Repositories for Research Data and Trusted Research Environments

Martin Weise

with adapted slides from Andreas Rauber

Data Science Group
Technical University of Vienna

Motivation

Data is the new oil

… or "the new water"

… or "the new light"

The Economist
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world's most valuable resource
amazon
Google
Data and the new rules of competition

v Brussels, Theresa May, 2017. The World's Most Valuable Resource, in *The Economist*. Edition May 6th.

Motivation

Data is the new oil ... if properly managed!
... or “the new water”
... or “the new light”

Otherwise, it’s an oilspill
... or flood
... or blinding flash of lighting

The Economist
MAY 6TH–12TH 2017
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea’s unfinished revolution
Biology, but without the cells
The world’s most valuable resource
Data and the new rules of competition
amazon
Google

v Brussels, Theresa May, 2017. The World’s Most Valuable Resource, in *The Economist*. Edition May 6th.

Motivation

Research depends on data in virtually all disciplines:

- **Value** of data determinded through
  - Exhaustive collection
  - Data (pre-)processing
  - Volume, e.g. meta-studies
  - Reproducability as core principle of science
- Proper **data management** enables
  - Speedup of research (avoiding repeated collection, processing)
  - Robust resarch (larger data pools)
  - Increased quality (reproducability, comparability)

Agenda

1. Introduction
2. Background
3. Repositories for Research Data
   - DSpace
   - Gitlab
   - InvenioRDM
   - DBRepo
4. Trusted Research Environments
   - RemoteNEPS
   - SAIL Gateway
   - DEXHELPP
   - OSSDIP
5. Future Work
6. Conclusion

1. Introduction

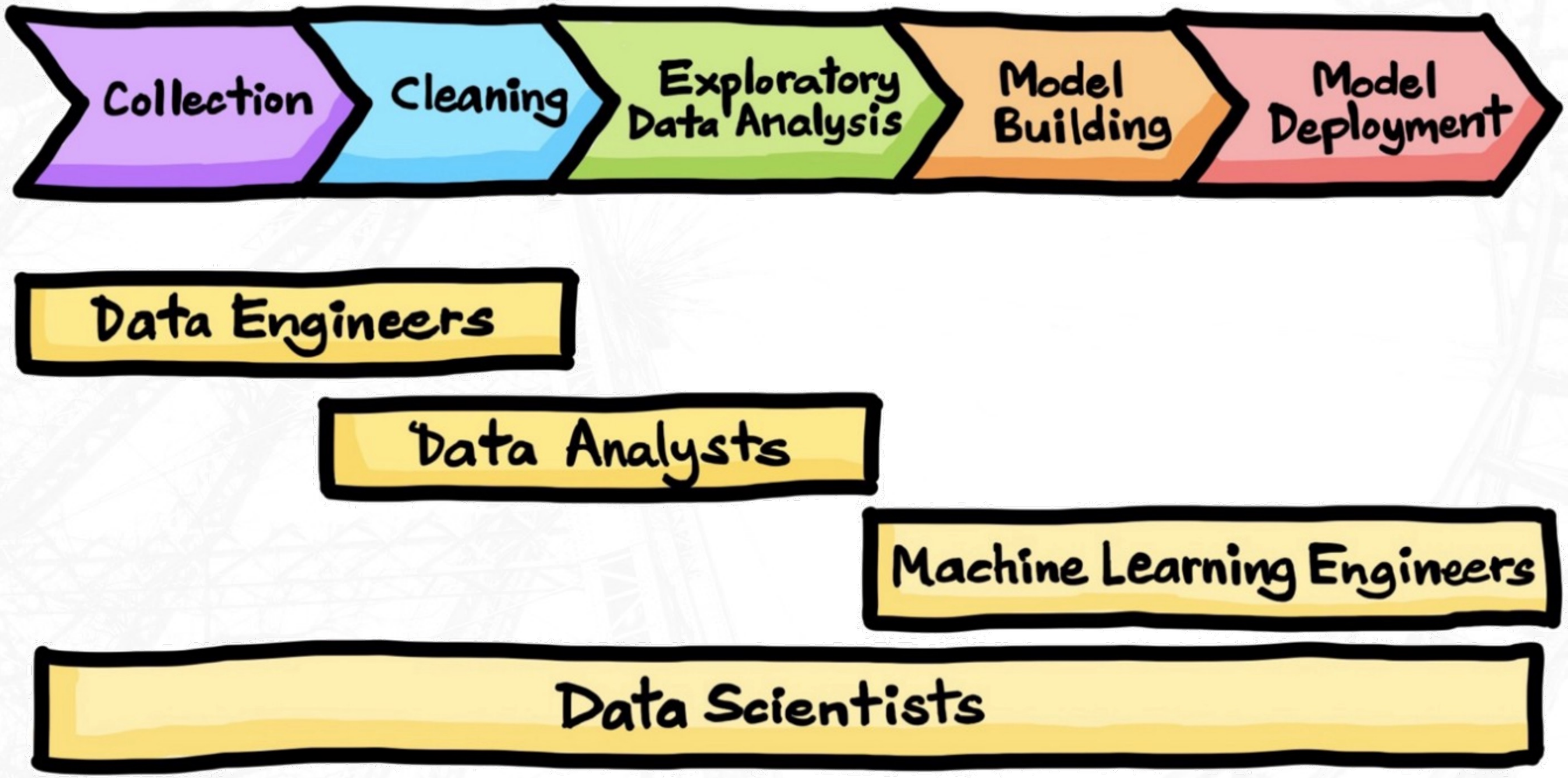
PreDoc Researcher in the first year

- **MSc** in *Software Engineering & Internet Computing* 2022 (Technical University of Vienna)
- **BSc** in *Software & Informaiton Engineering* 2019 (Technical University of Vienna)

Research interests:

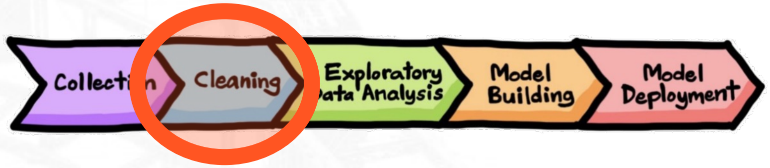
- Making sensitive data FAIR

2. Background



Nantasenamat, C. & Jee, K., 2020. [Online]. URL: https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b, accessed 2022-09-09

2. Background



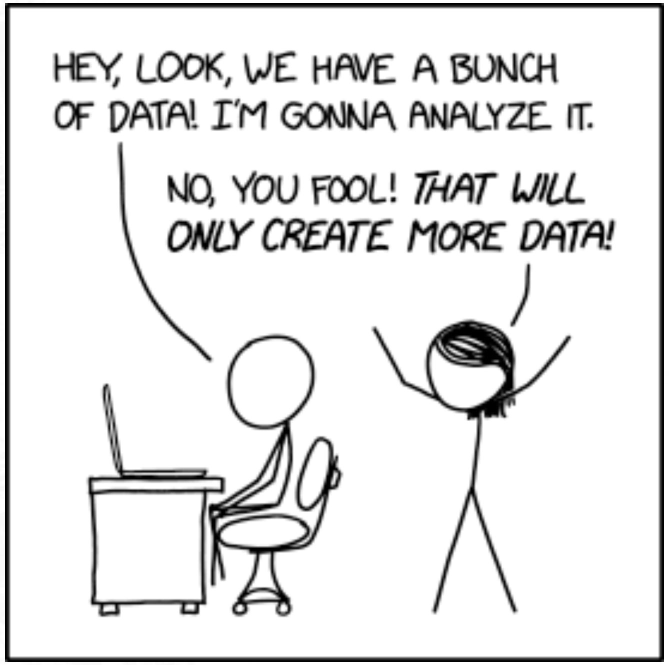
Cleaning and preparing data takes about 80% of the total engineering effort

- Real-world **data** may be
  - *incomplete*, lacking attribute values, contains only aggregated data,
  - *noisy*, containing errors or outliers,
  - inconsistent, discrepancy in names
- **Preparation** generates a subset of the data, potentially increasing utility
  - *attribute selection*, relevant data, anomaly removal, duplicate elimination
  - *reducing data*, sampling or instance selection
- **Outcome**
  - *recovery* of incomplete data
  - *purify data*, correcting errors, removing outliers
  - *resolve conflicts* using domain knowledge, expert decisions

Zhang, S., Zhang, C. & Yang, Q., 2003. Data preparation for data mining, in *Applied Artificial Intelligence*, 17(5-6), p.375-381, DOI: 10.1080/713827180

2. Background

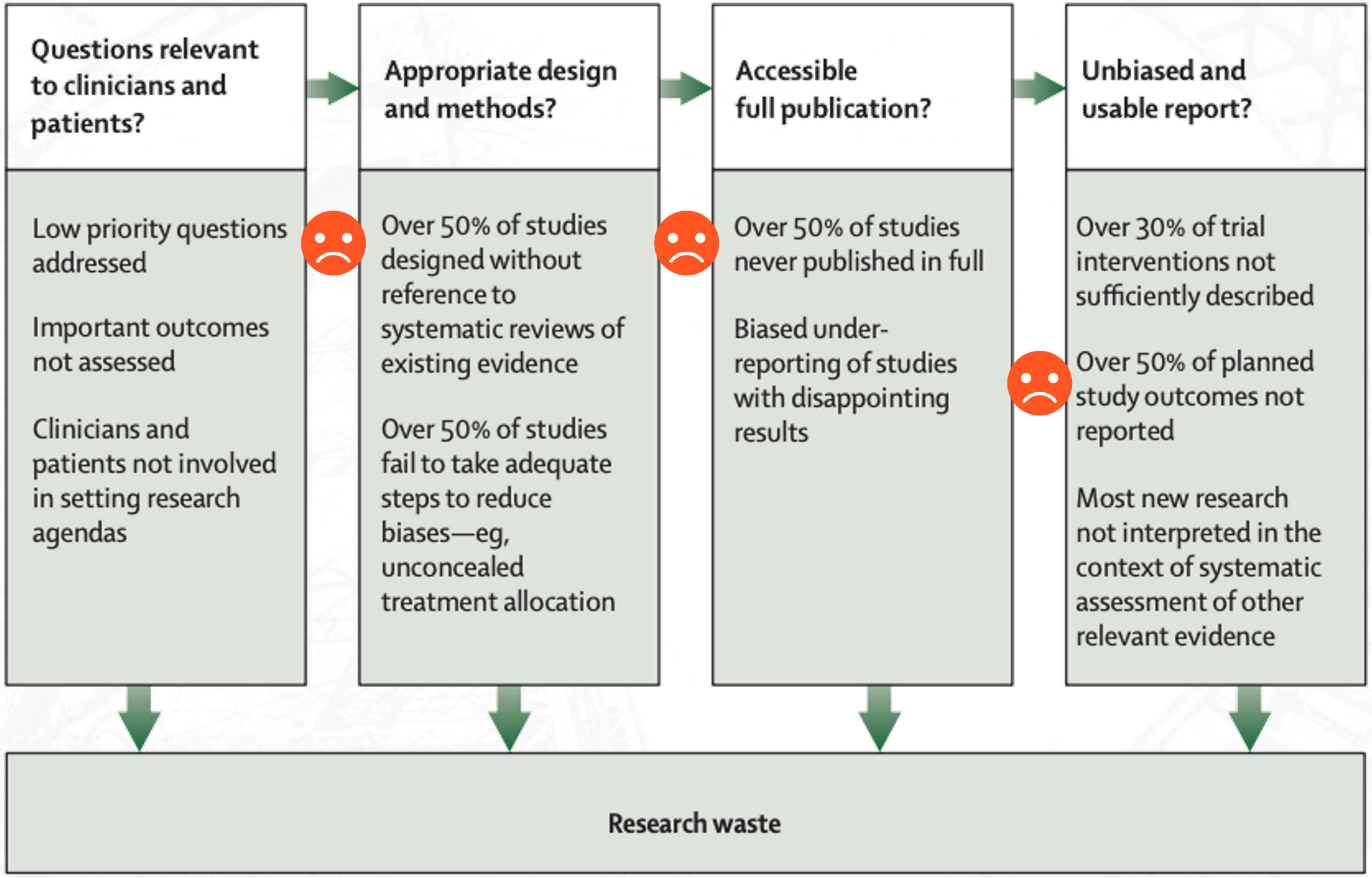
After all this effort, people will be happy to re-use data, **right**?


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

ALT: “It's important to make sure your analysis destroys as much information as it produces.”

Munroe, R., 2022. Data Trap. [Online]. URL: https://xkcd.com/2582/, accessed 2022-09-11

2. Background



Questions relevant to clinicians and patients?

Low priority questions addressed

Important outcomes not assessed

Clinicians and patients not involved in setting research agendas

Appropriate design and methods?

Over 50% of studies designed without reference to systematic reviews of existing evidence

Over 50% of studies fail to take adequate steps to reduce biases—eg, unconcealed treatment allocation

Accessible full publication?

Over 50% of studies never published in full

Biased under-reporting of studies with disappointing results

Unbiased and usable report?

Over 30% of trial interventions not sufficiently described

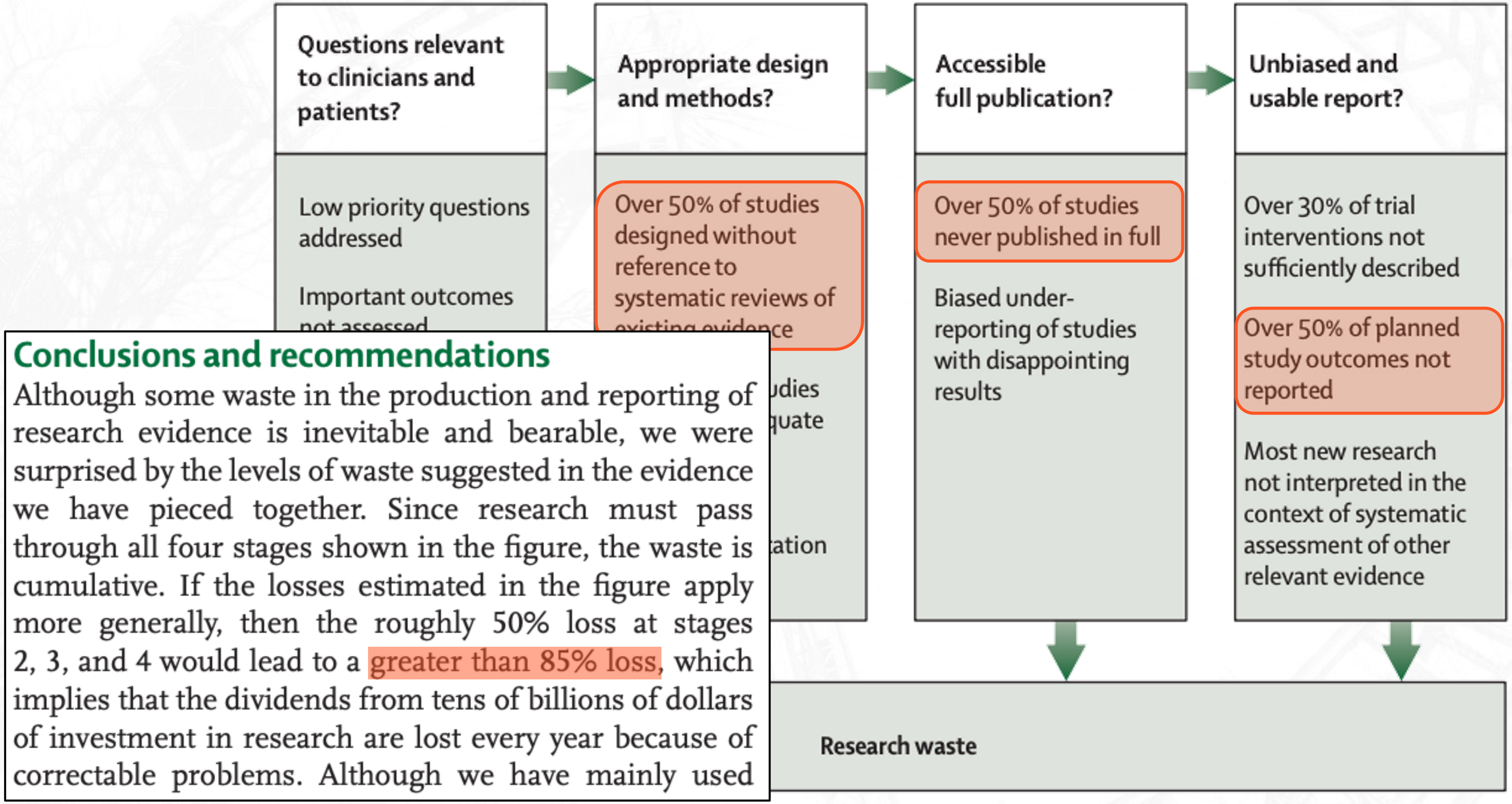
Over 50% of planned study outcomes not reported

Most new research not interpreted in the context of systematic assessment of other relevant evidence

Research waste

Chalmers, I., Glasziou, P., 2009. Avoidable Waste in the Production and Reporting of Research Evidence, in *The Lancet,* 374(9683), p.86-9. DOI: 10.1016/S0140-6736(09)60329-9

2. Background



Questions relevant to clinicians and patients?

Low priority questions addressed

Important outcomes not assessed

Appropriate design and methods?

Over 50% of studies designed without reference to systematic reviews of existing evidence

udies

quate

ation

Accessible full publication?

Over 50% of studies never published in full

Biased under-reporting of studies with disappointing results

Unbiased and usable report?

Over 30% of trial interventions not sufficiently described

Over 50% of planned study outcomes not reported

Most new research not interpreted in the context of systematic assessment of other relevant evidence

Research waste

Conclusions and recommendations

Although some waste in the production and reporting of research evidence is inevitable and bearable, we were surprised by the levels of waste suggested in the evidence we have pieced together. Since research must pass through all four stages shown in the figure, the waste is cumulative. If the losses estimated in the figure apply more generally, then the roughly 50% loss at stages 2, 3, and 4 would lead to a greater than 85% loss, which implies that the dividends from tens of billions of dollars of investment in research are lost every year because of correctable problems. Although we have mainly used

Chalmers, I., Glasziou, P., 2009. Avoidable Waste in the Production and Reporting of Research Evidence. *Lancet*. 374(9683), p.86-9. DOI: 10.1016/S0140-6736(09)60329-9

2. Background

Why even share research data?

- **Increase trust** in the work, allow reproduce and validate findings
- Information is **valuable** to the research community
- **Contribute** work beyond the original findings
- Allow others to **re-use** and build on top of their data

2. Background

Why even share research data?

- **Increase trust** in the work, allow reproduce and validate findings
- Information is **valuable** to the research community
- **Contribute** work beyond the original findings
- Allow others to **re-use** and build on top of their data

We contiously use the understanding gained by major thinkers in order to make intellectual progress.

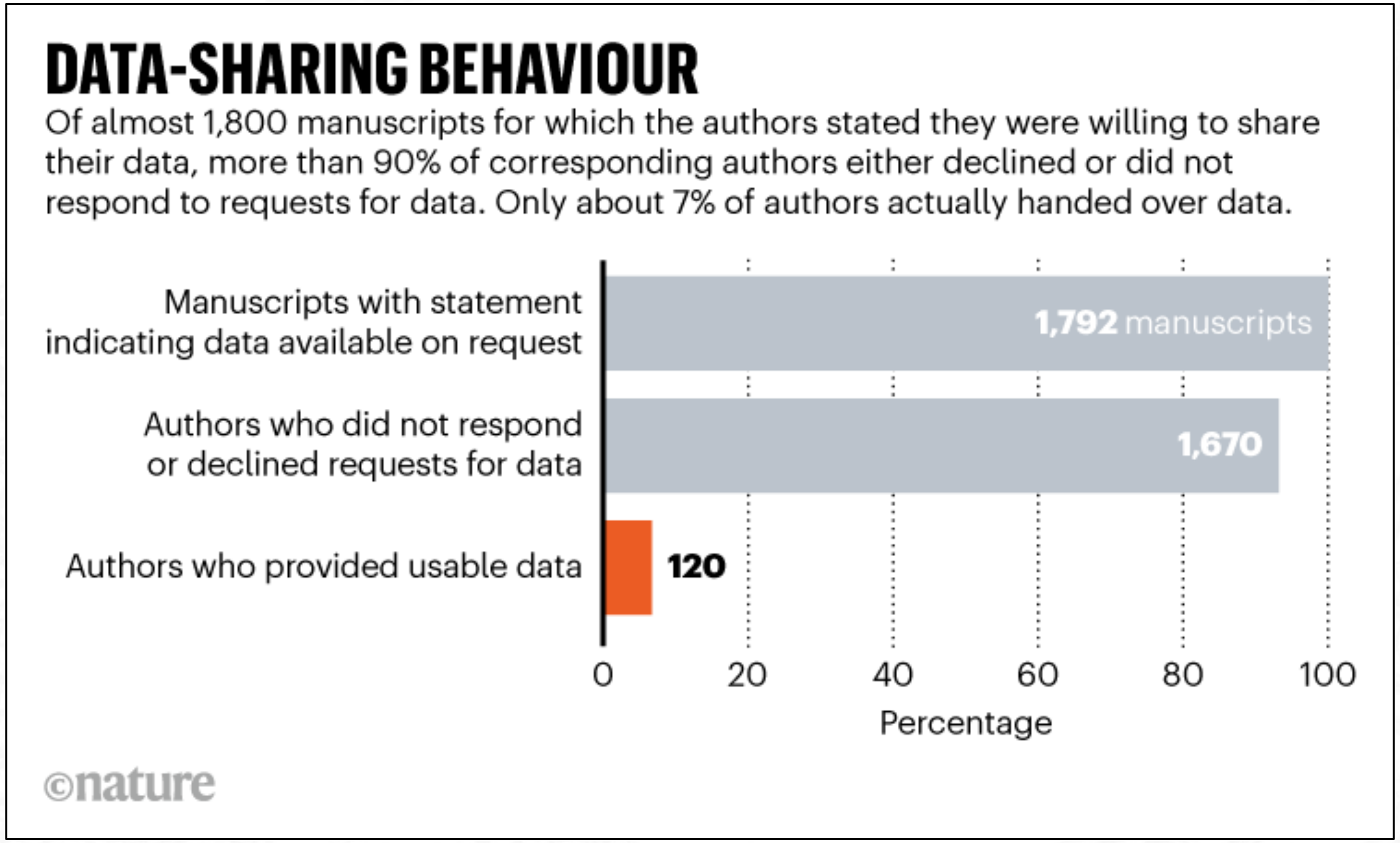
> “If I have seen further it is by standing on the shoulders of Giants.”
>
> —Isaac Newton, 1675.

