structured data
- Private cloud hosted relational databases
- DB is created directly in repository framework
- DB is populated and used within repository
- Metadata is generated and exposed
- Databases and data are searchable
- Data is versioned & time-stamped: reproducibility, re-use, provenance
- Data is cite-able at arbitrary levels of granularity (RDA WGDC recommendations)
- Data Management outsourced to repository infrastructure: easier for researchers, higher quality data mgt, higher security, …





Findability

- Metadata DB – initial schema (+ API, schema for responses, …)



Outline

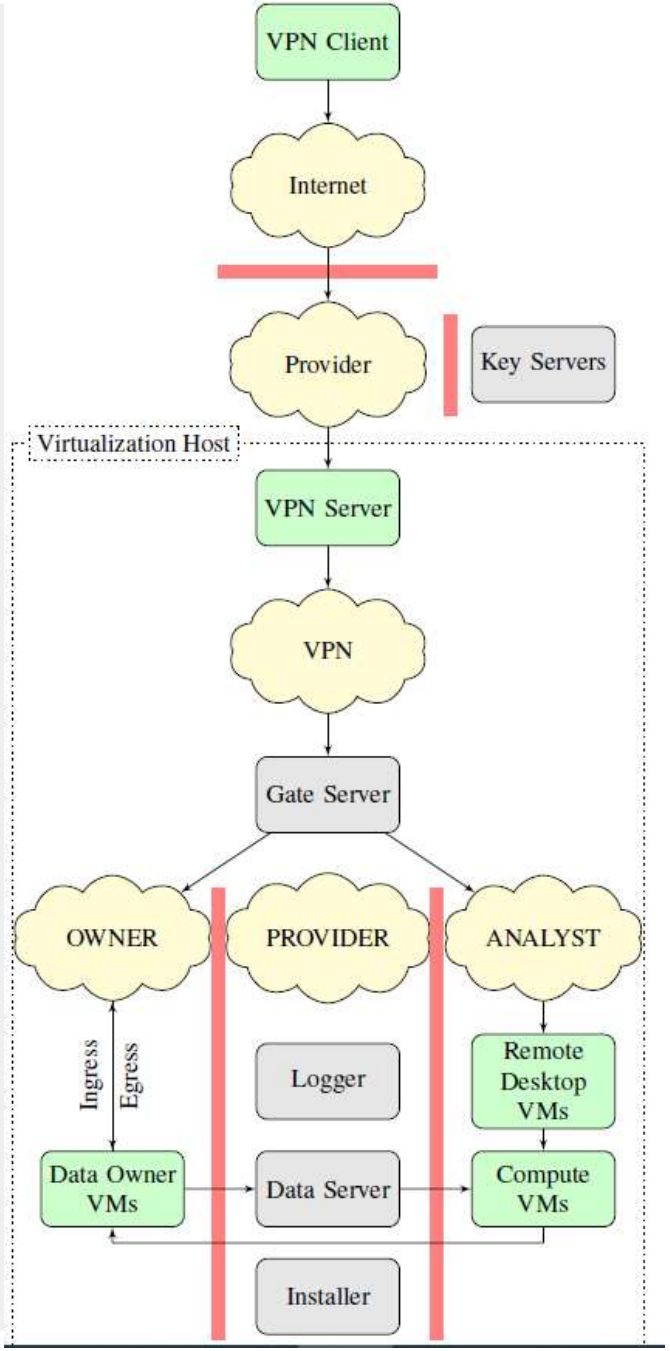
- The ***F*** in FAIR
- FAIR ≠ Open

FAIR ≠ Open

- FAIR data does not imply/require Open data!
- Accessibility means it is clear HOW data can be accessed (again, ideally also for machines!)
- An extreme example: FAIRness for medical data (not anonymized, not pre-aggregated, …)
- Full control by data owner on WHO has access WHEN, for WHICH purpose
  - Only selected people may see the data
  - It may only be used for analysing pre-defined questions
  - Access only during limited periods of time
  - Data may not be transferred to anybody
- →Data visiting instead of data sharing

Secure Data Infrastructure

- **FAIRness for "closed" data!**
- Sensitive Data (privacy, IPR, …)
- **Data Visiting** instead of Data Sharing
- **Data owner maintains full control over data:**
  - Access by **whom**, for which period of **time**,
  - to which **subset** of data
  - for which analysis **goal** / research question
- Data infrastructure acts as data processor
- Secured IT system
  - Air-gapped (virtual) machines with data excerpts
  - Access solely via remote desktop
  - Complete monitoring of all interactions
- Controlled processes
- Data identification, dynamic citation, reproducibility
- Open source reference implementation (**OSSDIP**)



