

FAIR technical infrastructures  
currently in development

Andreas Rauber  
TU Wien

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

Andreas Rauber  
TU Wien

# Outline

- The ***F*** in FAIR
- FAIR  $\neq$  Open

# Findability

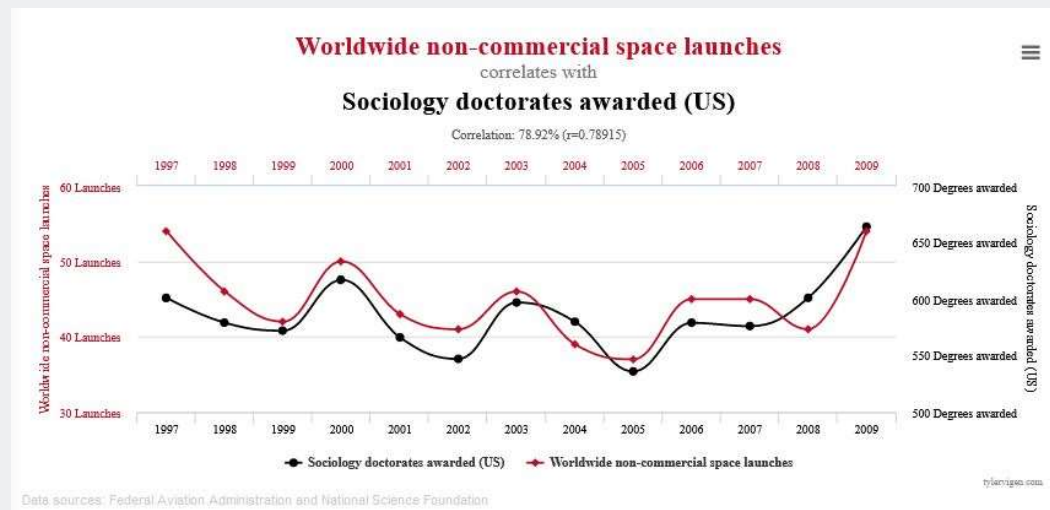
- Being able to find stuff (data)
- “Google for Data” – what do we need?
  - Description of the data
  - A full-text search engine (e.g. Elastic Search)
- Done?

# 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...

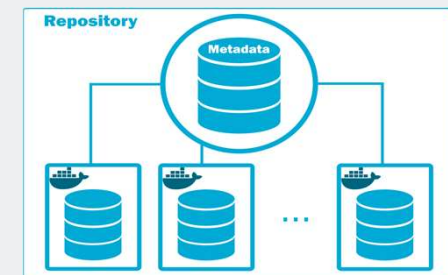
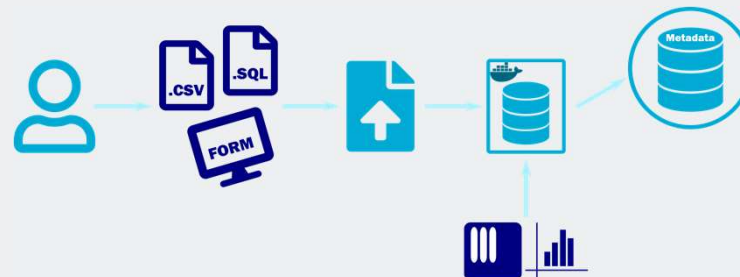
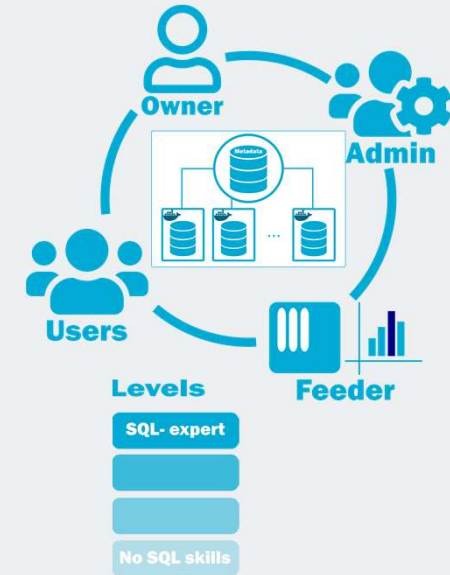