2. Background

Is this necessary, why not just **reasonably request** data from researchers?

2. Background

Is this necessary, why not just **reasonably request** data from researchers?



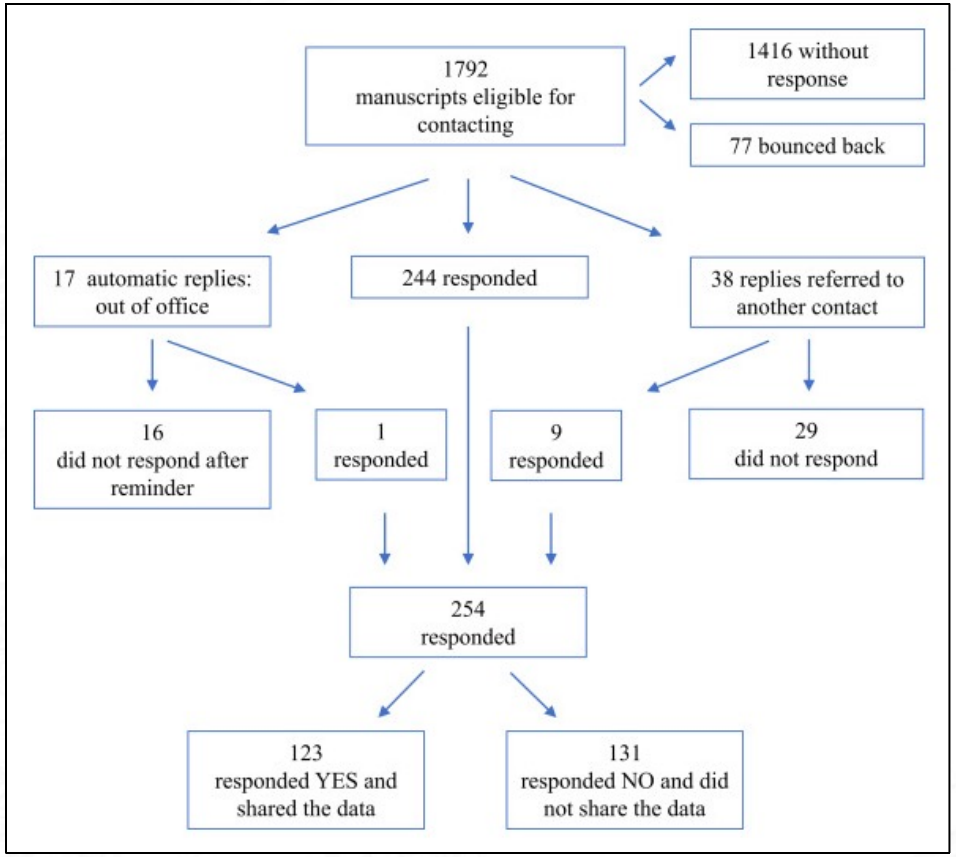
DATA-SHARING BEHAVIOUR

Of almost 1,800 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for data. Only about 7% of authors actually handed over data.

Watson, C., 2022. Many researchers say they'll share data - but don't, in *Nature*, 606, p.853, DOI: 10.1038/d41586-022-01692-1

2. Background

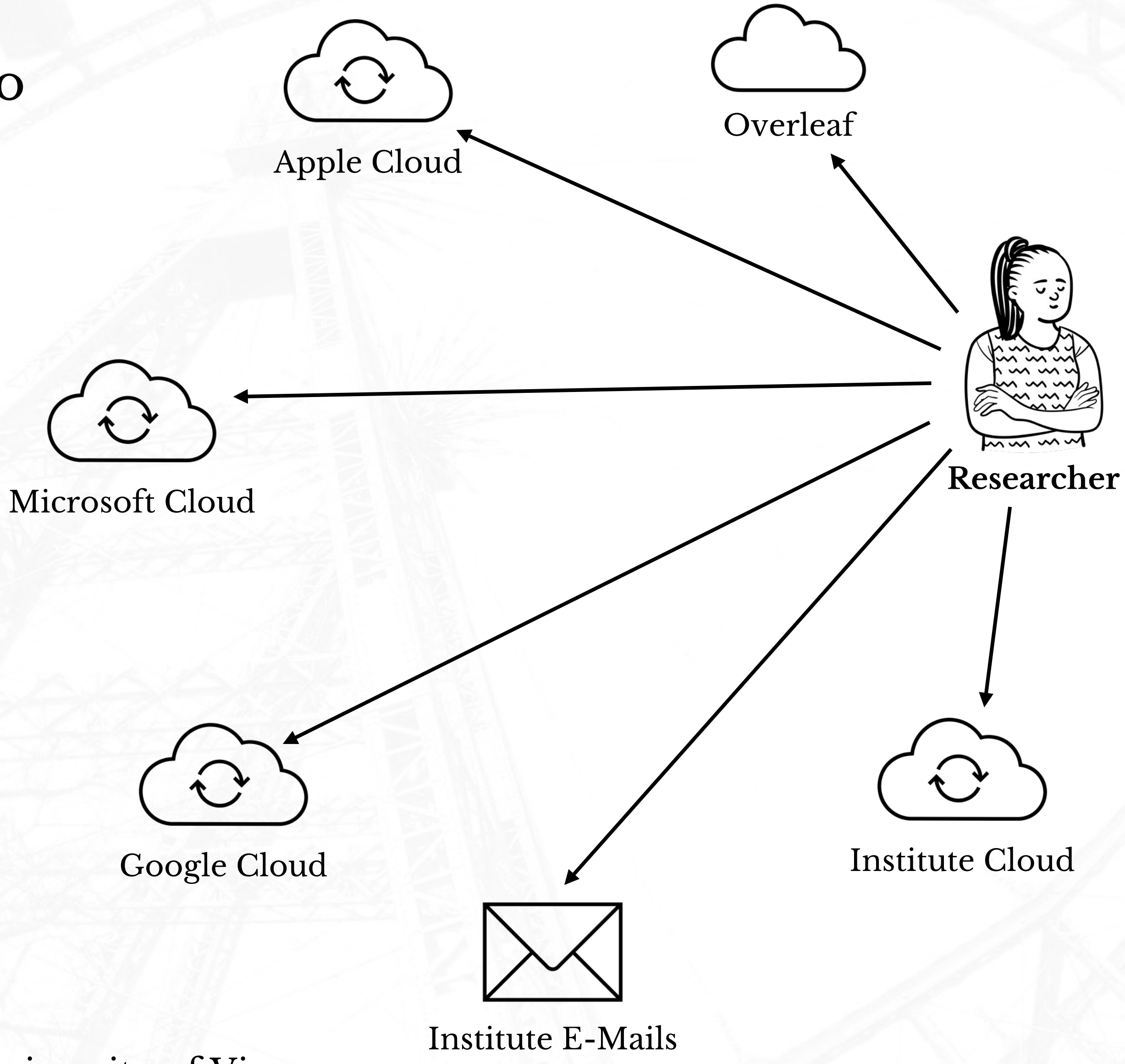
Is this necessary, why not just **reasonably request** data from researchers?


1792
manuscripts eligible for contacting
1416 without response
77 bounced back
17 automatic replies: out of office
244 responded
38 replies referred to another contact
16
did not respond after reminder
1
responded
9
responded
29
did not respond
254
responded
123
responded YES and shared the data
131
responded NO and did not share the data

Gabelica, M., Bojčić, R. & Puljakc, L., 2022. Many Researchers were not Compliant with their Published Data Sharing Statement: a Mixed-methods Study, in *Journal of Clinical Epidemiology*, DOI: 10.1016/j.jclinepi.2022.05.019

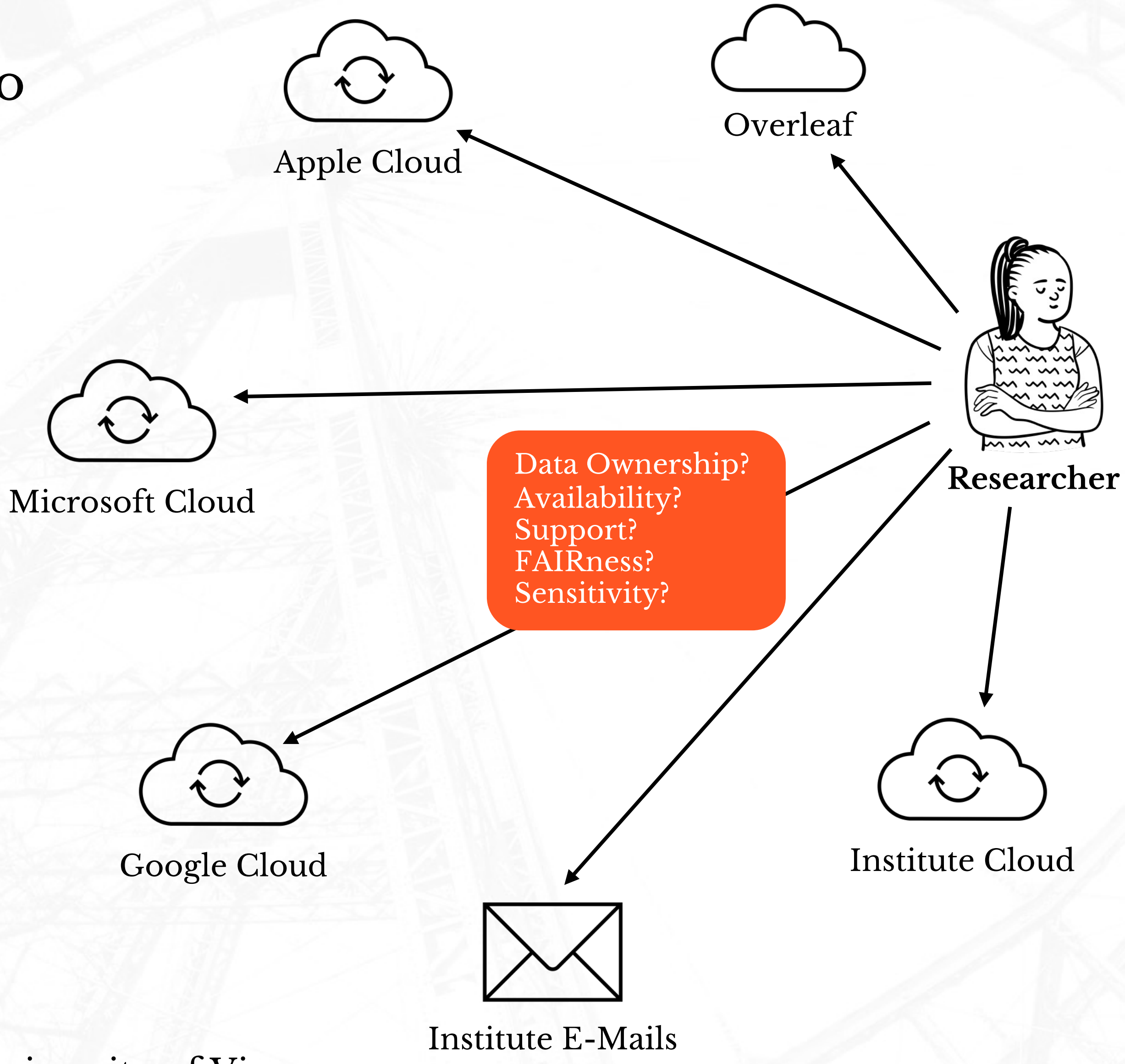
2. Background

Common Scenario



2. Background

Common Scenario



Data Ownership?
Availability?
Support?
FAIRness?
Sensitivity?

3. Repositories for Research Data

What are **research data repositories?**

OpenNEURO

GSA for Human
Genome Sequence Archive

UK Data Service

HARVARD
Dataverse

UniProt

DRYAD

zenodo

Science Data Bank
科学数据银行

IDR

OSF

NDEx

BGS British Geological Survey

figshare

TU Wien Research Data

MENDELEY DATA

3. Repositories for Research Data

Data Management

- Currently largely on researcher's shoulders
- Need **separation of concerns** (excerpt)
  - *Researchers*, **work with data**, domain expertise
  - *Data Stewards*, **curation**, preservation, FAIR
  - *IT-Department*, hardware/software infrastructure, security, backup
  - *Legal-Department*, licenses, NDAs
  - *Admin*, reporting, GDPR, complicance
- **Dedicated infrastructure** to ensure data is **properly managed** and value realized
- Infrastructure for **research data management** connected to internal information systems, funders, etc.

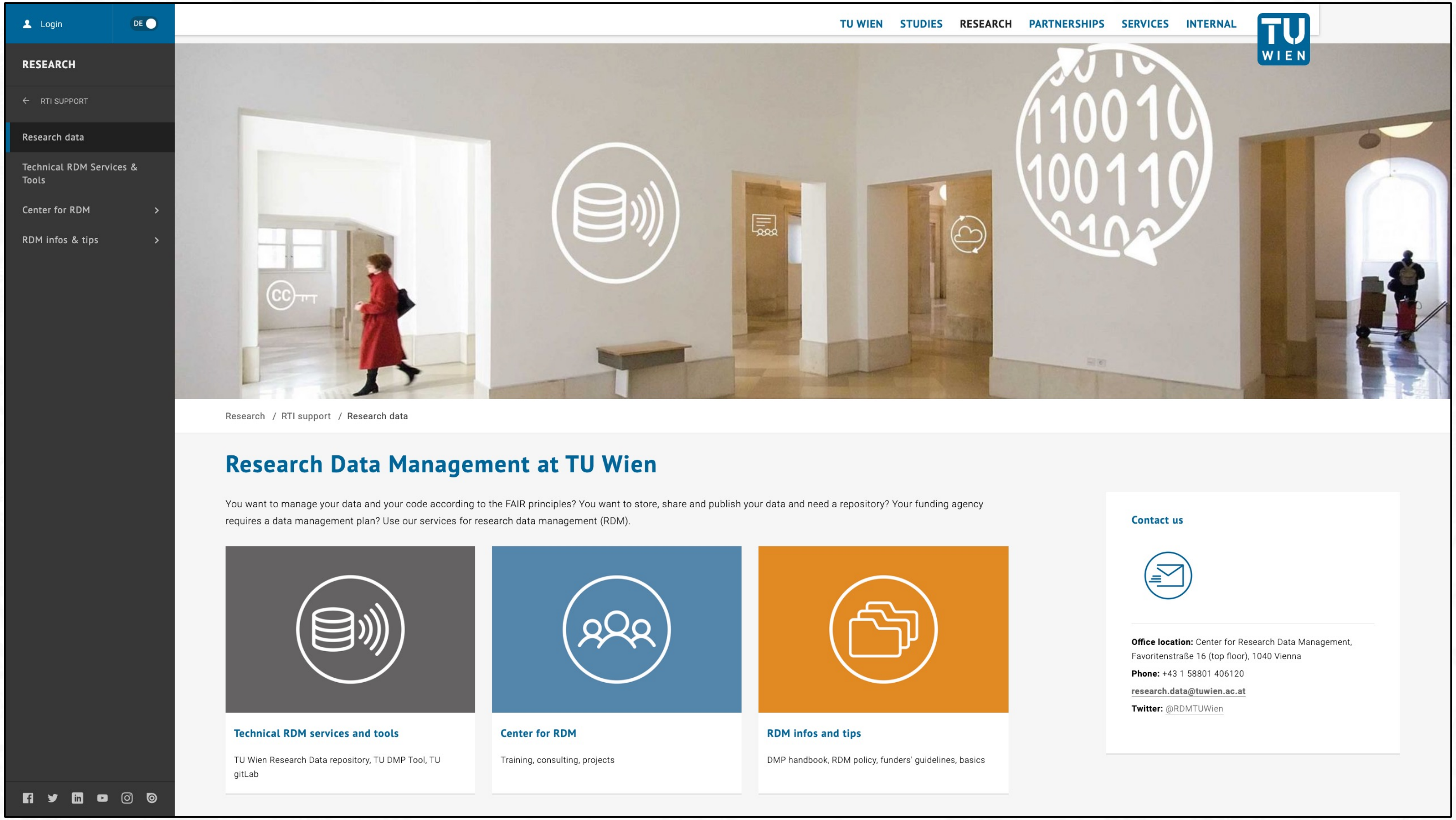
Research Data Management is a joint effort!

3. Repositories for Research Data

Steps towards research infrastructure at TU Vienna:

- Involved stakeholders, regular round tables (rectorate, support offices, research departments)
- Established **policy** on RDM
- Established **Center for Research Data Management**
- Devised plan for
  - National and European projects
    - **FAIR Data Austria**
    - Austrian Data Labs and Services
    - EOSC-* projects (European Open Science Cloud)
    - Involvement in **Research Data Alliance** (RDA), **EGI**, …
  - Setting up and rolling out infrastructure

3. Repositories for Research Data



Research Data Management at TU Wien

You want to manage your data and your code according to the FAIR principles? You want to store, share and publish your data and need a repository? Your funding agency requires a data management plan? Use our services for research data management (RDM).

Technical RDM services and tools

TU Wien Research Data repository, TU DMP Tool, TU gitLab

Center for RDM

Training, consulting, projects

RDM infos and tips

DMP handbook, RDM policy, funders' guidelines, basics

Contact us

Office location: Center for Research Data Management, Favoritenstraße 16 (top floor), 1040 Vienna

Phone: +43 1 58801 406120

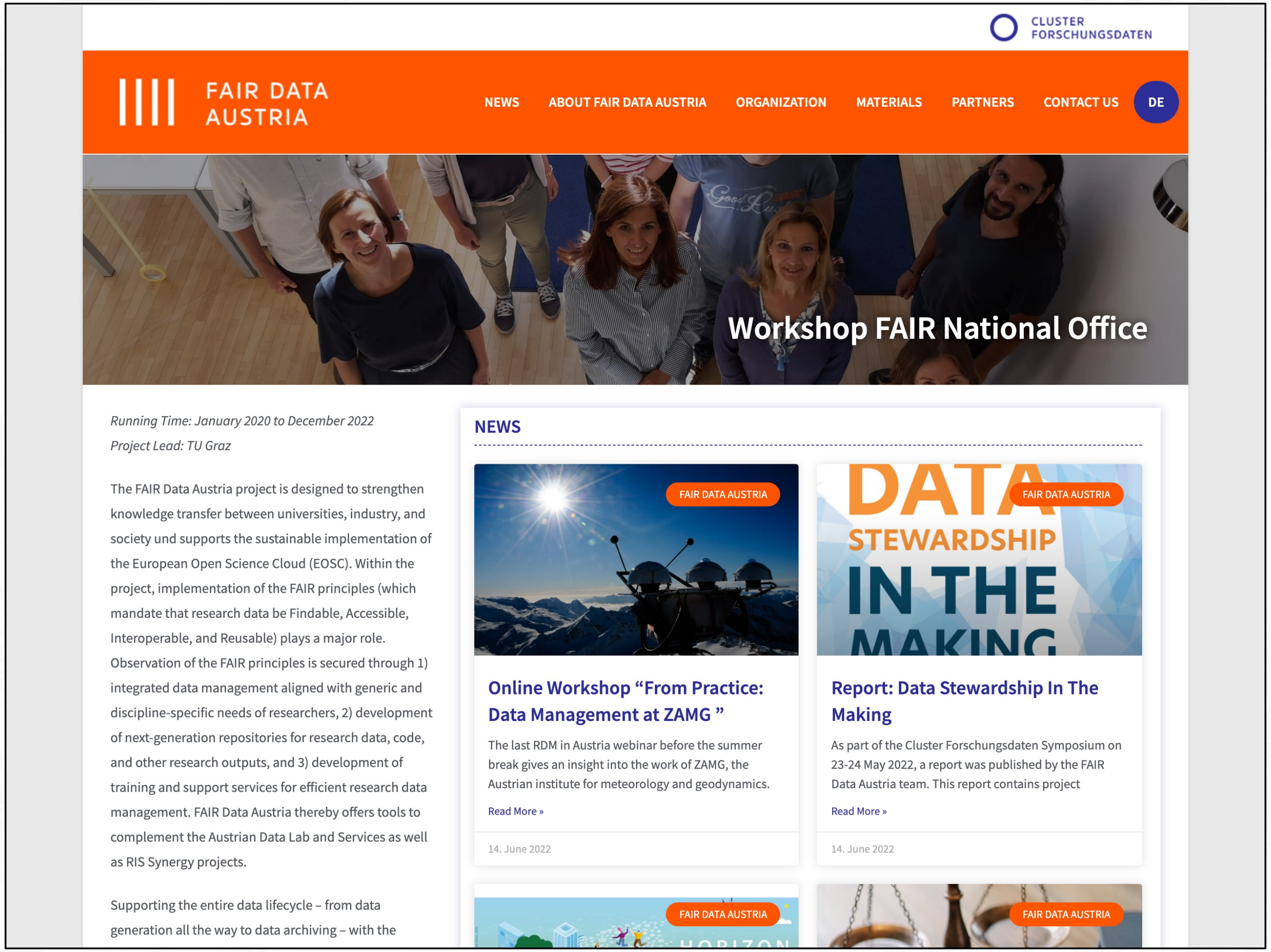
research.data@tuwien.ac.at

Twitter: @RDMTUWien

Research Data. [Online]. URL: https://www.tuwien.at/forschung/fti-support/forschungsdaten, accessed 2022-09-09

Martin Weise, Technical University of Vienna

3. Repositories for Research Data



Running Time: January 2020 to December 2022

Project Lead: TU Graz

The FAIR Data Austria project is designed to strengthen knowledge transfer between universities, industry, and society und supports the sustainable implementation of the European Open Science Cloud (EOSC). Within the project, implementation of the FAIR principles (which mandate that research data be Findable, Accessible, Interoperable, and Reusable) plays a major role. Observation of the FAIR principles is secured through 1) integrated data management aligned with generic and discipline-specific needs of researchers, 2) development of next-generation repositories for research data, code, and other research outputs, and 3) development of training and support services for efficient research data management. FAIR Data Austria thereby offers tools to complement the Austrian Data Lab and Services as well as RIS Synergy projects.