TU WIEN Informatics ifs

Secure Data Infrastructure (OSSDIP) - Processes

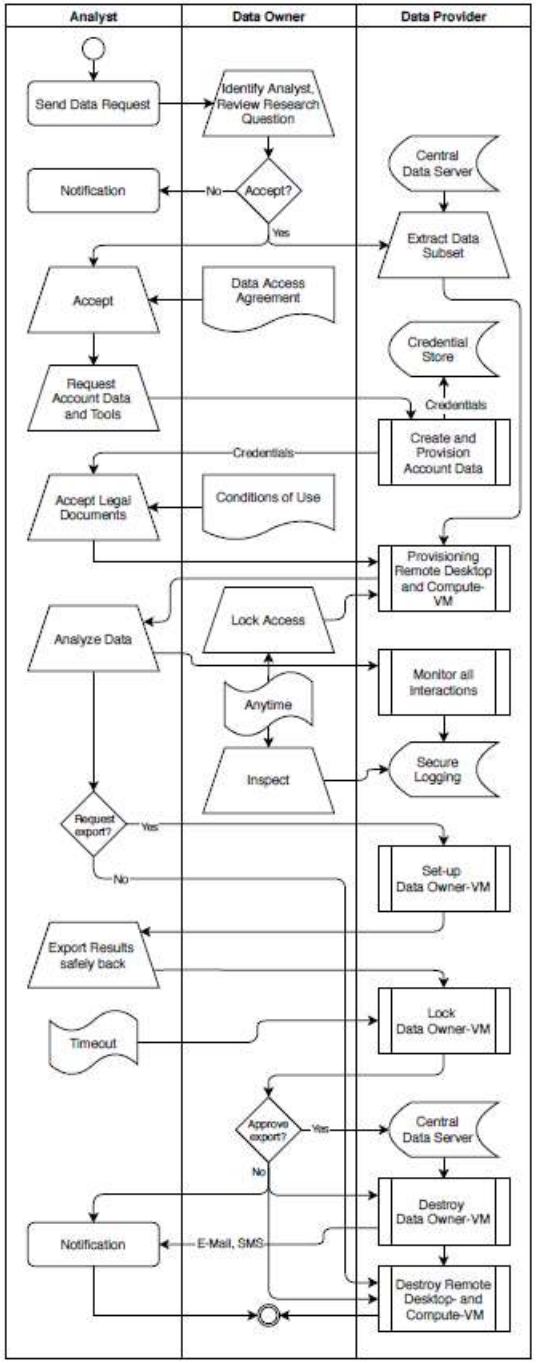
Data ingest

1. Agreement on data delivery (Data processing agreement) incl. metadata:
   - List of attributes incl. description, primary key (FAIR)
   - Number of records
   - Data format (CSV,Separator, NULL-encoding, Boolean-encoding, Date encoding, …)
   - Short description of data set
2. Request for account (name, e-Mail, mobile phone nr. for SMS)
3. Provisioning of account data by infrastructure to provider
4. Transfer of data to specific Provider-VM by provider
5. Notification of transfer concluded
6. Transfer / ingest of data from VM to server, destruction of VM
7. Provisioning of metadata on portal (“FAIRness”)

Secure Data Infrastructure (OSSDIP)

Access (selected subset of steps)