## Findability

- 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

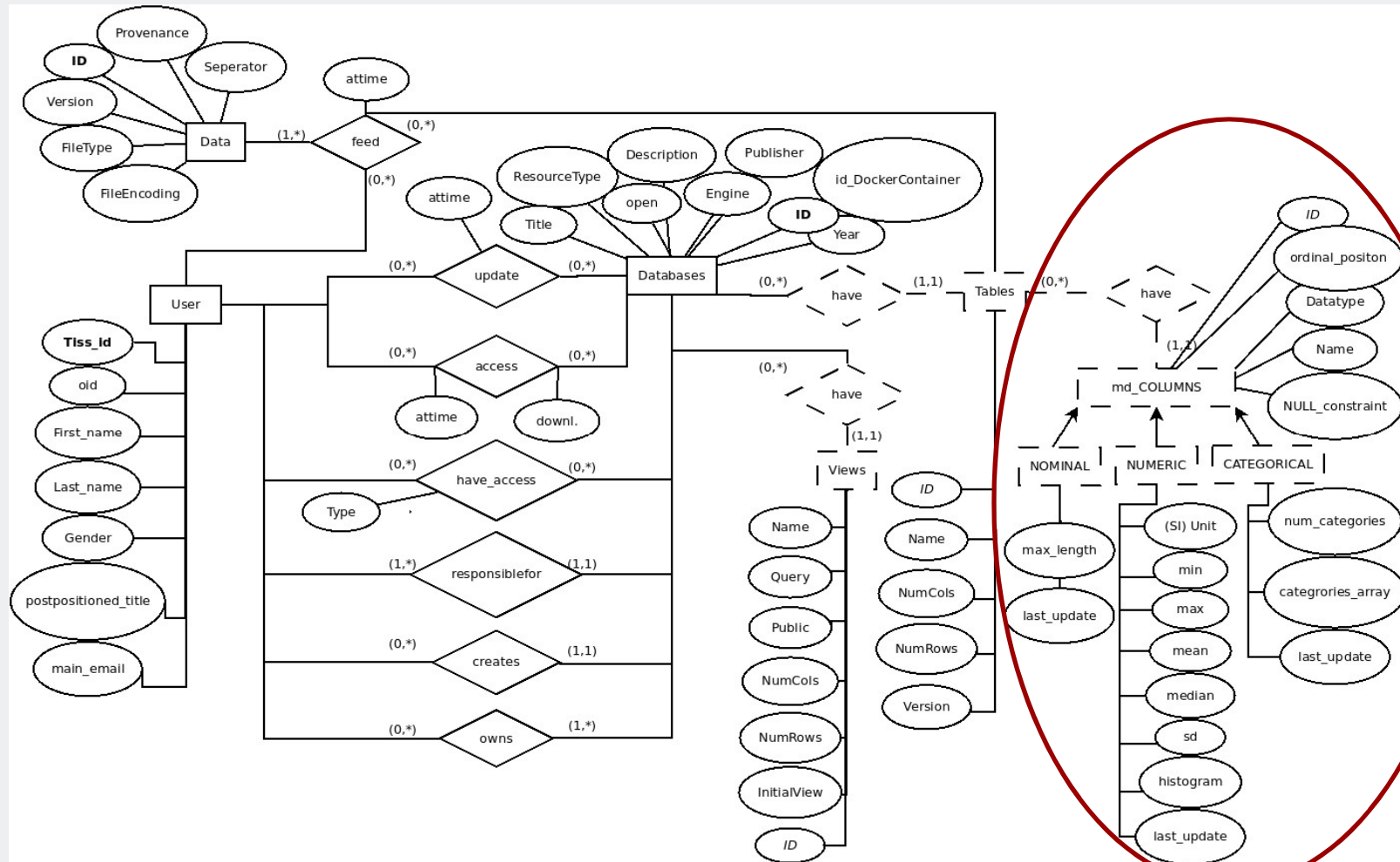
- Structured data
- Private cloud hosted relational databases
- DB is created directly in repository framework
- DB is populated and used within repository
- Metadata is generated and exposed
- Databases and data are searchable
- Data is versioned & time-stamped: reproducibility, re-use, provenance
- 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