Supporting the entire data lifecycle – from data generation all the way to data archiving – with the

NEWS

Online Workshop "From Practice: Data Management at ZAMG "

The last RDM in Austria webinar before the summer break gives an insight into the work of ZAMG, the Austrian institute for meteorology and geodynamics.

Read More »

14. June 2022

Report: Data Stewardship In The Making

As part of the Cluster Forschungsdaten Symposium on 23-24 May 2022, a report was published by the FAIR Data Austria team. This report contains project

Read More »

14. June 2022

FAIR Data Austria. [Online]. URL: https://forschungsdaten.at/en/fair-data-austria/, accessed 2022-09-09

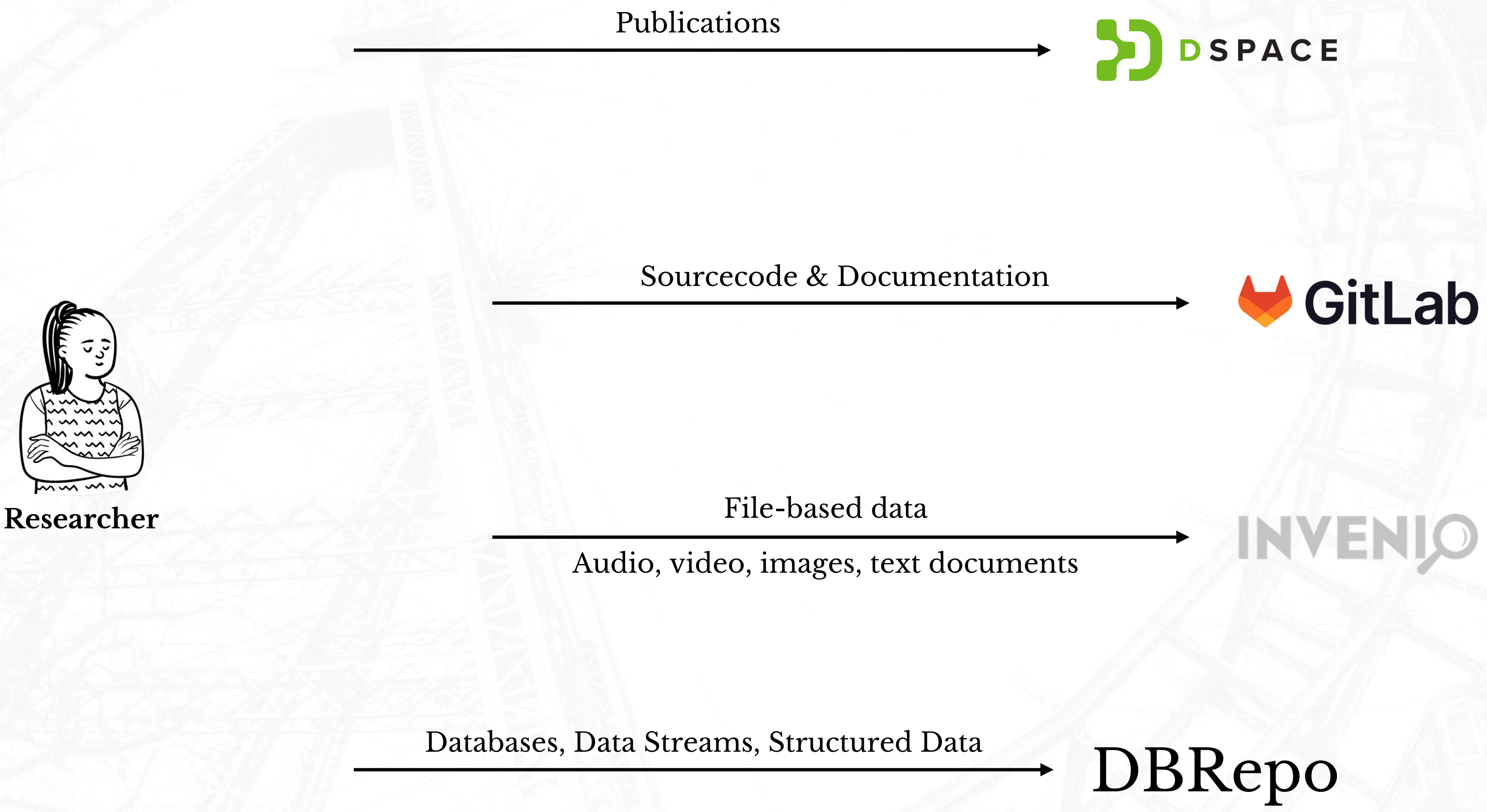
3. Repositories for Research Data

Repository Infrastructure

1. Repositories for **all kinds** of research material
    - Input, output, interim
    - Open and closed / sensitive data
2. Provide **visibility**
    - Citation, impact
    - FAIR compliant
3. Be largely **transparent** to researchers
    - Integration with TU and external infrastructure
4. Be **trustworthy**

3. Repositories for Research Data

Architecture and vision

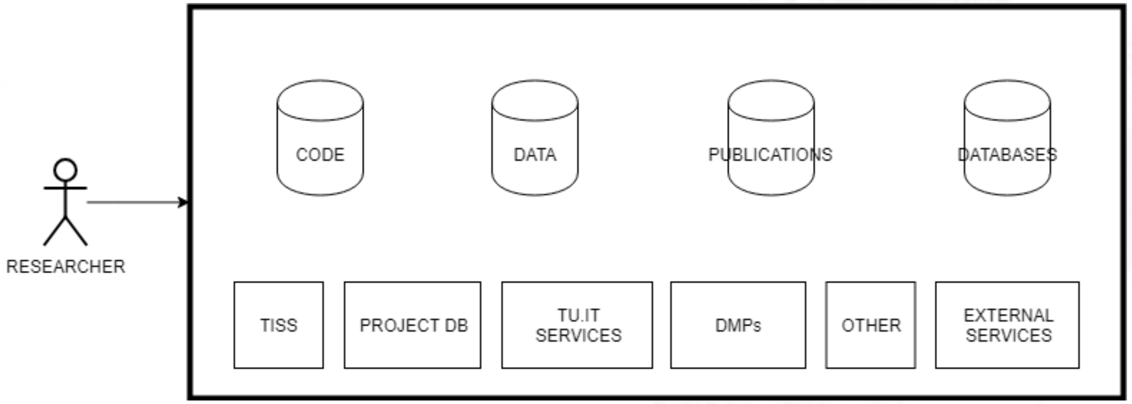


Publications
DSPACE
Sourcecode & Documentation
GitLab
Researcher
File-based data
Audio, video, images, text documents
INVENIO
Databases, Data Streams, Structured Data
DBRepo

3. Repositories for Research Data

Repository Infrastructure

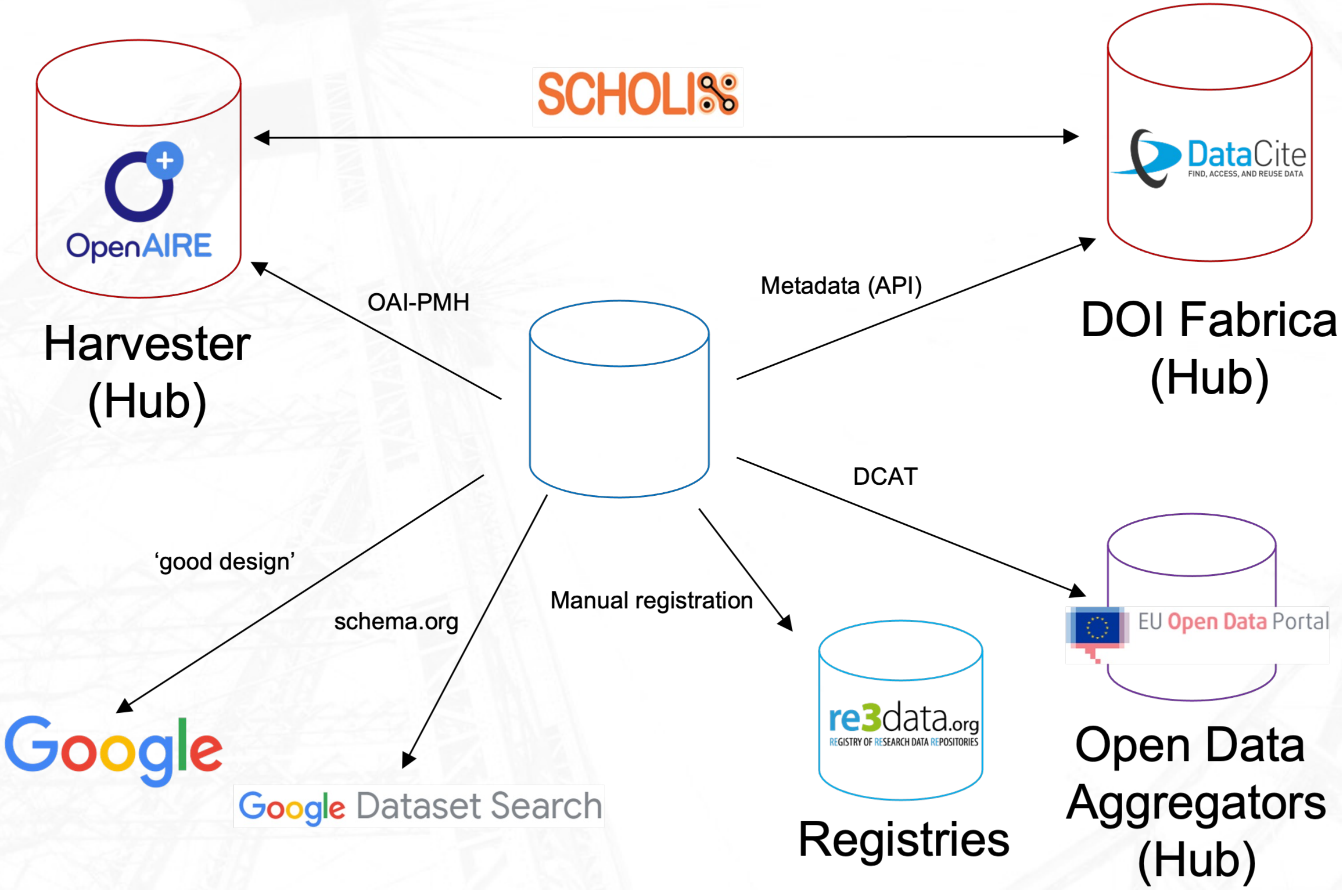
- Provide unified services for research data management
- Cooperation
  - Center for Research Data Management
  - TU.it
  - Campus Development Services
  - TU library
  - Research Unit Data Science (formerly IFS)
  - Central administration, legal department



Systems integrated appear as a ‘single repository’

3. Repositories for Research Data

External Visibility: nobody searches in your repo to find data at first



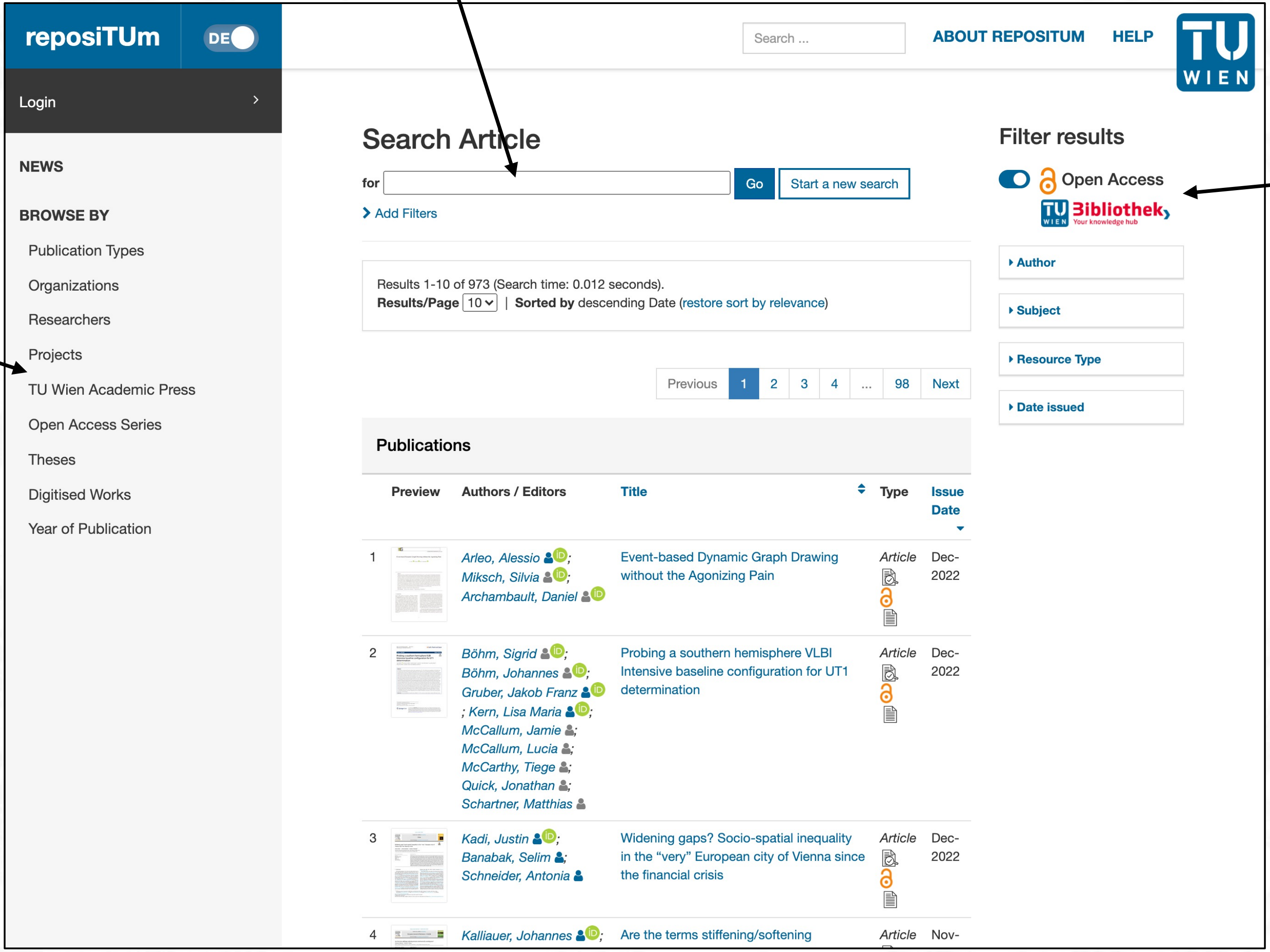
Martin Weise, Technical University of Vienna

3.1. ReposiTUm (DSpace)

Query Search

Faceted Browsing

Accessability Filter



reposiTUm DE

Login ›

Search ...

ABOUT REPOSITUM HELP

NEWS

BROWSE BY

- Publication Types
- Organizations
- Researchers
- Projects
- TU Wien Academic Press
- Open Access Series
- Theses
- Digitised Works
- Year of Publication

Search Article

for [ ] Go | Start a new search

› Add Filters

Results 1-10 of 973 (Search time: 0.012 seconds).
Results/Page 10 | **Sorted by** descending Date (restore sort by relevance)

Previous 1 2 3 4 ... 98 Next

Publications

	Preview	Authors / Editors	Title	Type	Issue Date
1		Arleo, Alessio; Miksch, Silvia; Archambault, Daniel	Event-based Dynamic Graph Drawing without the Agonizing Pain	Article	Dec-2022
2		Böhm, Sigrid; Böhm, Johannes; Gruber, Jakob Franz ; Kern, Lisa Maria; McCallum, Jamie; McCallum, Lucia; McCarthy, Tiege; Quick, Jonathan; Schartner, Matthias	Probing a southern hemisphere VLBI Intensive baseline configuration for UT1 determination	Article	Dec-2022
3		Kadi, Justin; Banabak, Selim; Schneider, Antonia	Widening gaps? Socio-spatial inequality in the “very” European city of Vienna since the financial crisis	Article	Dec-2022
4		Kalliauer, Johannes;	Are the terms stiffening/softening	Article	Nov-