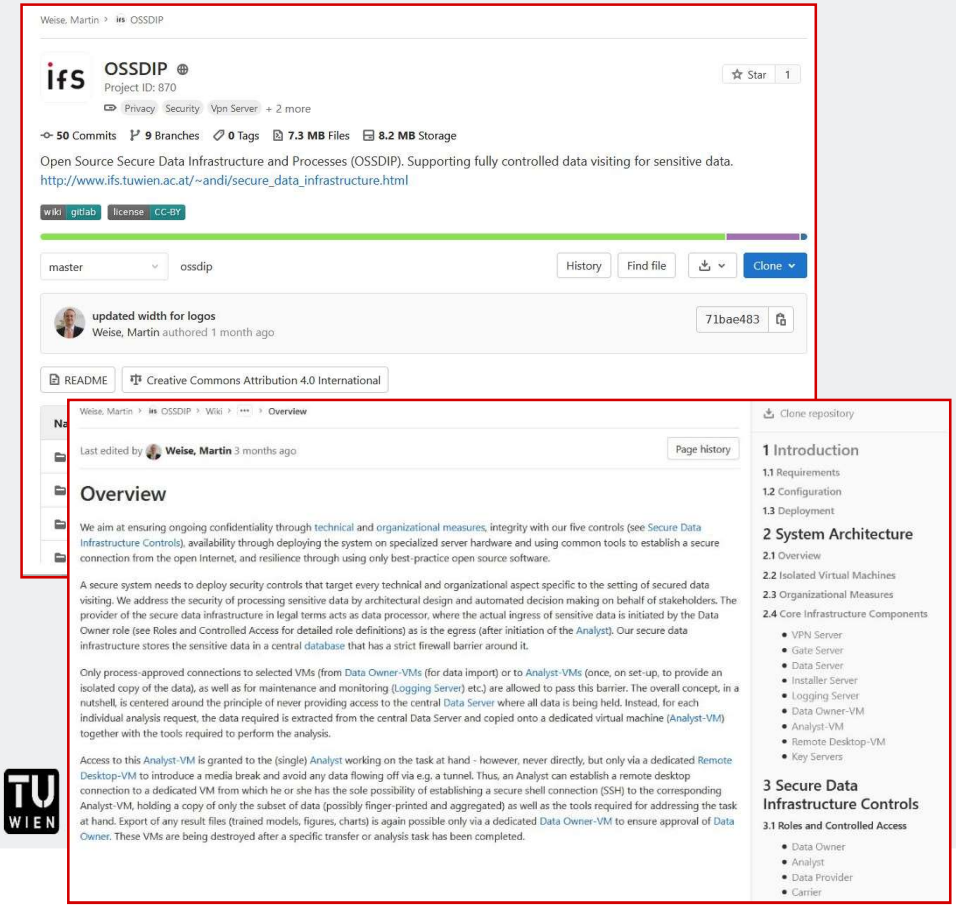
1. Researcher sends **request** to data owner (*Person, question, required data*)
2. In case of **permission** being **granted: subset of data**, at specific **aggregation level**, potentially with **fingerprint** is extracted onto a VM for a dedicated **researcher** for a dedicated **time period** to address the **question** posed
3. Legal agreement on data access permission
4. Provisioning of account info, conditions of use (no download, no de-anonymization, …)
5. Provisioning of VNC and Compute VMs with dedicated SW and data
6. Monitoring of all interactions on machine on secured logging server (log-files + video)
7. Transfer of results via dedicated Provider-VM
8. Destruction of VNC and Compute VMs

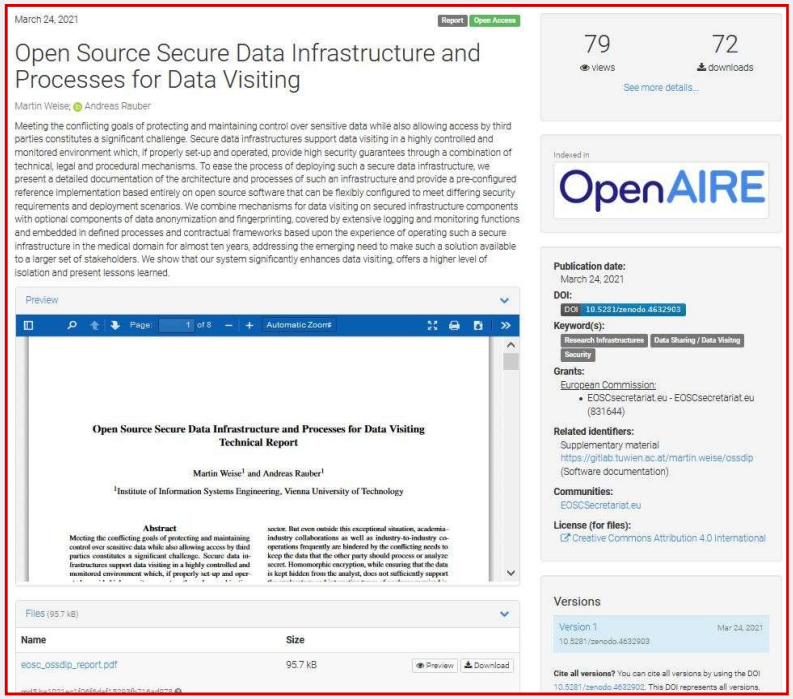


TU WIEN Informatics ifs

Secure Data Infrastructure (OSSDIP)

- (First version of) reference implementation:
  (Co-funded by EOSC-Secretariat, EOSC-Life)
  - Report: https://zenodo.org/record/4632903
  - Source: https://gitlab.tuwien.ac.at/martin.weise/ossdip





Secure Data Infrastructure (OSSDIP)

Deployment

- Easy deployment via Ansible Playbooks
- Proof-of-concept relying on virtualization

Next steps

- Exposure of Metadata DB
- Exposure of Query Store
- Different application scenarios
  - Medical / Industry settings
  - In-house settings
  - Archives, “DP-Settings”
- Further security aspects (logging, homomorphic encryption, secure multiparty computation, watermarking, …)

Conclusions

- The ***F*** in FAIR
  - Need to consider specifics of data search
  - Lots of structured information, automatically, for machines
- FAIR ≠ Open
  - Data visiting instead of data sharing
  - Non-open data can be made FAIR!
- The technical aspects are, basically, easy