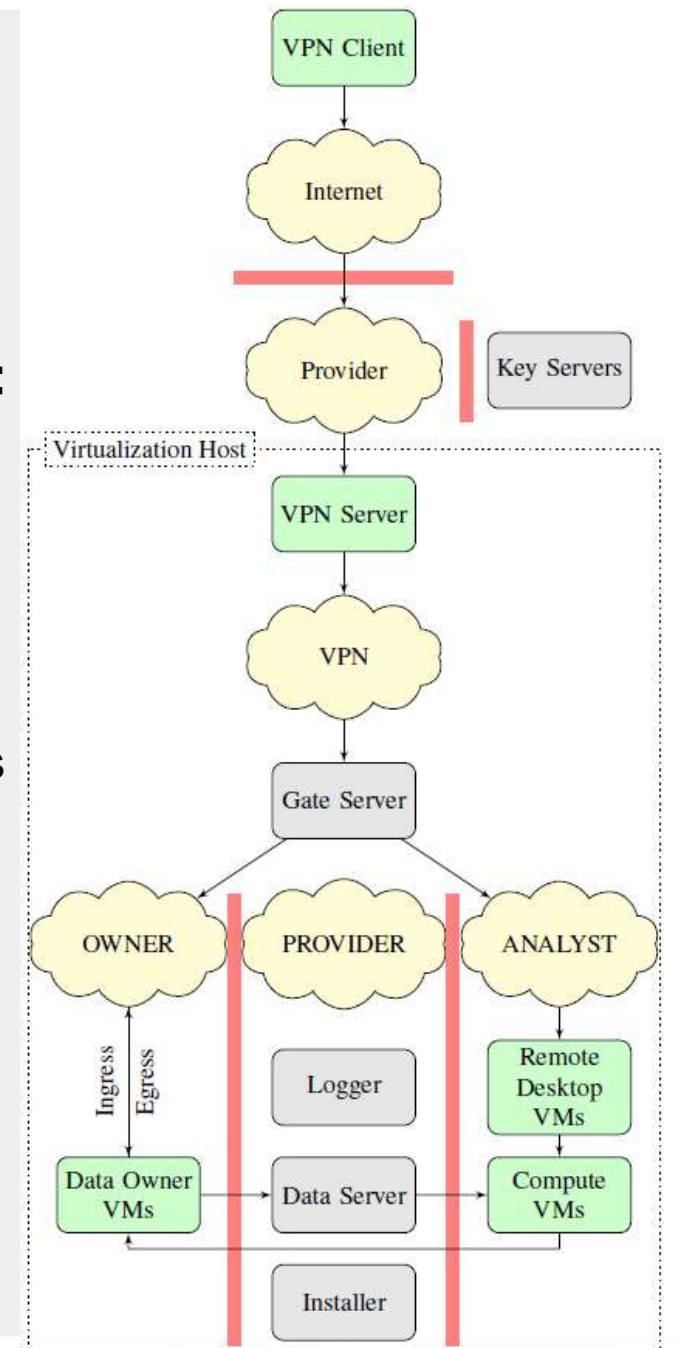
- The *F* in FAIR
- FAIR  $\neq$  Open

## 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**)



# Secure Data Infrastructure (OSSDIP) - Processes

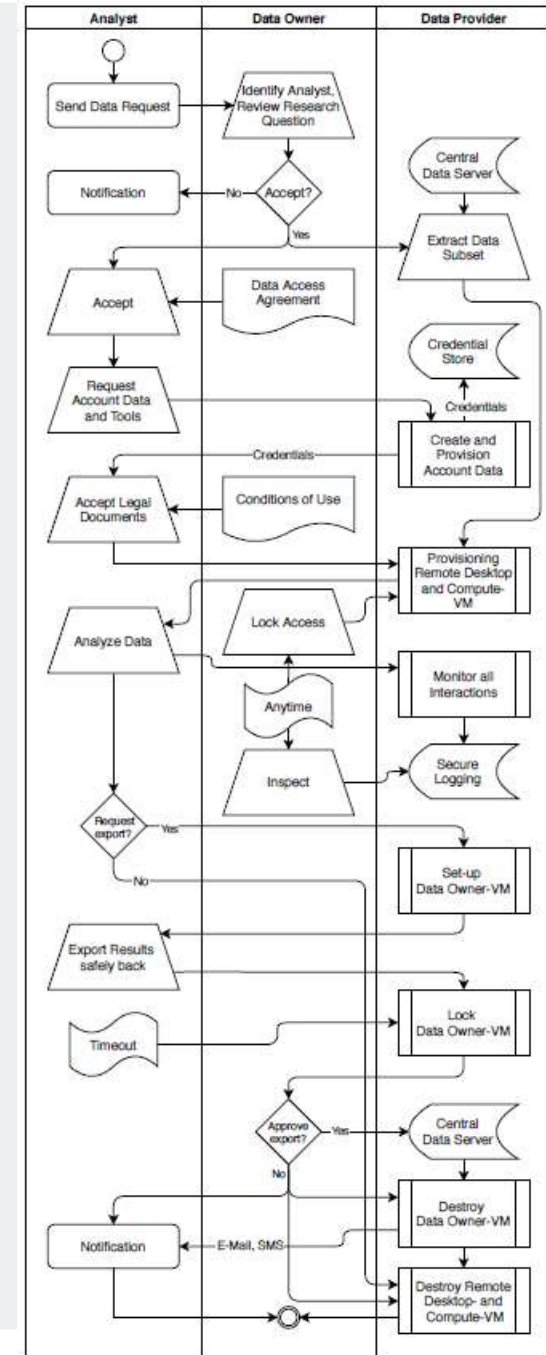
## Data ingest

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





## 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>

Weise, Martin > ossdip

ifs

OSSDIP

Project ID: 870

Privacy

Security

Vpn Server

+ 2 more

☆ Star

<- 50 Commits

9 Branches

0 Tags

7.3 MB Files

8.2 MB Storage

Open Source Secure Data Infrastructure and Processes (OSSDIP). Supporting fully controlled data visiting for sensitive data.

[http://www.ifs.tuwien.ac.at/~andi/secure\\_data\\_infrastructure.html](http://www.ifs.tuwien.ac.at/~andi/secure_data_infrastructure.html)

wiki

gitlab

license

CC-BY

master

ossdip

History

Find file

Clone

updated width for logos

Weise, Martin authored 1 month ago

71bae483


README

Creative Commons Attribution 4.0 International

Weise, Martin

SSSPD | Will | ...

Overview

Last edited by  **Weise, Martin** 3 months ago

Page history

## Overview

We aim at ensuring ongoing confidentiality through [technical and organizational measures](#), integrity with our five controls (see [Secure Data Infrastructure Controls](#)), availability through deploying the system on specialized server hardware and using common tools to establish a secure connection from the open Internet, and resilience through using only best-practice open source software.

A secure system needs to deploy security controls that target every technical and organizational aspect specific to the setting of secured data visiting. We address the security of processing sensitive data by architectural design and automated decision making on behalf of stakeholders. The provider of the secure data infrastructure in legal terms acts as data processor, where the actual ingress of sensitive data is initiated by the Data Owner role (see Roles and Controlled Access for detailed role definitions) as is the egress (after initiation of the Analyst). Our secure data infrastructure stores the sensitive data in a central [database](#) that has a strict firewall barrier around it.

Only process-approved connections to selected VMs (from [Data Owner-VMs](#) (for data import) or to [Analyst-VMs](#) (once, on set-up, to provide an isolated copy of the data), as well as for maintenance and monitoring ([Logging Server](#)) etc.) are allowed to pass this barrier. The overall concept, in a nutshell, is centered around the principle of never providing access to the central [Data Server](#) where all data is being held. Instead, for each individual analysis request, the data required is extracted from the central [Data Server](#) and copied onto a dedicated virtual machine ([Analyst-VM](#)) together with the tools required to perform the analysis.