Filter results

Open Access

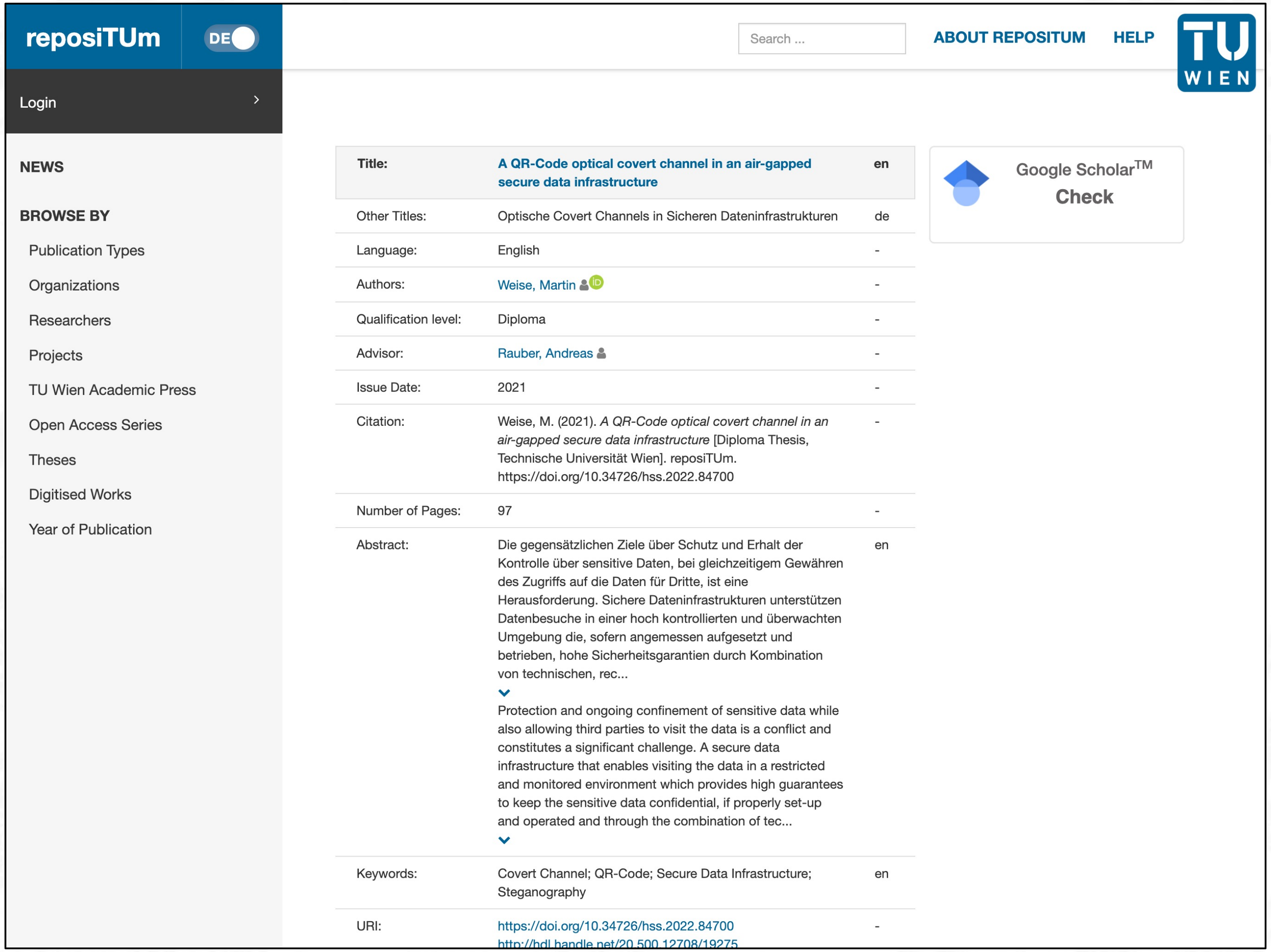
TU Bibliothek Your knowledge hub

- Author
- Subject
- Resource Type
- Date issued

ReposiTUm. [Online]. URL: https://repositum.tuwien.at/, accessed 2022-09-09

Martin Weise, Technical University of Vienna

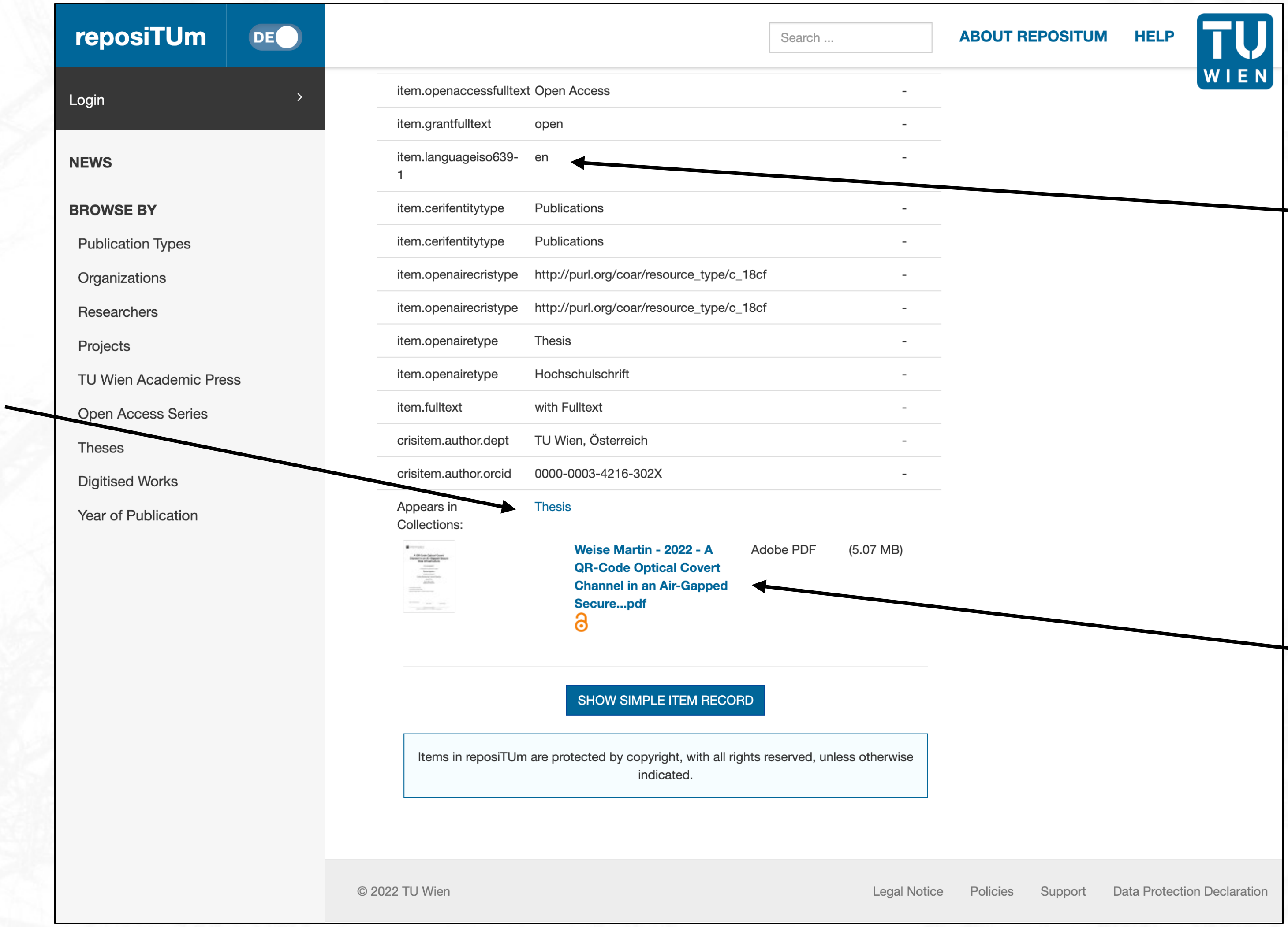
3.1. ReposiTUm (DSpace)



Title:	**A QR-Code optical covert channel in an air-gapped secure data infrastructure**	en
Other Titles:	Optische Covert Channels in Sicheren Dateninfrastrukturen	de
Language:	English	-
Authors:	Weise, Martin	-
Qualification level:	Diploma	-
Advisor:	Rauber, Andreas	-
Issue Date:	2021	-
Citation:	Weise, M. (2021). *A QR-Code optical covert channel in an air-gapped secure data infrastructure* [Diploma Thesis, Technische Universität Wien]. reposiTUm. https://doi.org/10.34726/hss.2022.84700	-
Number of Pages:	97	-
Abstract:	Die gegensätzlichen Ziele über Schutz und Erhalt der Kontrolle über sensitive Daten, bei gleichzeitigem Gewähren des Zugriffs auf die Daten für Dritte, ist eine Herausforderung. Sichere Dateninfrastrukturen unterstützen Datenbesuche in einer hoch kontrollierten und überwachten Umgebung die, sofern angemessen aufgesetzt und betrieben, hohe Sicherheitsgarantien durch Kombination von technischen, rec... Protection and ongoing confinement of sensitive data while also allowing third parties to visit the data is a conflict and constitutes a significant challenge. A secure data infrastructure that enables visiting the data in a restricted and monitored environment which provides high guarantees to keep the sensitive data confidential, if properly set-up and operated and through the combination of tec...	en
Keywords:	Covert Channel; QR-Code; Secure Data Infrastructure; Steganography	en
URI:	https://doi.org/10.34726/hss.2022.84700 http://hdl.handle.net/20.500.12708/19275	-

Weise, M, 2021. A QR-Code Optical Covert Channel in an Air-Gapped Secure Data Infrastructure. [Thesis], p.97, DOI: 10.34726/hss.2022.84700

3.1. ReposiTUm (DSpace)



item.openaccessfulltext	Open Access	-
item.grantfulltext	open	-
item.languageiso639-1	en	-
item.cerifentitytype	Publications	-
item.cerifentitytype	Publications	-
item.openairecristype	http://purl.org/coar/resource_type/c_18cf	-
item.openairecristype	http://purl.org/coar/resource_type/c_18cf	-
item.openairetype	Thesis	-
item.openairetype	Hochschulschrift	-
item.fulltext	with Fulltext	-
crisitem.author.dept	TU Wien, Österreich	-
crisitem.author.orcid	0000-0003-4216-302X	-
Appears in Collections:	Thesis	

Weise Martin - 2022 - A QR-Code Optical Covert Channel in an Air-Gapped Secure...pdf Adobe PDF (5.07 MB)

SHOW SIMPLE ITEM RECORD

Items in reposiTUm are protected by copyright, with all rights reserved, unless otherwise indicated.

© 2022 TU Wien Legal Notice Policies Support Data Protection Declaration

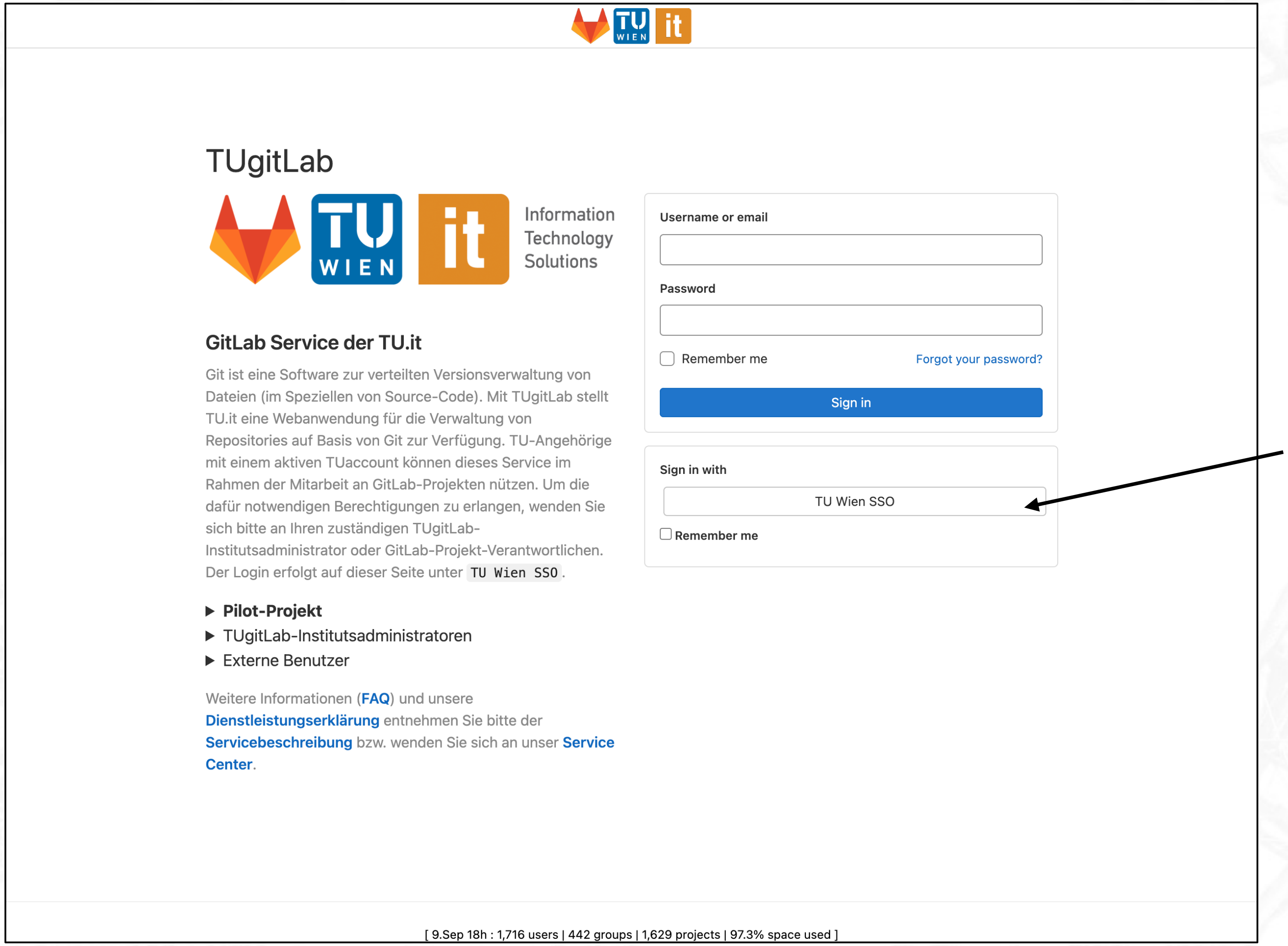
Groupings

Controlled Vocabulary (=machine-readable)

Direct .pdf ressource access

Weise, M, 2021. A QR-Code Optical Covert Channel in an Air-Gapped Secure Data Infrastructure. [Thesis], p.97, DOI: 10.34726/hss.2022.84700

3.2. Gitlab



TUgitLab

GitLab Service der TU.it

Git ist eine Software zur verteilten Versionsverwaltung von Dateien (im Speziellen von Source-Code). Mit TUgitLab stellt TU.it eine Webanwendung für die Verwaltung von Repositories auf Basis von Git zur Verfügung. TU-Angehörige mit einem aktiven TUaccount können dieses Service im Rahmen der Mitarbeit an GitLab-Projekten nützen. Um die dafür notwendigen Berechtigungen zu erlangen, wenden Sie sich bitte an Ihren zuständigen TUgitLab-Institutsadministrator oder GitLab-Projekt-Verantwortlichen. Der Login erfolgt auf dieser Seite unter `TU Wien SSO`.

- ▶ **Pilot-Projekt**
- ▶ TUgitLab-Institutsadministratoren
- ▶ Externe Benutzer

Weitere Informationen (**FAQ**) und unsere **Dienstleistungserklärung** entnehmen Sie bitte der **Servicebeschreibung** bzw. wenden Sie sich an unser **Service Center**.

Username or email:

Password:

□ Remember me

□ Remember me

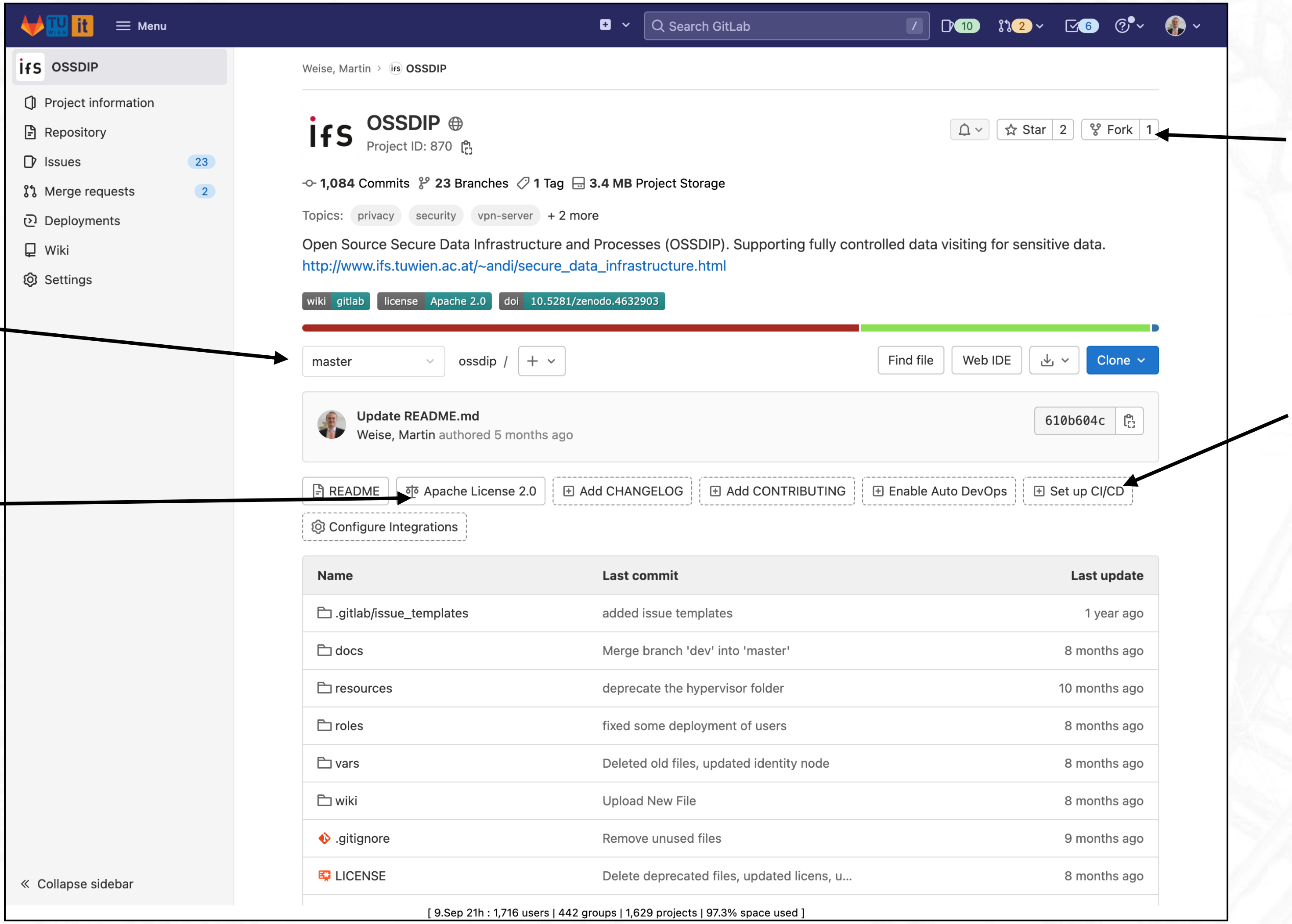
[9.Sep 18h : 1,716 users | 442 groups | 1,629 projects | 97.3% space used]

Single Sign-On (e.g. eduroam is also SSO)

Gitlab. [Online]. URL: https://gitlab.tuwien.ac.at/, accessed 2022-09-09

Martin Weise, Technical University of Vienna

3.2. Gitlab



OSSDIP

- Project information
- Repository
- Issues 23
- Merge requests 2
- Deployments
- Wiki
- Settings

« Collapse sidebar

Weise, Martin › OSSDIP

OSSDIP
Project ID: 870

Star 2 · Fork 1

1,084 Commits · 23 Branches · 1 Tag · 3.4 MB Project Storage

Topics: privacy · security · vpn-server · + 2 more

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

wiki gitlab · license Apache 2.0 · doi 10.5281/zenodo.4632903

master · ossdip / · + · Find file · Web IDE · Clone

Update README.md
Weise, Martin authored 5 months ago · 610b604c

README · Apache License 2.0 · Add CHANGELOG · Add CONTRIBUTING · Enable Auto DevOps · Set up CI/CD · Configure Integrations

Name	Last commit	Last update
.gitlab/issue_templates	added issue templates	1 year ago
docs	Merge branch 'dev' into 'master'	8 months ago
resources	deprecate the hypervisor folder	10 months ago
roles	fixed some deployment of users	8 months ago
vars	Deleted old files, updated identity node	8 months ago
wiki	Upload New File	8 months ago
.gitignore	Remove unused files	9 months ago
LICENSE	Delete deprecated files, updated licens, u...	8 months ago

[9.Sep 21h : 1,716 users | 442 groups | 1,629 projects | 97.3% space used]

Contribute

Software Versioning

License

Continous Integration

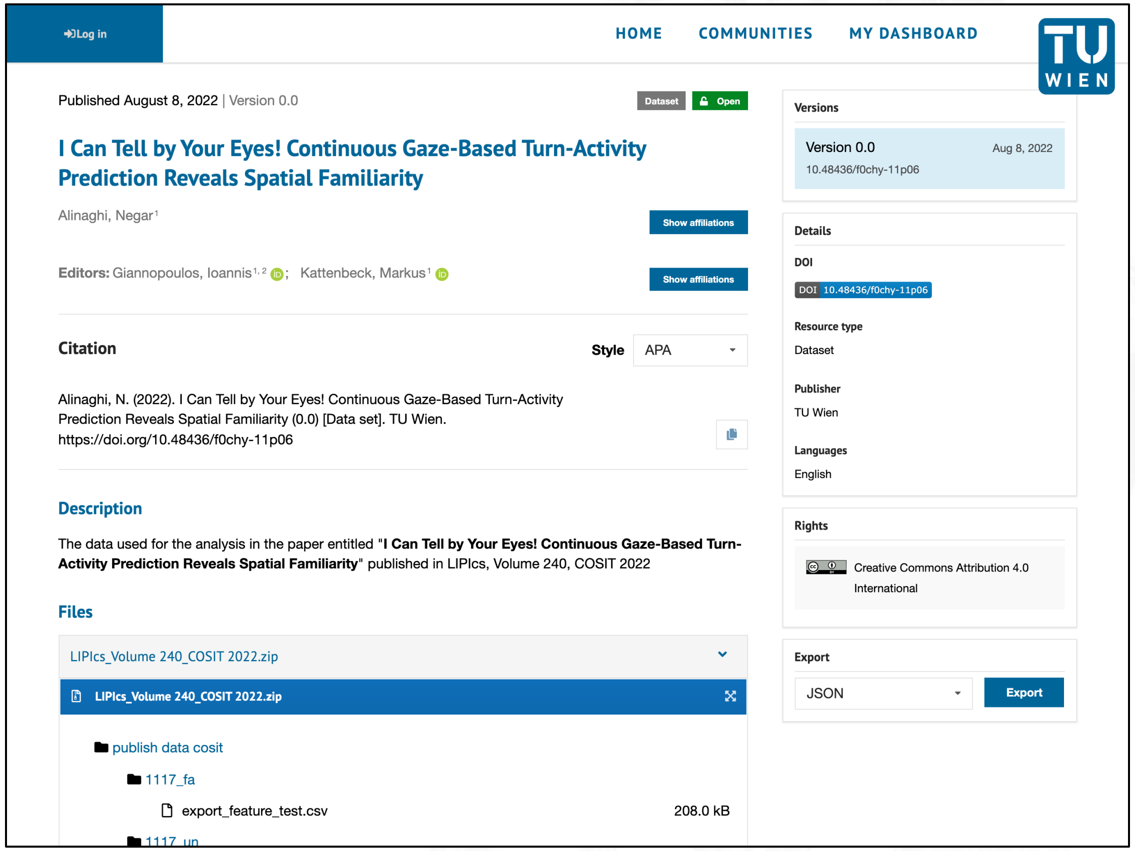
OSSDIP. [Online]. URL: https://gitlab.tuwien.ac.at/martin.weise/ossdip, accessed 2022-09-09

Martin Weise, Technical University of Vienna

3.3. Research Data Repository (InvenioRDM)

- Make digital objects FAIR
- Suitable for **research data**
- Not for publications
  - Other system exists (ReposiTUm)
- Running since December 2020

More details will be presented later today!



Published August 8, 2022 | Version 0.0

I Can Tell by Your Eyes! Continuous Gaze-Based Turn-Activity Prediction Reveals Spatial Familiarity

Alinaghi, Negar[1]

Editors: Giannopoulos, Ioannis[1,2]; Kattenbeck, Markus[1]

Citation

Alinaghi, N. (2022). I Can Tell by Your Eyes! Continuous Gaze-Based Turn-Activity Prediction Reveals Spatial Familiarity (0.0) [Data set]. TU Wien. https://doi.org/10.48436/f0chy-11p06

Description

The data used for the analysis in the paper entitled "**I Can Tell by Your Eyes! Continuous Gaze-Based Turn-Activity Prediction Reveals Spatial Familiarity**" published in LIPIcs, Volume 240, COSIT 2022

Files

LIPIcs_Volume 240_COSIT 2022.zip

LIPIcs_Volume 240_COSIT 2022.zip

- publish data cosit
  - 1117_fa
    - export_feature_test.csv 208.0 kB
  - 1117_un

Versions

Version 0.0 Aug 8, 2022
10.48436/f0chy-11p06

Details

DOI
DOI 10.48436/f0chy-11p06

Resource type
Dataset

Publisher
TU Wien

Languages
English

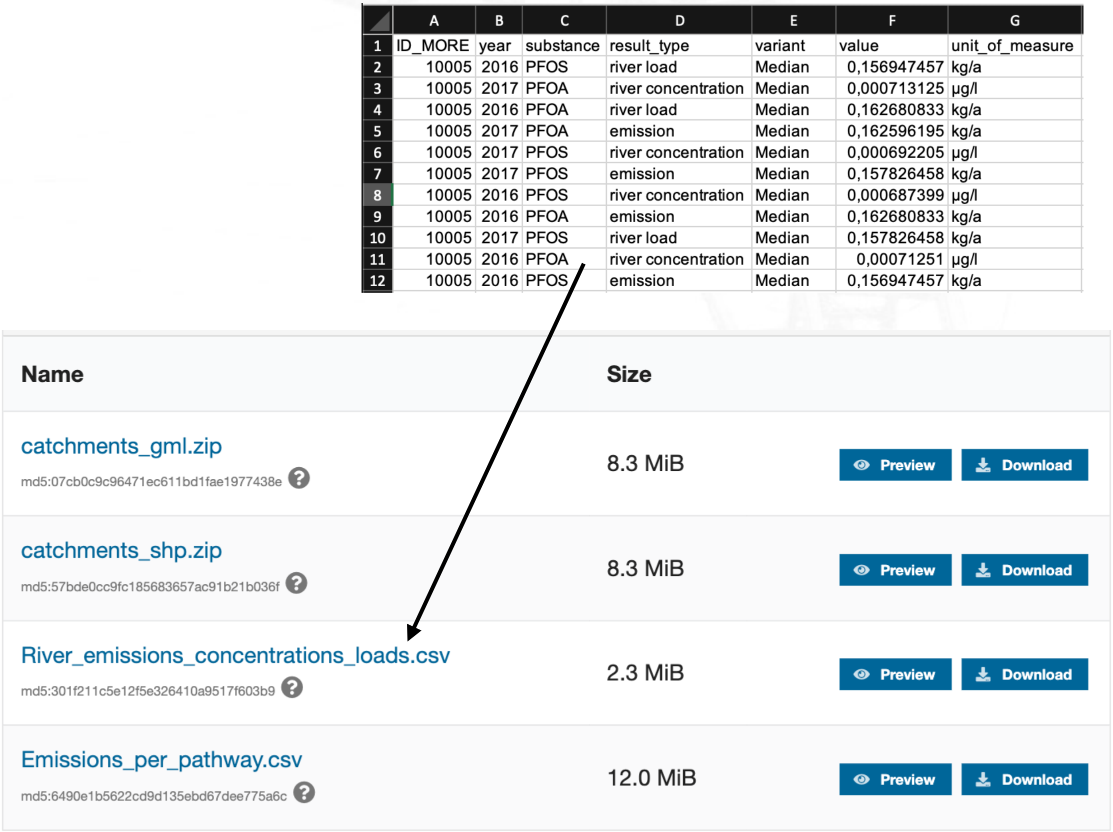
Rights

Creative Commons Attribution 4.0 International

TU Wien Research Data. [Online]. URL: https://researchdata.tuwien.ac.at/, accessed 2022-09-09

3.4. DBRepo

- Invenio can handle collection of files well
- How about relational data in databases?
  - Releasing a data dump every x amount of time?
  - Adding continous data streams, e.g. IoT?
  - How to update / correct data in those databases?
  - Allow reproduction of any subset?
- Private cloud-based environment?
- Dump the data after the end of a project into some repository – to fulfill some grant agreement?



	A	B	C	D	E	F	G
1	ID_MORE	year	substance	result_type	variant	value	unit_of_measure
2	10005	2016	PFOS	river load	Median	0,156947457	kg/a
3	10005	2017	PFOA	river concentration	Median	0,000713125	µg/l
4	10005	2016	PFOA	river load	Median	0,162680833	kg/a
5	10005	2017	PFOA	emission	Median	0,162596195	kg/a
6	10005	2017	PFOS	river concentration	Median	0,000692205	µg/l
7	10005	2017	PFOS	emission	Median	0,157826458	kg/a
8	10005	2016	PFOS	river concentration	Median	0,000687399	µg/l
9	10005	2016	PFOA	emission	Median	0,162680833	kg/a
10	10005	2017	PFOS	river load	Median	0,157826458	kg/a
11	10005	2016	PFOA	river concentration	Median	0,00071251	µg/l
12	10005	2016	PFOS	emission	Median	0,156947457	kg/a

Name	Size	
catchments_gml.zip md5:07cb0c9c96471ec611bd1fae1977438e ?	8.3 MiB	Preview Download
catchments_shp.zip md5:57bde0cc9fc185683657ac91b21b036f ?	8.3 MiB	Preview Download
River_emissions_concentrations_loads.csv md5:301f211c5e12f5e326410a9517f603b9 ?	2.3 MiB	Preview Download
Emissions_per_pathway.csv md5:6490e1b5622cd9d135ebd67dee775a6c ?	12.0 MiB	Preview Download

Kittlaus, S., 2022. Modelled emissions, river loads and river concentrations for PFOA and PFOS in 2016/2017 in Austrian surface waters (1.0) [Data set]. DOI: 10.48436/jpzv9-c8w75

3.4. DBRepo

- Cloud hosted repository for structured research data
- Supports data versioning & citeability via query store and dynamic data citations (Recommendations of RDA WGDC)
- **Microservice architecture**
- Each database encapsulated in a Docker container
- Central **metadata database**
- APIs for different levels of SQL-knowledge:
  - Web interface, support for CSV import,...
  - REST, message queue for data streams

3.4. DBRepo

Database Name

Database Visibility

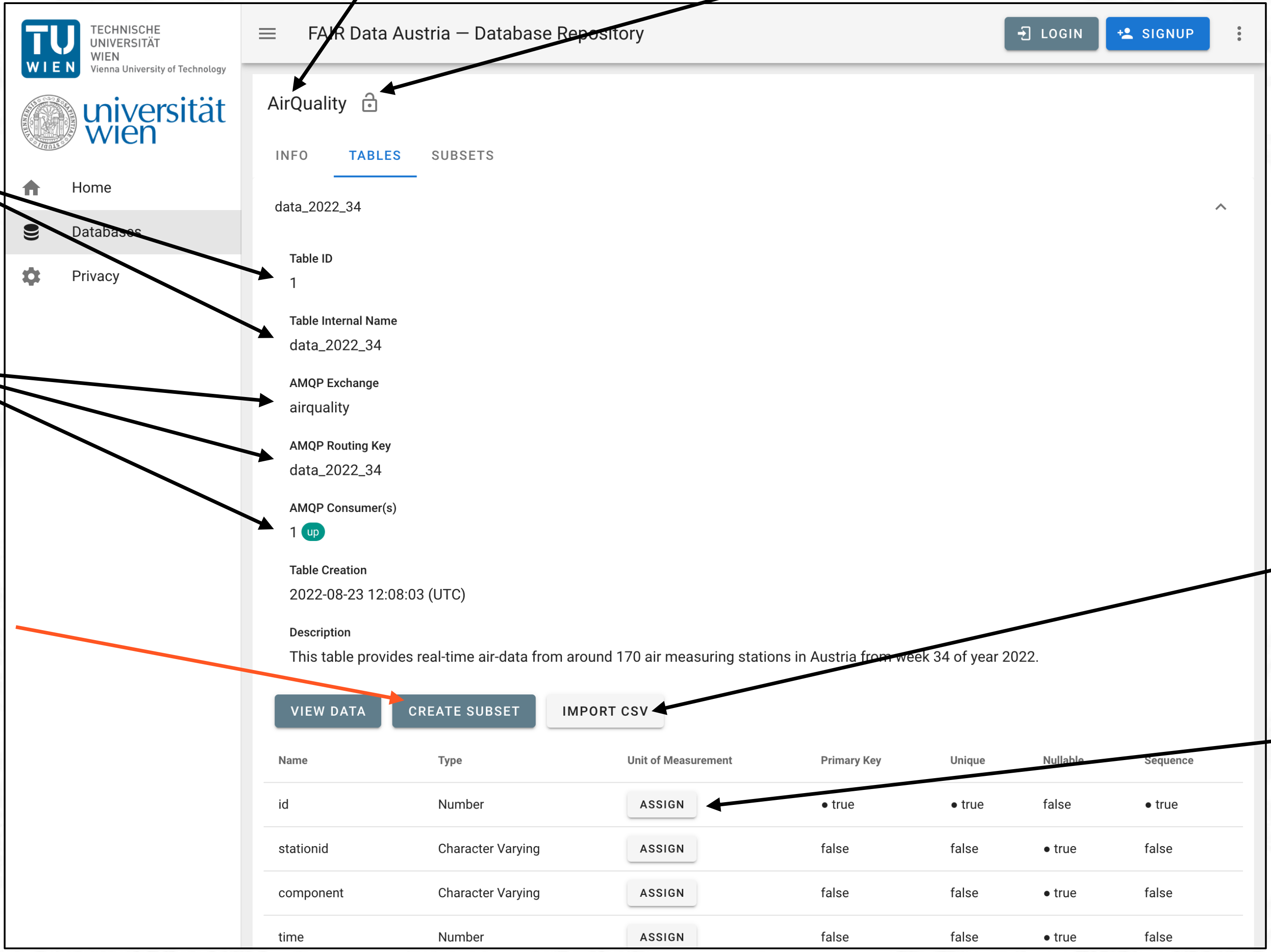
Table information

AMQP information

Create a subset of the data

.csv import

Add metadata


TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology
universität wien
Home
Databases
Privacy
FAIR Data Austria — Database Repository
LOGIN
SIGNUP
INFO
TABLES
SUBSETS
VIEW DATA
CREATE SUBSET
IMPORT CSV

AirQuality

data_2022_34

Table ID
1

Table Internal Name
data_2022_34

AMQP Exchange
airquality

AMQP Routing Key
data_2022_34

AMQP Consumer(s)
1 up

Table Creation
2022-08-23 12:08:03 (UTC)

Description
This table provides real-time air-data from around 170 air measuring stations in Austria from week 34 of year 2022.

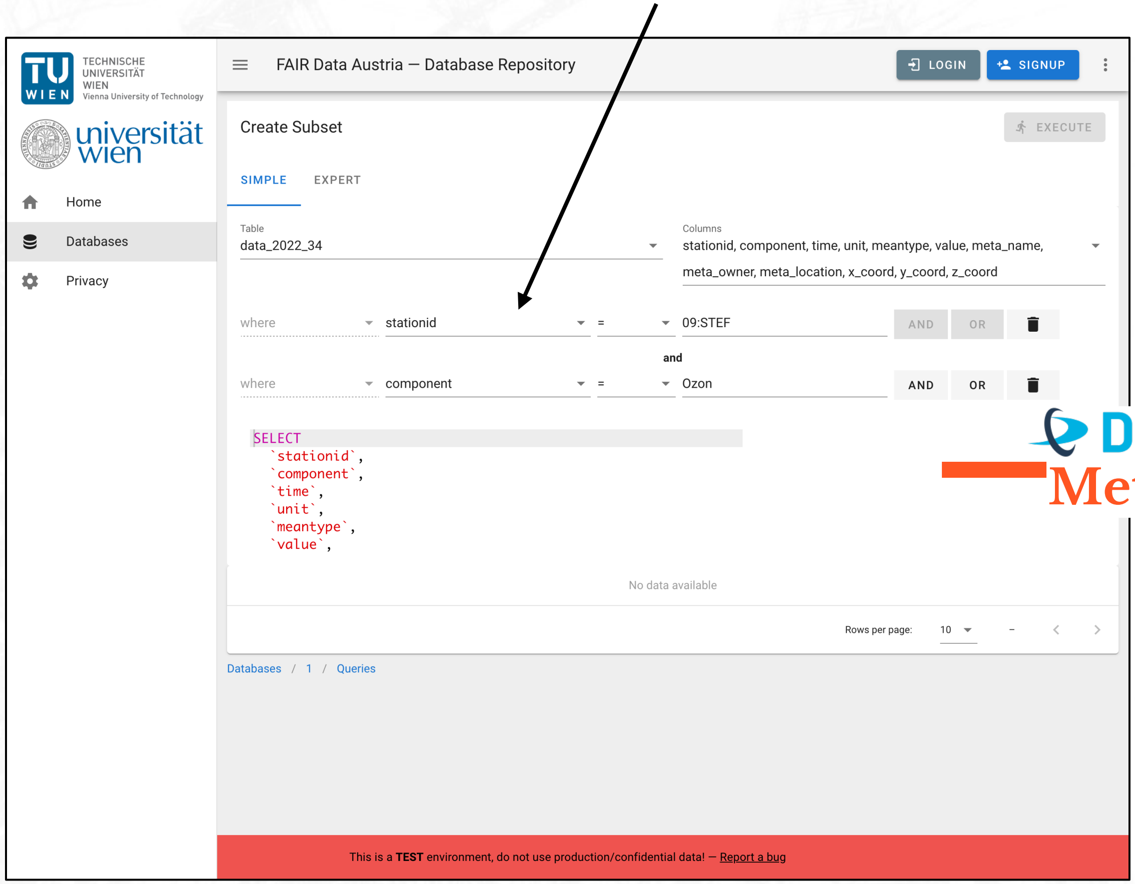
Name	Type	Unit of Measurement	Primary Key	Unique	Nullable	Sequence
id	Number	ASSIGN	● true	● true	false	● true
stationid	Character Varying	ASSIGN	false	false	● true	false
component	Character Varying	ASSIGN	false	false	● true	false
time	Number	ASSIGN	false	false	● true	false

Taha, J. & Weise, M., 2022. AirQuality. [Dataset]. URL: https://dbrepo.ossdip.at/container/1/database/1/table, accessed 2022-09-10

Martin Weise, Technical University of Vienna

3.4. DBRepo

Query Builder for simple subsets



TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology
universität wien
Home
Databases
Privacy
FAIR Data Austria — Database Repository
LOGIN
SIGNUP
EXECUTE
SIMPLE
EXPERT

Create Subset

Table: data_2022_34

Columns: stationid, component, time, unit, meantype, value, meta_name, meta_owner, meta_location, x_coord, y_coord, z_coord

where stationid = 09:STEF AND OR

and

where component = Ozon AND OR

```
SELECT
  `stationid`,
  `component`,
  `time`,
  `unit`,
  `meantype`,
  `value`,
```

No data available

Rows per page: 10 –

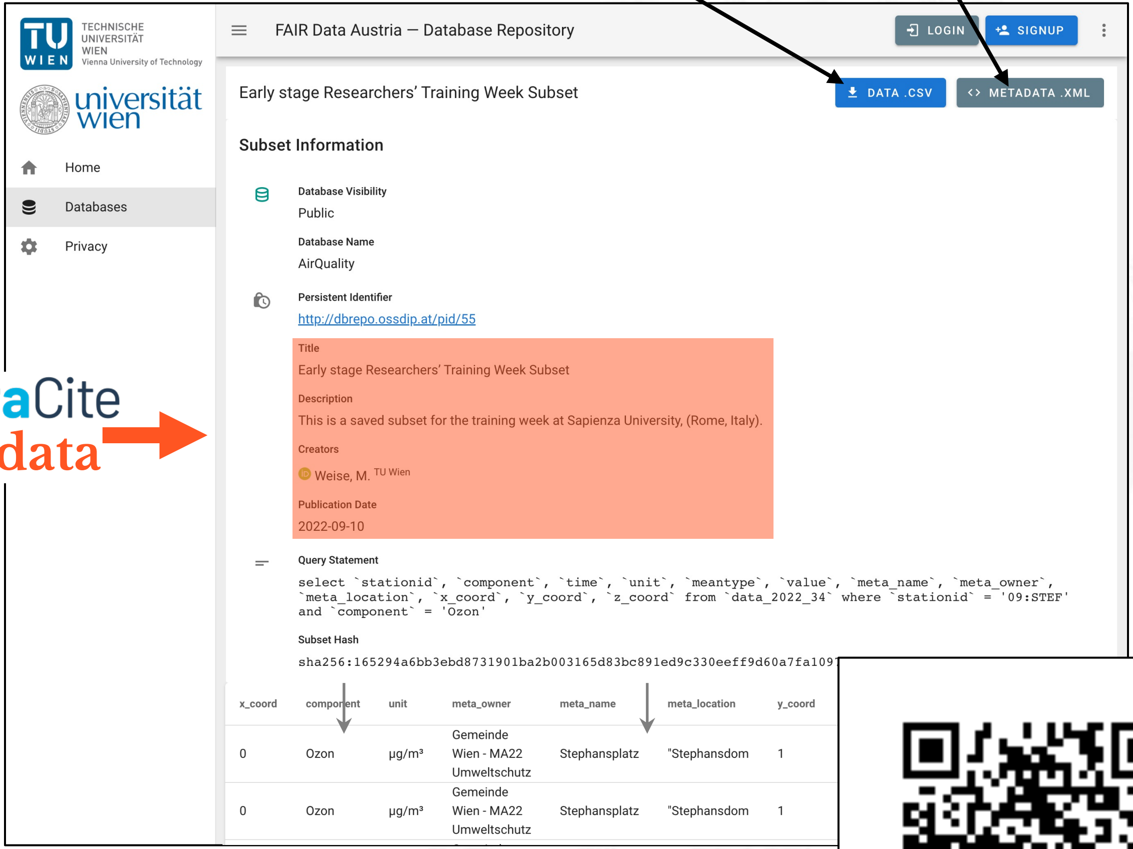
Databases / 1 / Queries

This is a **TEST** environment, do not use production/confidential data! — Report a bug

DataCite Metadata

Export .csv

Export DataCite Metadata



TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology
universität wien
Home
Databases
Privacy
FAIR Data Austria — Database Repository
LOGIN
SIGNUP
DATA .CSV
METADATA .XML

Early stage Researchers' Training Week Subset

Subset Information

Database Visibility
Public

Database Name
AirQuality

Persistent Identifier
http://dbrepo.ossdip.at/pid/55

Title
Early stage Researchers' Training Week Subset

Description
This is a saved subset for the training week at Sapienza University, (Rome, Italy).