Access to this [Analyst-VM](#) is granted to the (single) [Analyst](#) working on the task at hand – however, never directly, but only via a dedicated [Remote Desktop-VM](#) to introduce a media break and avoid any data flowing off via e.g. a tunnel. Thus, an Analyst can establish a remote desktop connection to a dedicated VM from which he or she has the sole possibility of establishing a secure shell connection (SSH) to the corresponding [Analyst-VM](#), holding a copy of only the subset of data (possibly finger-printed and aggregated) as well as the tools required for addressing the task at hand. Export of any result files (trained models, figures, charts) is again possible only via a dedicated [Data Owner-VM](#) to ensure approval of [Data Owner](#). These VMs are being destroyed after a specific transfer or analysis task has been completed.

## 1 Introduction

- 1.1 Requirements
- 1.2 Configuration
- 1.3 Deployment

## 2 System Architecture

- 2.1 Overview
- 2.2 Isolated Virtual Machines
- 2.3 Organizational Measures
- 2.4 Core Infrastructure Components
  - VPN Server
  - Gate Server
  - Data Server
  - Installer Server
  - Logging Server
  - Data Owner-VM
  - Analyst-VM
  - Remote Desktop-VM
  - Key Servers

## 3 Secure Data Infrastructure Controls

- 3.1 Roles and Controlled Access
  - Data Owner
  - Analyst
  - Data Provider
  - Carrier

March 24, 2021
Report
Open Access

# Open Source Secure Data Infrastructure and Processes for Data Visiting

Martin Weisse Andreas Rauber

Meeting the conflicting goals of protecting and maintaining control over sensitive data while also allowing access by third parties constitutes a significant challenge. Secure data infrastructures support data visiting in a highly controlled and monitored environment which, if properly set-up and operated, provide high security guarantees through a combination of technical, legal and procedural mechanisms. To ease the process of deploying such a secure data infrastructure, we present a detailed documentation of the architecture and processes of such an infrastructure and provide a pre-configured reference implementation based entirely on open source software that can be flexibly configured to meet differing security requirements and deployment scenarios. We combine mechanisms for data visiting on secured infrastructure components with optional components of data anonymization and fingerprinting, covered by extensive logging and monitoring functions and embedded in defined procedures and contractual frameworks based upon the experience of operating such a secure infrastructure in the medical domain for almost ten years, addressing the emerging need to make such a solution available to a larger set of stakeholders. We show that our system significantly enhances data visiting offers a higher level of isolation and presents less learned.

View
Download
Automatic Zoom

## Open Source Secure Data Infrastructure and Processes for Data Visiting Technical Report

Martin Weisse<sup>1</sup> and Andreas Rauber<sup>1</sup>

<sup>1</sup>Institute of Information Systems Engineering, Vienna University of Technology

**Abstract**

Meeting the conflicting goals of protecting and maintaining control over sensitive data while also allowing access by third parties constitutes a significant challenge. Secure data infrastructures support data visiting in a highly controlled and monitored environment which, if properly set up and operated, But even outside this exceptional situation, academic-industry collaborations as well as industry-to-industry collaborations frequently are hindered by the conflicting needs to keep the data that the other party should process or analyze secret. Homomorphic encryption, while ensuring that the data is kept hidden from the analyst, does not sufficiently support

Files (957 kB)
Name
Size
eosc\_ossdip\_report.pdf
957 kB
Previous
Download

79 views
72 downloads
See more details...

Indexed in

**Publication date:**  
March 24, 2021  
  
**DOI:**  
[DOI: 10.5270/zenedo.4632903](#)

**Keywords:**  
Research Infrastructures
Data Sharing / Data Visiting
Security

**Grants:**  
European Commission  
• EOSCsecretariat.eu - EOSCsecretariat.eu (831644)

**Related identifiers:**  
Supplementary material  
<https://gitlab.tuwien.ac.at/martin.weisse/ossdp>  
(Software documentation)

**Communities:**  
EOSCsecretariat.eu

**Licence (or files):**  
Creative Commons Attribution 4.0 International

**Versions**

Version 1	Mar 24, 2021
10.5270/zenedo.4632903	

**Cite all versions?** You can cite all versions by using the DOI

# 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  $\neq$  Open
  - Data visiting instead of data sharing
  - Non-open data can be made FAIR!
- The technical aspects are, basically, easy