Creators
Weise, M. TU Wien

Publication Date
2022-09-10

Query Statement

```
select `stationid`, `component`, `time`, `unit`, `meantype`, `value`, `meta_name`, `meta_owner`,
`meta_location`, `x_coord`, `y_coord`, `z_coord` from `data_2022_34` where `stationid` = '09:STEF'
and `component` = 'Ozon'
```

Subset Hash

sha256:165294a6bb3ebd8731901ba2b003165d83bc891ed9c330eeff9d60a7fa109

x_coord	component	unit	meta_owner	meta_name	meta_location	y_coord
0	Ozon	µg/m³	Gemeinde Wien - MA22 Umweltschutz	Stephansplatz	"Stephansdom	1
0	Ozon	µg/m³	Gemeinde Wien - MA22 Umweltschutz	Stephansplatz	"Stephansdom	1

Have a look yourself

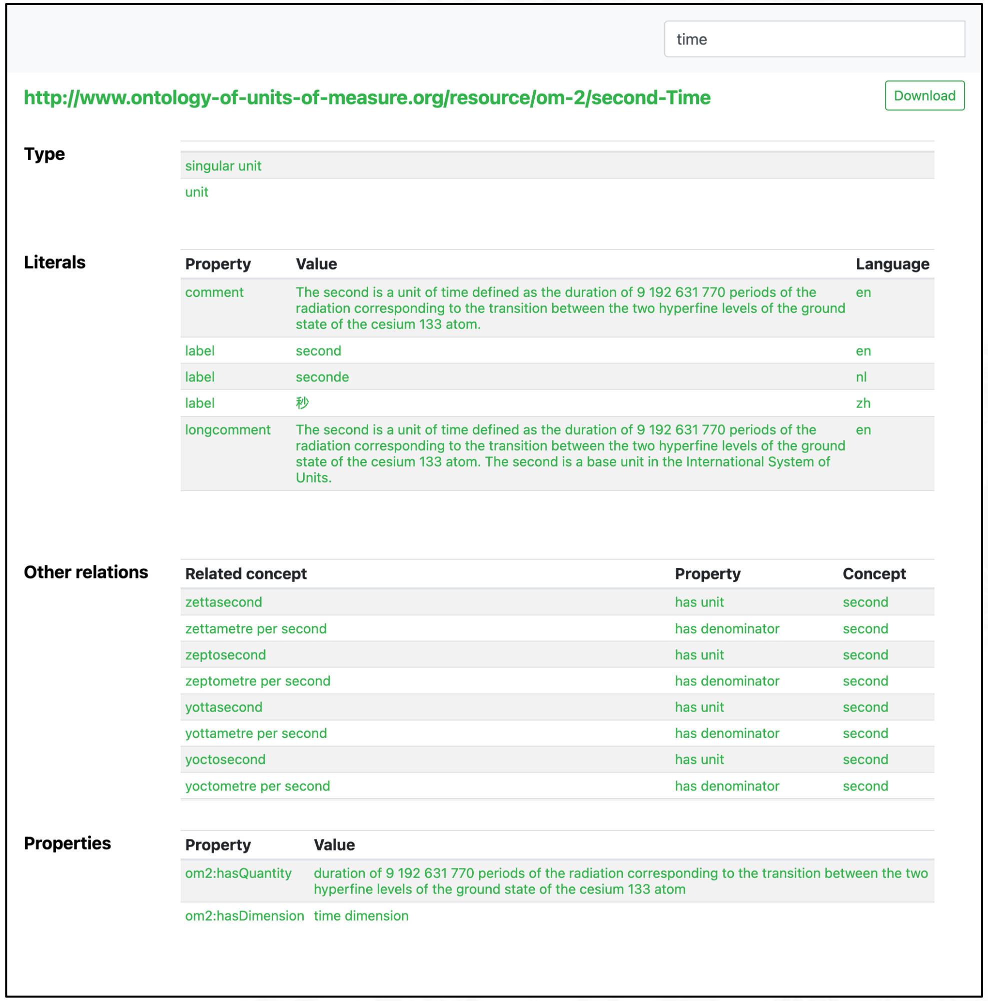
Weise, M., 2022. Early stage Researchers' Training Week Subset. [Dataset]. URL: https://dbrepo.ossdip.at/pid/55, accessed 2022-09-10

Martin Weise, Technical University of Vienna

3.4. DBRepo

Metadata makes databases findable (as in FAIR)

- Container name
- Database name, description, -license
- Table name
  - Column name, -type, -uniqueness, -nullability, -date format
  - Column measurement unit (controlled vocabulary)
  - Statistical properties



http://www.ontology-of-units-of-measure.org/resource/om-2/second-Time

Type

- singular unit
- unit

Literals

Property	Value	Language
comment	The second is a unit of time defined as the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom.	en
label	second	en
label	seconde	nl
label	秒	zh
longcomment	The second is a unit of time defined as the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom. The second is a base unit in the International System of Units.	en

Other relations

Related concept	Property	Concept
zettasecond	has unit	second
zettametre per second	has denominator	second
zeptosecond	has unit	second
zeptometre per second	has denominator	second
yottasecond	has unit	second
yottametre per second	has denominator	second
yoctosecond	has unit	second
yoctometre per second	has denominator	second

Properties

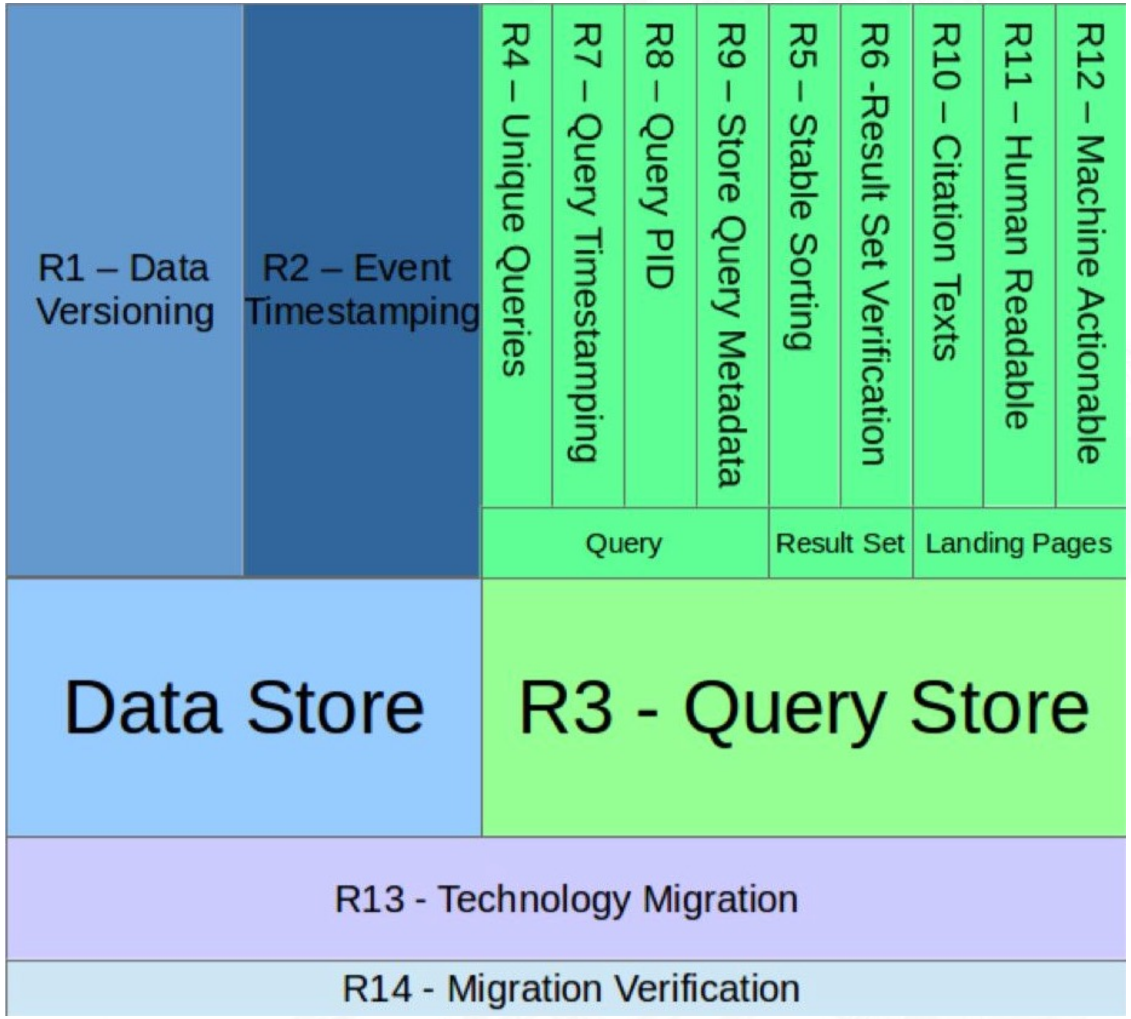
Property	Value
om2:hasQuantity	duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom
om2:hasDimension	time dimension

Foodvoc. [Online]. URL: http://www.ontology-of-units-of-measure.org/resource/om-2/second-Time, accessed 2022-09-10

3.4. DBRepo

Persistent identification of **arbitrary subsets of data**

- Each query issued to the database is saved in the Query Store
- Attaching **metadata** to a query statement, following the DataCite schema
- Mirror the query metadata to DBRepo's central database
- Allows precise data citation
- **Implements recommendations** of the RDA WGDC



R1 – Data Versioning

R2 – Event Timestamping

R4 – Unique Queries

R7 – Query Timestamping

R8 – Query PID

R9 – Store Query Metadata

R5 – Stable Sorting

R6 -Result Set Verification

R10 – Citation Texts

R11 – Human Readable

R12 – Machine Actionable

Query

Result Set

Landing Pages

Data Store

R3 - Query Store

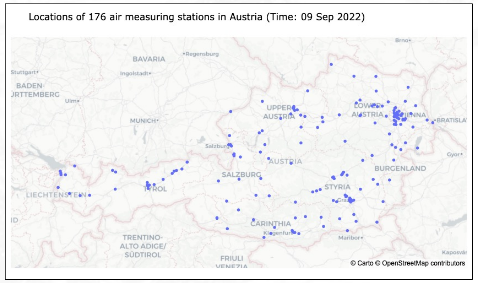
R13 - Technology Migration

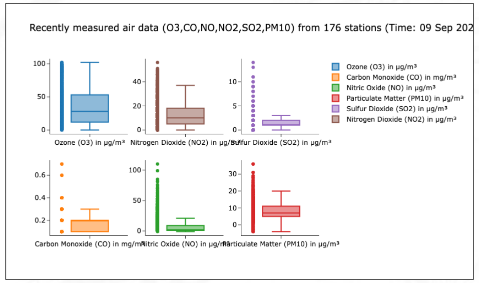
R14 - Migration Verification

Rauber, A., et al, 2021. Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data, in *Harvard Data Science Review*, 3(4). DOI: 10.1162/99608f92.be565013

3.4. DBRepo

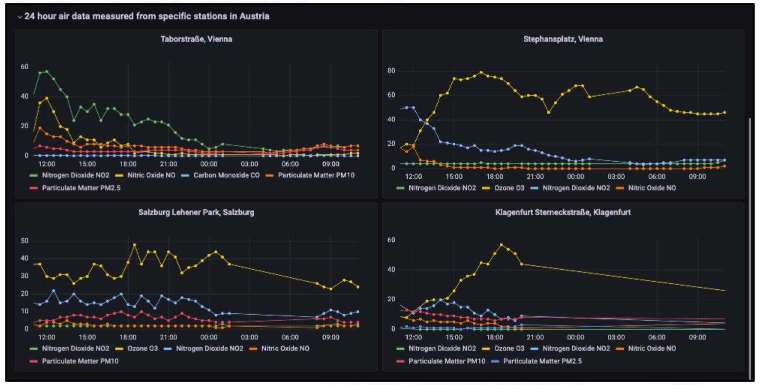
Simple use-case:

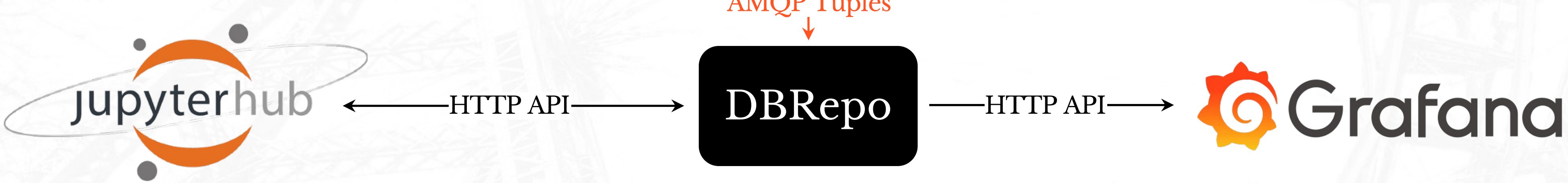

Locations of 176 air measuring stations in Austria (Time: 09 Sep 2022)


Recently measured air data (O3,CO,NO,NO2,SO2,PM10) from 176 stations (Time: 09 Sep 202

https://luft.umweltbundesamt.at/

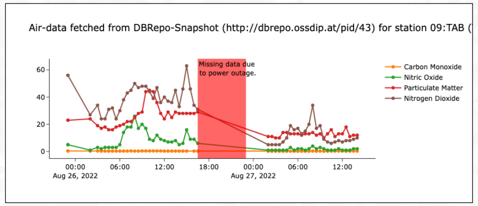
AMQP Tuples

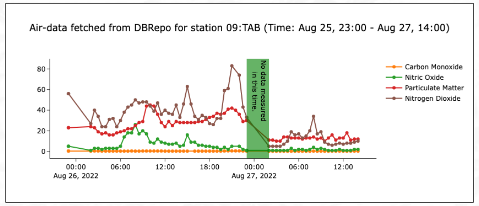


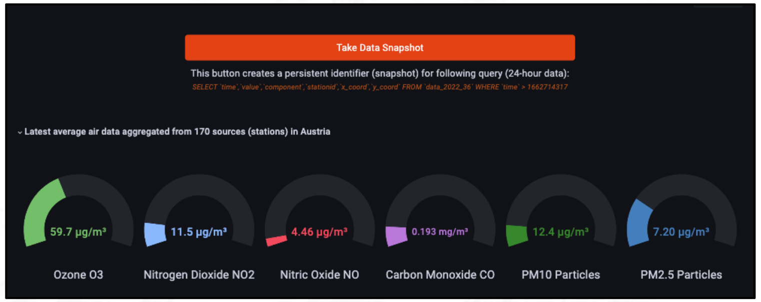

jupyterhub ←HTTP API→ DBRepo —HTTP API→ Grafana

https://s125.dl.hpc.tuwien.ac.at/user/retropotato/lab

http://s125.dl.hpc.tuwien.ac.at:8080/grafana/d/R4B-UWZVz/dbrepo-airquality-austria


Air-data fetched from DBRepo-Snapshot (http://dbrepo.ossdip.at/pid/43) for station 09:TAB (


Air-data fetched from DBRepo for station 09:TAB (Time: Aug 25, 23:00 - Aug 27, 14:00)



This button creates a persistent identifier (snapshot) for following query (24-hour data):

```
SELECT `time`,`value`,`component`,`stationid`,`x_coord`,`y_coord` FROM `data_2022_36` WHERE `time` > 1662714317
```

Latest average air data aggregated from 170 sources (stations) in Austria

3.4. DBRepo Further Reading

Material

- https://indico.egi.eu/event/5882/contributions/16724/ (EGI’22 Poster)
- https://doi.org/10.5281/zenodo.6637333 (IDCC’22 paper)
- https://doi.org/10.17605/OSF.IO/B7NX5 (iPRES’21 paper)

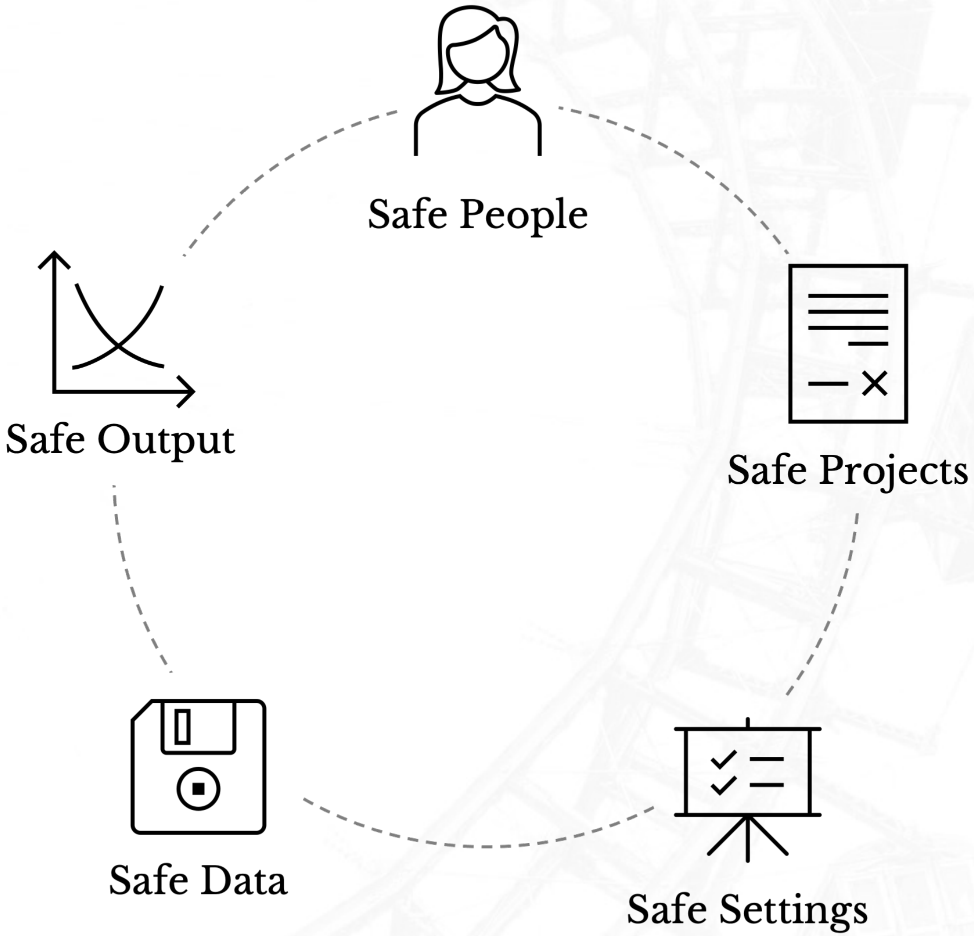
Resources

- https://dbrepo.ossdip.at (sandbox)
- https://dbrepo-docs.ossdip.at (documentation, getting started guide)
- https://gitlab.phaidra.org/fair-data-austria-db-repository/fda-services (source code)

4. Trusted Research Environments

”five safes” dimensions:

1. *Safe projects*, **appropriateness** of the usage of the data
2. *Safe people*, **identify** users that access senstive data, legal bindings
3. *Safe data*, **appropriate** data de-identification, research questions formulated
4. *Safe settings*, **necessity** of security and transparency
5. *Safe outputs*, **approved**, aggregated research results can be exported



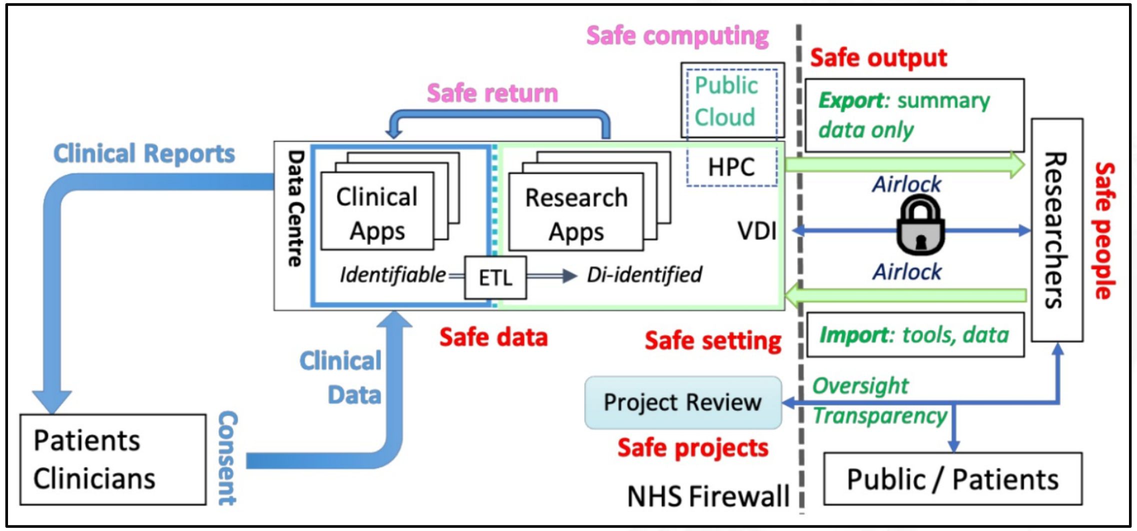
Desai, T., 2016. Five Safes: Designing Data Access for Research. [Online]. URL: https://uwe-repository.worktribe.com/output/914745/five-safes-designing-data-access-for-research, accessed 2022-09-10

4. Trusted Research Environments

UK Health Data
Research Alliance

- United Kingdom Health Data Research Alliance (UKHDRA)
- Confederation of leading organizations in the healthcare field
- Extend “five safes”

6. Safe return, allow de-identified research results to be re-identified and securely mapped back to the original data set



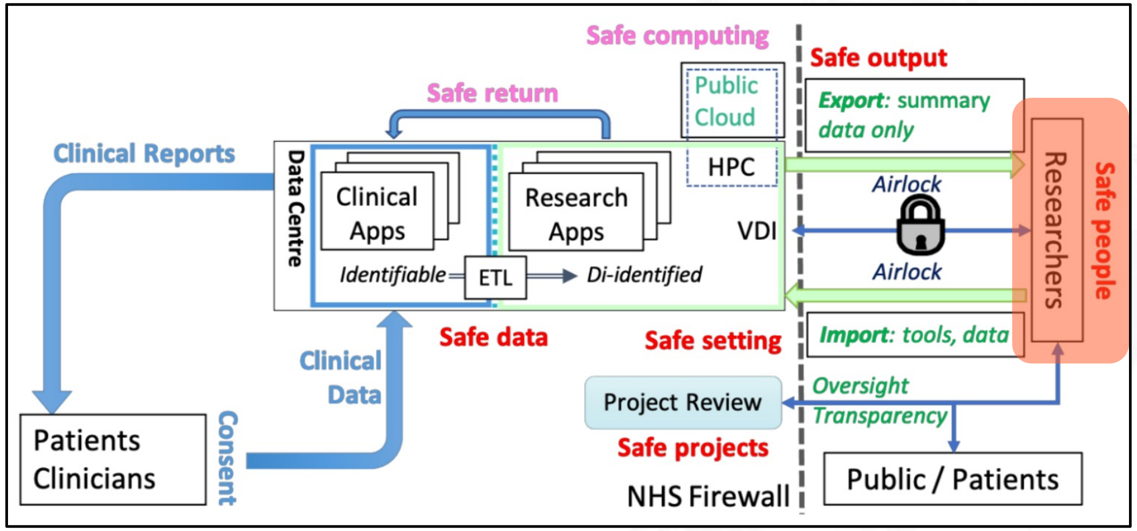
Ensure public trust by implementing UKHDRA’s recommendations on TRE+ and independent accreditation and audit

UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

UK Health Data Research Alliance

1. *Safe people*
   - Control & ensure interests of people where data
     - Collected
     - Analyzed
   - Sign legally-binding documents (e.g. NDAs)
   - Analysts must undergo information governance training (once approved, access all)

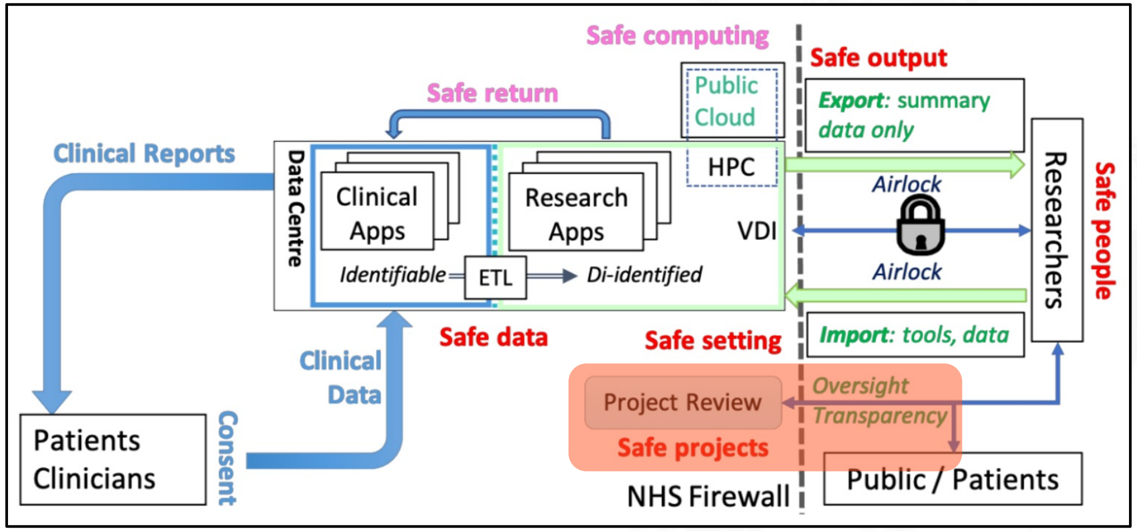


UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

UK Health Data Research Alliance

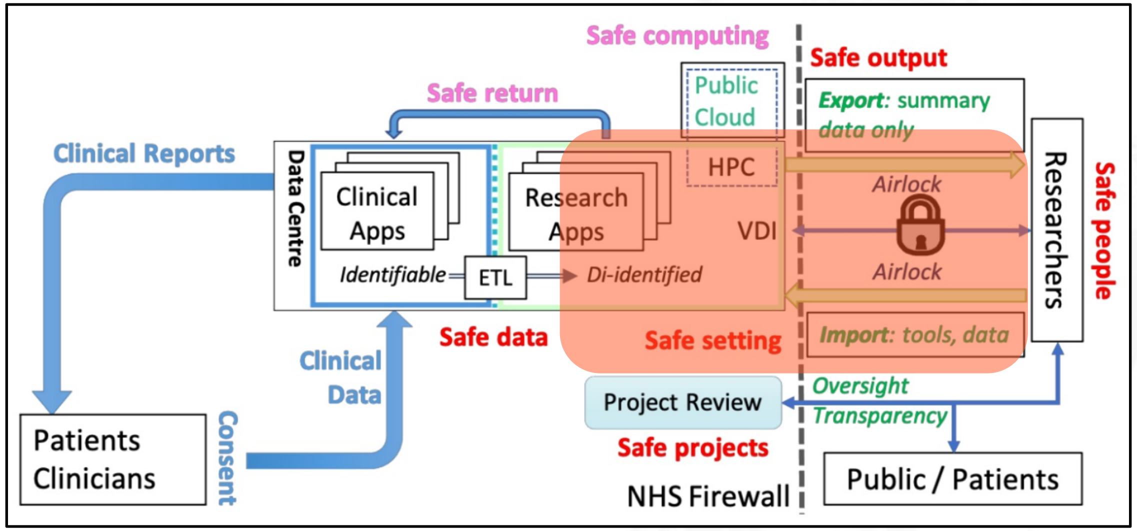
2. *Safe projects*
   - **Appropriate** use of sensitive data
   - Mandatory possibility to external audits
   - **Ethics board** (review the project proposal and gives a clearance)
   - Want to **improve the maturity** of the project management processes



UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

3. *Safe setting*
   - Ensure a straightforward to use, secure environment for the sensitive data to reside in
   - Defined and transparent process
   - Trusted administrators with permission to exchange data through the air-gap
   - System-internal barriers
   - Protected individual data cannot be exported
   - Actions of authorized analysts are monitored

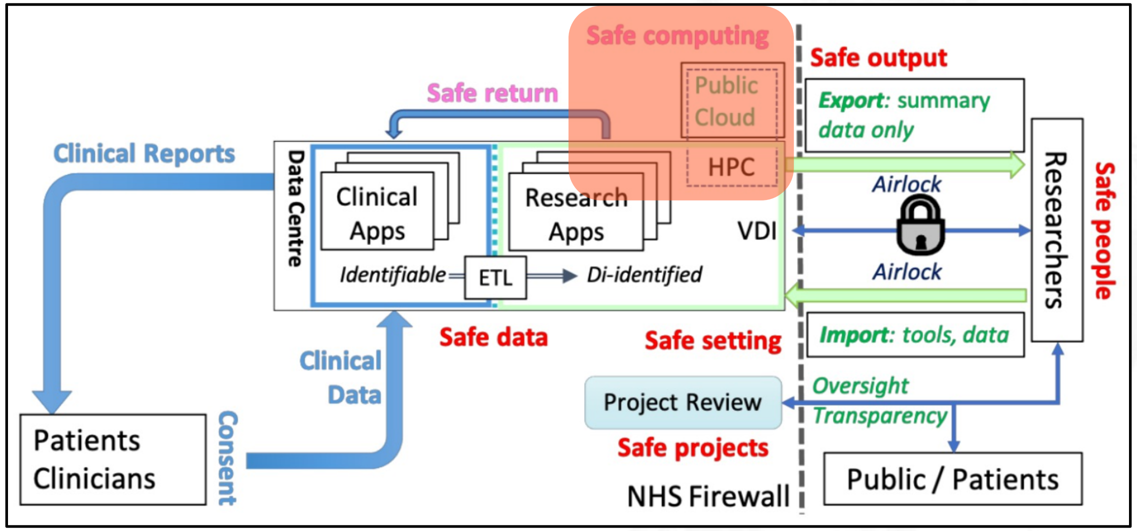


UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

UK Health Data Research Alliance

4. *Safe computing*
   - Outsourcing infrastructure (e.g. content delivery networks)
   - Overcome the risk of exposing sensitive data to public cloud providers
   - Additional safeguards that disallow any outsourced hardware or software
   - Access the sensitive data at any time must be implemented
   - This approach is well-studied and supported from major public cloud providers.

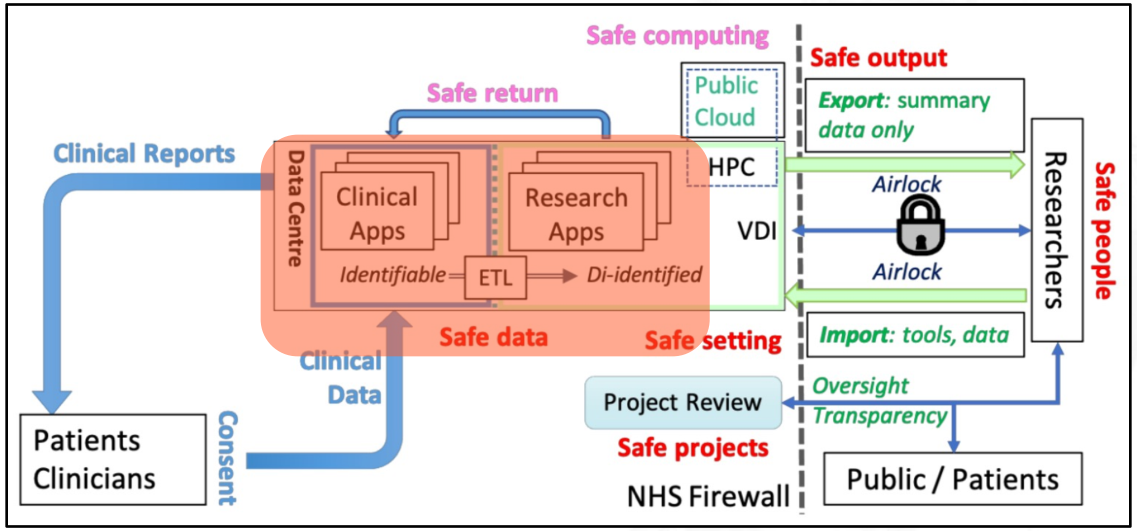


UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

UK Health Data Research Alliance

5. *Safe data*
   - Minimize risk of (accidental) re-identification of protected individuals when importing data into the TRE
   - De-identification software tools and encrypted (virtual) disks
   - Prioritizes the interests of the protected individual over the analyst through technical & organizational measures

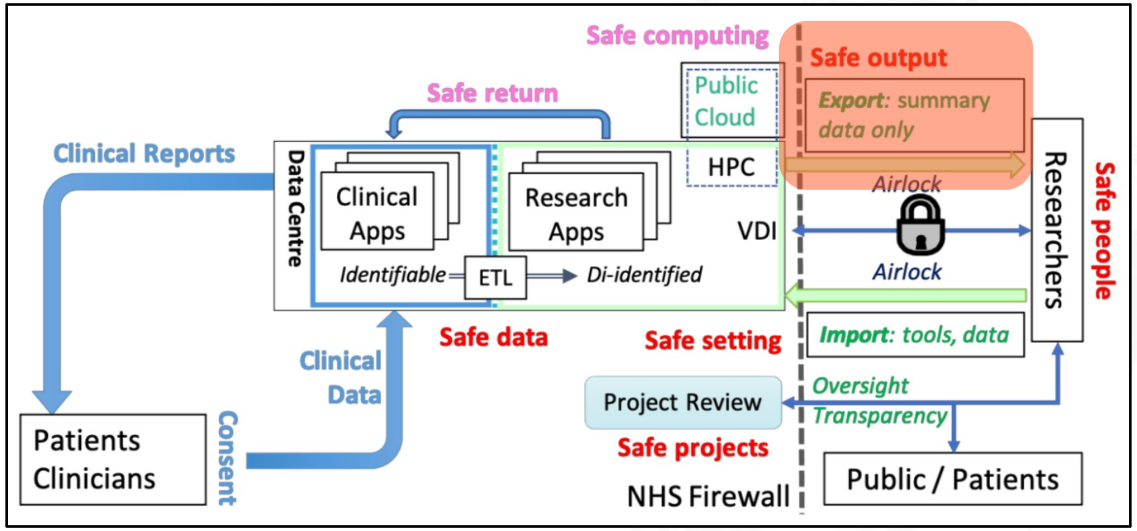


UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

UK Health Data Research Alliance

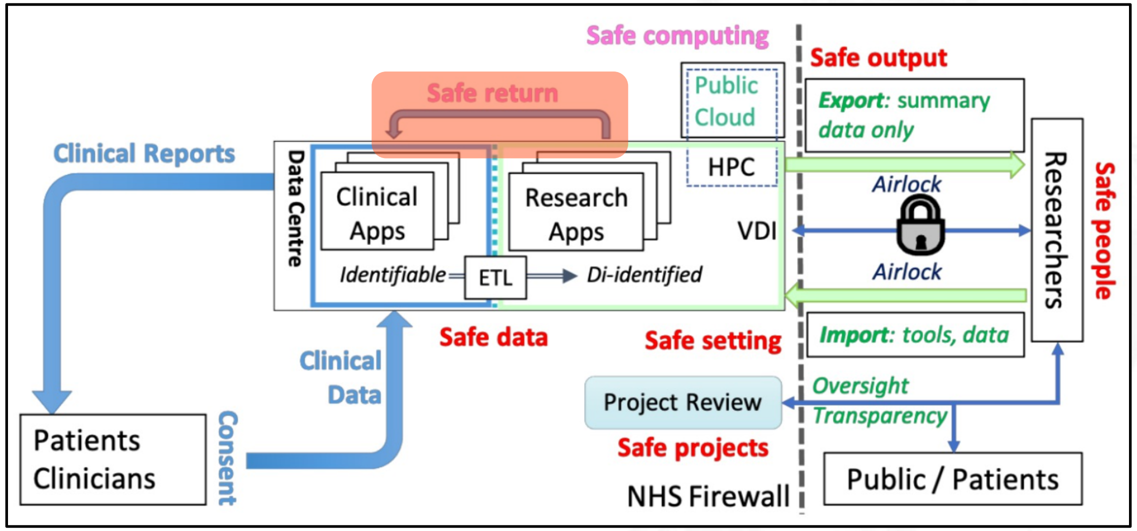
6. *Safe outputs*
   - Barrier ("air-gap") of the safe setting and the open internet
   - Data Ingress/egress process
     - Manual interaction to control risk of disclosure
   - Manages the communication to the outside world



UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4. Trusted Research Environments

7. *Safe return*
   - De-identified research output to be returned into the TRE where the sensitive data comes from and the identity of the protected individual is known
   - Gain information about the protected individual itself
   - Profit from allowing analysts to work with the data
   - Researchers can only access de-identified sensitive data while all their actions are overseen by a committee



Upon approval of both the patient and the ethics board, the research output can be re-identified and mapped back to the original data set to enrich value to it.

UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. DOI: 10.5281/zenodo.5767586

4.1. RemoteNEPS

NEPS

National Education Panel Study (NEPS)

- Remote access to data to **utilize** them better than local access
- Better than conventional methods
  - Remote execution
  - Job-submission systems

  } Queues, input-output (heavily) **delayed**
- Hosts a full-fledged secure data infrastructure
- Data access is **moderately** anonymized, the Analyst must sign an additional supplementary agreement
- Data export is possible via link after signing data us agreements (**heavily** anonymized and aggregated)

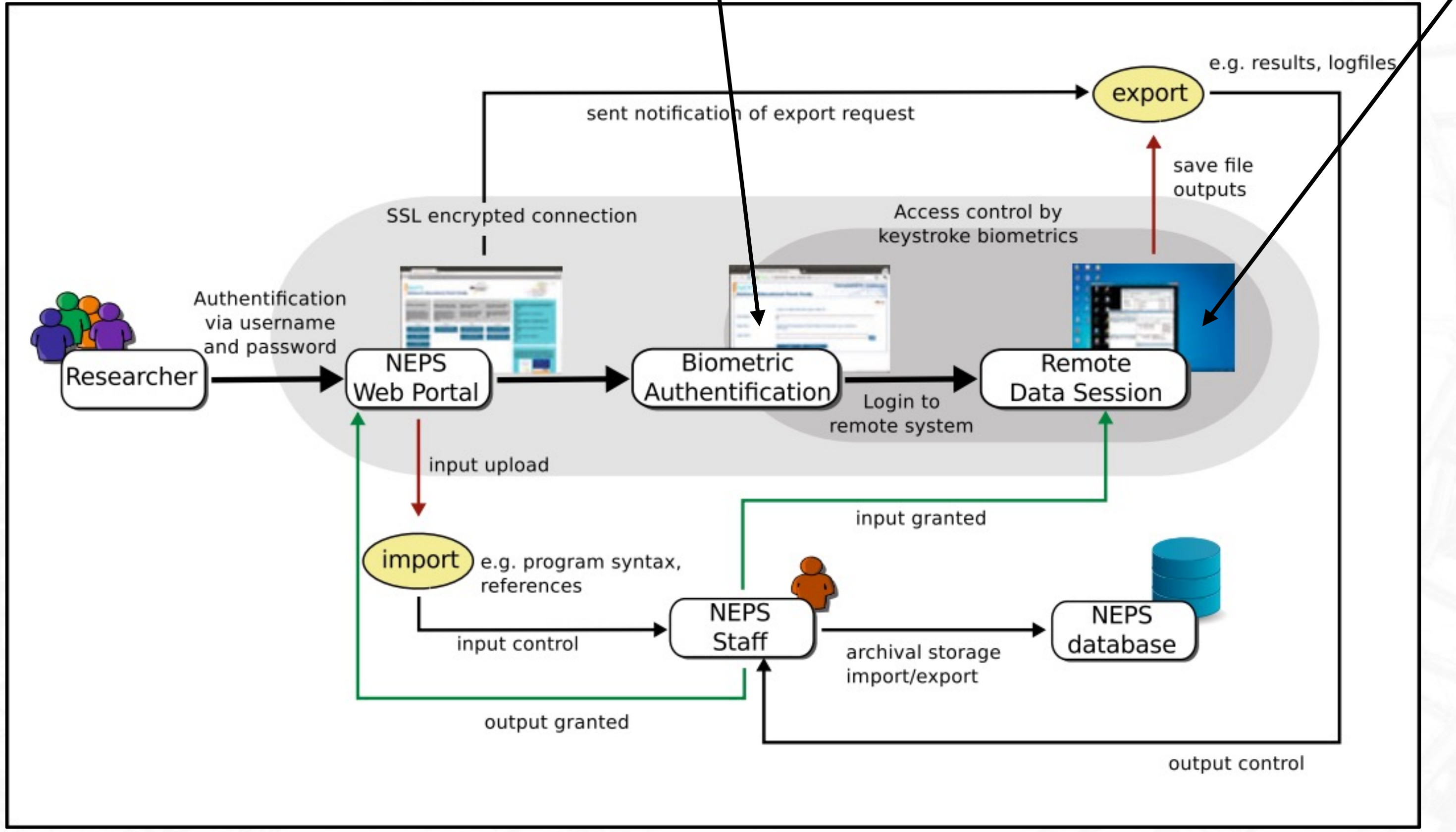
Skopek, J., 2016. RemoteNEPS - An Innovative Research Environment, in *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study*, p.611-626. DOI: 10.1007/978-3-658-11994-2_34

4.1. RemoteNEPS

Social Architecture

Second-factor
(prevent credential sharing)

SPSS, R, Office,
Windows 7



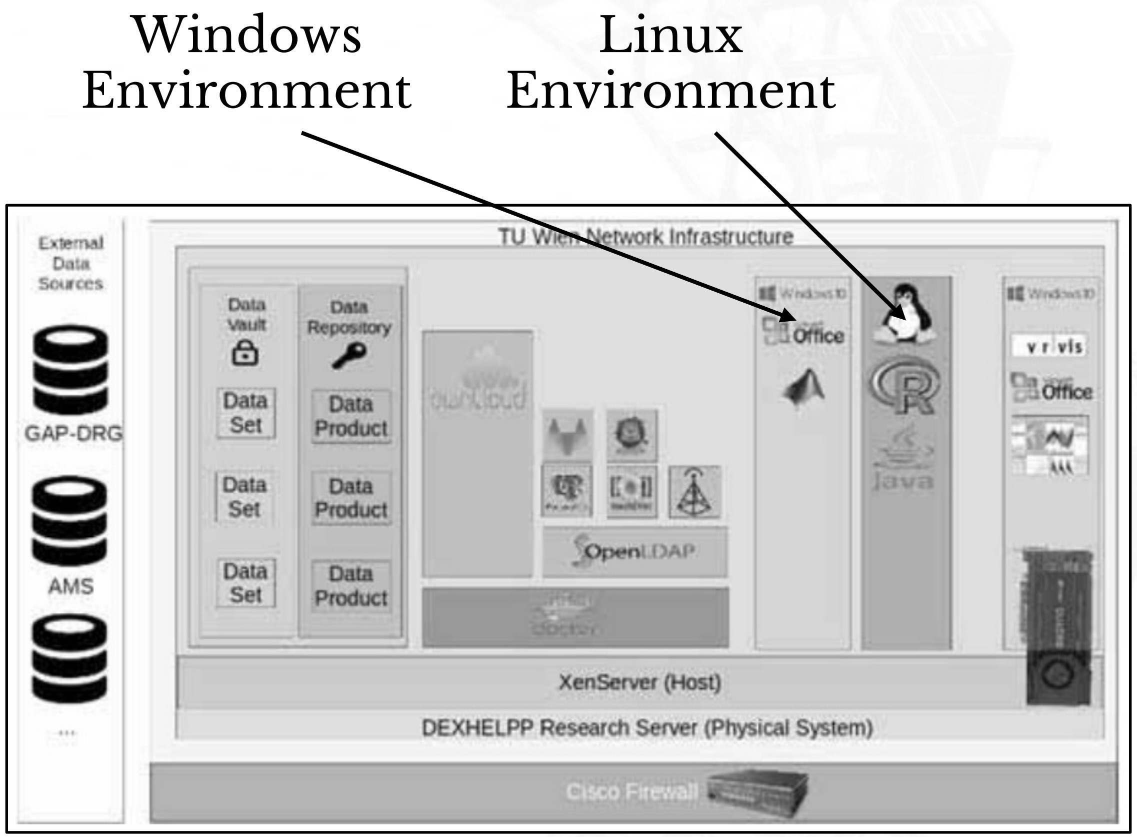
Skopek, J., 2016. RemoteNEPS - An Innovative Research Environment, in *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study*, p.611-626. DOI: 10.1007/978-3-658-11994-2_34

Martin Weise, Technical University of Vienna

4.2. DEXHELPP

Facilitates research in Austria for almost 10 years

- Provide **analysts** with a secure and controlled environment **without the need** to exfiltrate data out of the system
- Data owners on the other side **deposit** their data from **heterogeneous** sources in an encrypted vault and specify **fine-grained access** rights, e.g.
  - Entire data assets
  - Just specific subsets



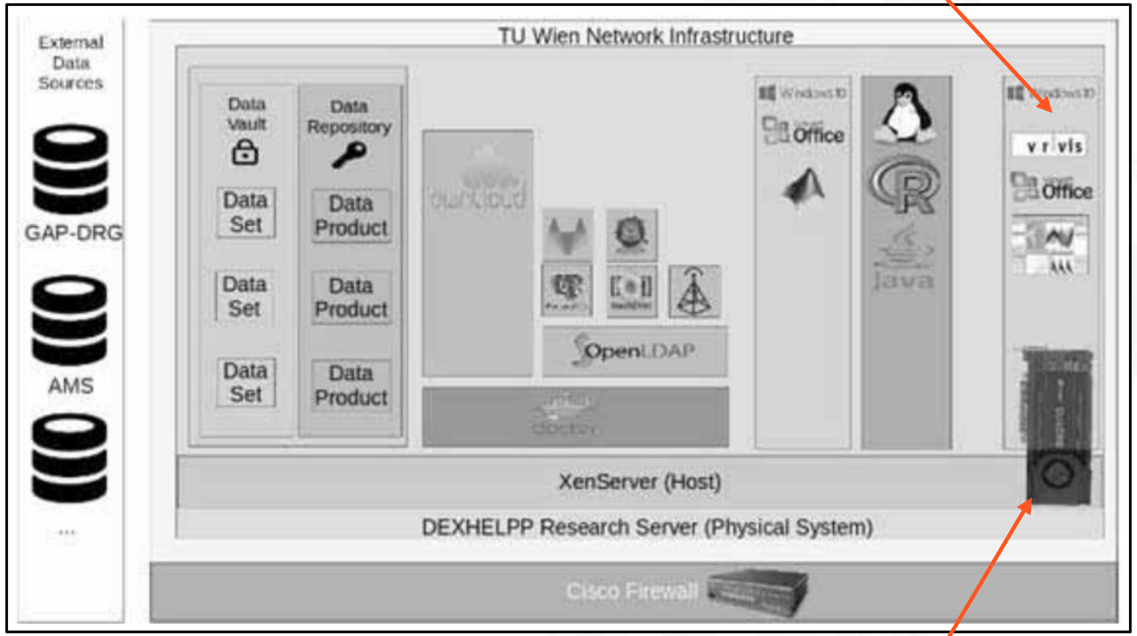
Windows Environment

Linux Environment

Popper, N., 2017. Planning Future Health: Developing Big Data and System Modelling Pipelines for Health System Research, in Simulation Notes Europe, 27(4), p.203-208 2017. DOI: 10.11128/sne.27.tn.10396

4.2. DEXHELPP

- Monitoring Node continuously **monitors** the access to the Data Endpoint (allows for **auditing** and inspection of the usage of the data at any time)
- **Docker environment** (e.g. PostgreSQL, RStudio, Web Applications)
- Analysts **working** on the Remote Desktop Node, connected through the VPN Client using two-factors
- Special hardware and hypervisor for GPU execution



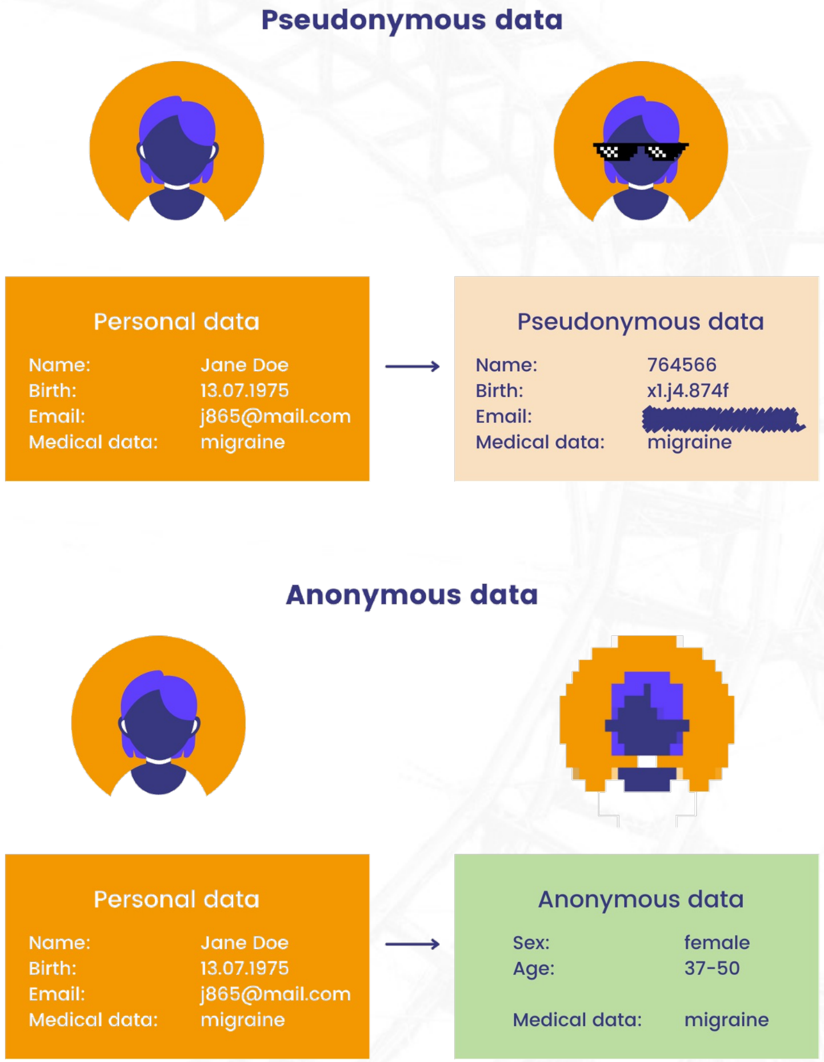
Accelerated Environment

Special Hardware

Popper, N., 2017. Planning Future Health: Developing Big Data and System Modelling Pipelines for Health System Research, in Simulation Notes Europe, 27(4), p.203-208 2017. DOI: 10.11128/sne.27.tn.10396

4.3. OSSDIP

- Sensitive Data (=due privacy issues, commercial interests, ...)
- Provide access for analysis, but ensure data is not leaked or misused
- Standard approach
  - Pseudonymization
  - Anonymization
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness



Pseudonymous data

Personal data
Name: Jane Doe
Birth: 13.07.1975
Email: j865@mail.com
Medical data: migraine

→

Pseudonymous data
Name: 764566
Birth: x1.j4.874f
Email: [redacted]
Medical data: migraine

Anonymous data

Personal data
Name: Jane Doe
Birth: 13.07.1975
Email: j865@mail.com
Medical data: migraine

→

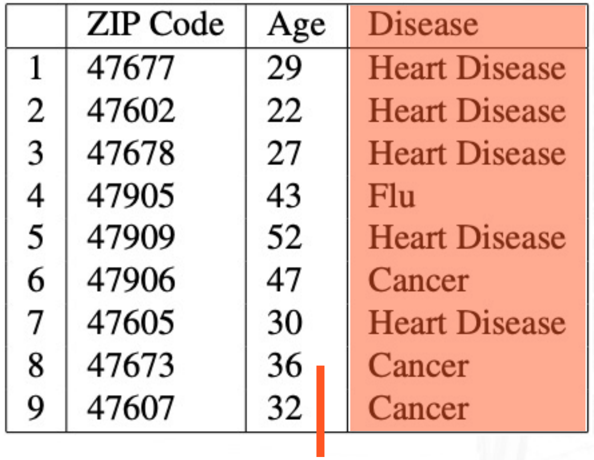
Anonymous data
Sex: female
Age: 37-50
Medical data: migraine

4.3. OSSDIP

EOSC-Life

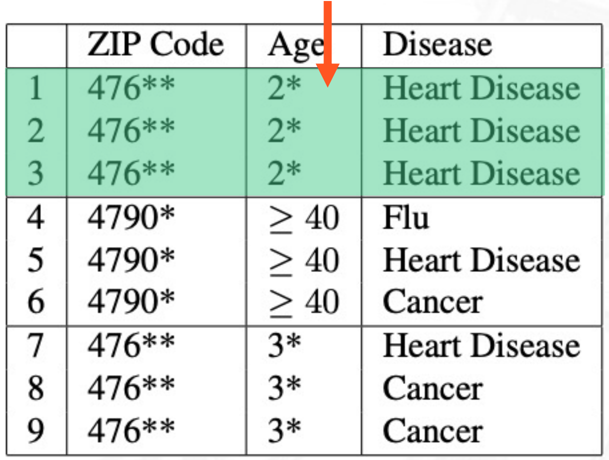
- Sensitive Data (=due privacy issues, commercial interests, …)
- Provide access for analysis, but ensure data is not leaked or misused
- Standard approach
  - Pseudonymization
  - Anonymization
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness

sensitive attribute



	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

3-anonymity



	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Alice knows **Bob** is 27yo, lives in 47678 and is in the first three entries

$\subseteq$ Bob has a **heart diseaese**

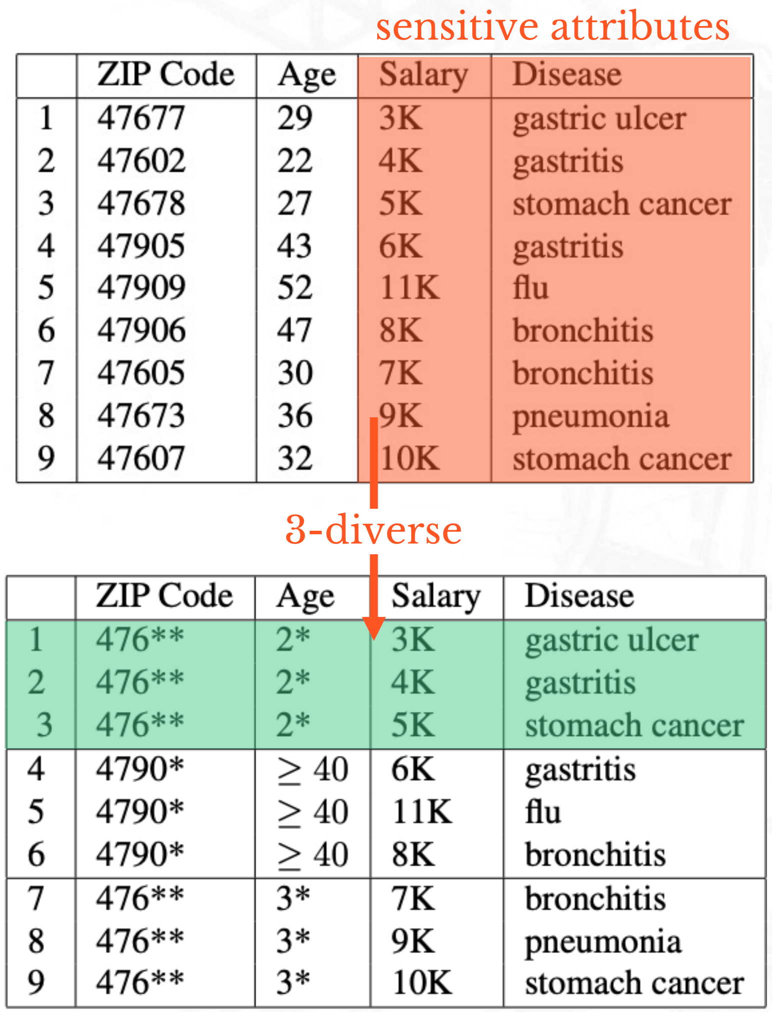
Samarati, P., 2001. Protecting Respondents Identities in Microdata Release, in *IEEE Transactions on Knowledge and Data Engineering, 13(6). DOI: 10.1109/69.971193*

4.3. OSSDIP

- Sensitive Data (=due privacy issues, commercial interests, ...)
- Provide access for analysis, but ensure data is not leaked or misused
- Standard approach
  - Pseudonymization
  - Anonymization
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness

Alice knows **Bob** is 27yo, lives in 47678 and is in the first three entries

⊆ **Bob** has a **low salary** and some **stomach disease**



sensitive attributes

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

3-diverse

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

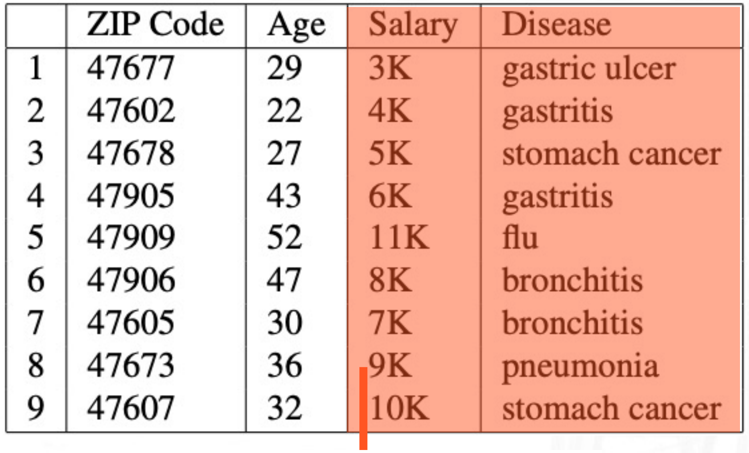
Machanavajjhala, A., et al 2006. L-diversity: Privacy Beyond k-anonymity, in *Proceedings of the 22nd International Conference on Data Engineering*. DOI: 10.1109/ICDE.2006.1

4.3. OSSDIP

EOSC-Life

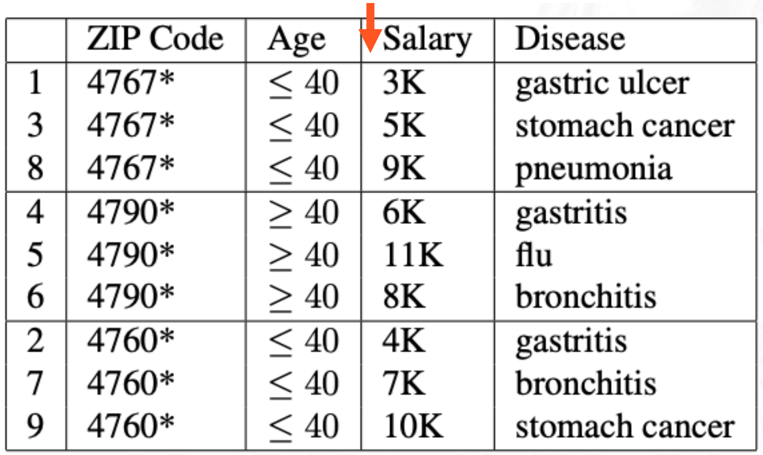
- Sensitive Data (=due privacy issues, commercial interests, ...)
- Provide access for analysis, but ensure data is not leaked or misused
- Standard approach
  - Pseudonymization
  - Anonymization
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness

sensitive attributes



	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

0.167-closeness (Salary)
0.278-closeness (Disease)



	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Alice knows **Bob** is 27yo, lives in 47678 and is in the first three entries

$\subseteq$ Bob has a ???

Ninghui, L., et al, 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, in *Proceedings of the 23rd International Conference on Data Engineering*. DOI: 10.1109/ICDE.2007.367856

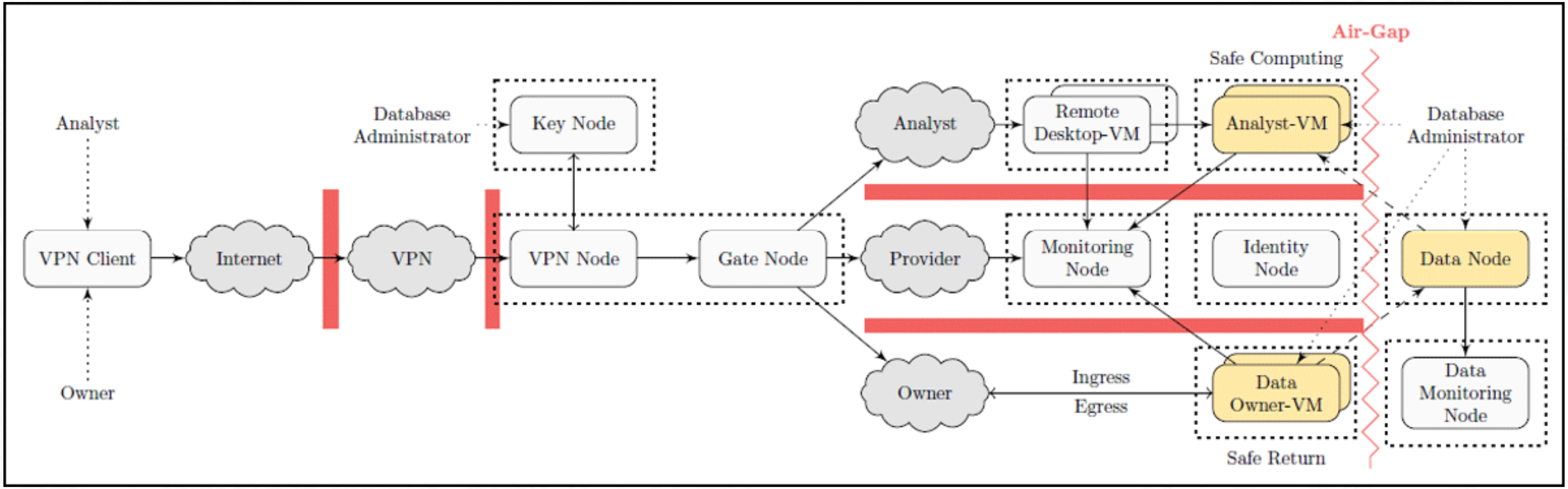
4.3. OSSDIP

- **Data Owner** maintains full control over data and use:
  - Who to **allow access**,
  - Over which **period of time**,
  - For which **subset of data**,
  - To answer which research question / analysis goals,
  - While monitoring what they are doing
- Based on experience of operating DEXHELPP for nearly 10 years

Weise, M., et al, 2022. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting, in *Data Science Journal*, 21(1). DOI: 10.5334/dsj-2022-004

4.3. OSSDIP

Technical Architecture:



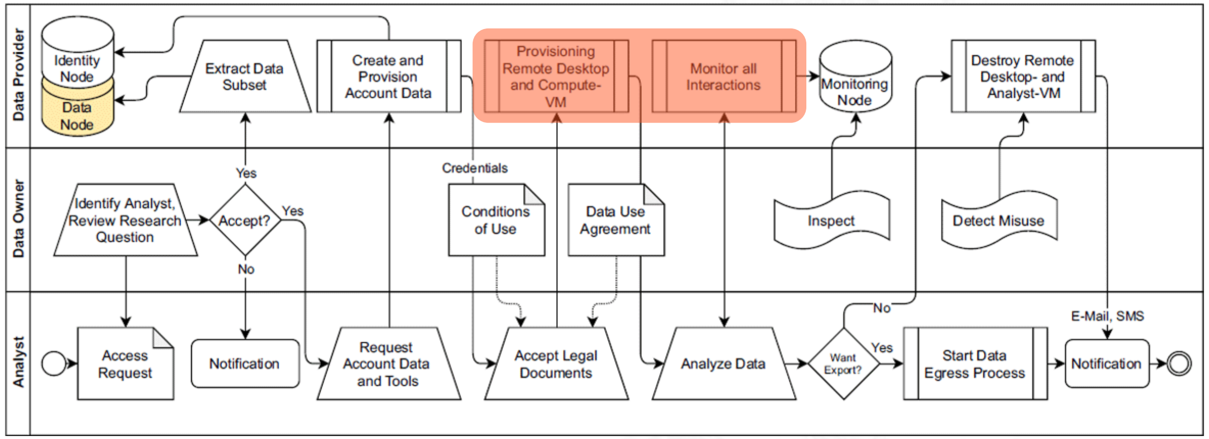
Weise, M., et al, 2022. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting, in *Data Science Journal*, 21(1). DOI: 10.5334/dsj-2022-004

Martin Weise, Technical University of Vienna

4.3. OSSDIP

(highlight)

1. Researcher sends **request** to Data Owner (*Person, question, required data*)
2. 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. Expose metadata of data subsets (FAIRness)
4. ...
5. Provisioning of VNC and Compute VMs with dedicated software and data

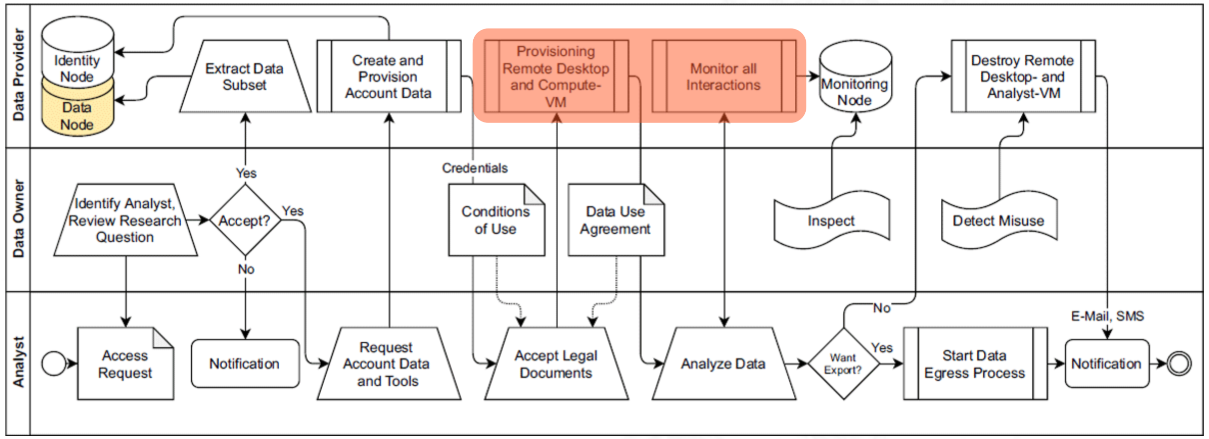


Weise, M., et al, 2022. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting, in *Data Science Journal*, 21(1). DOI: 10.5334/dsj-2022-004

4.3. OSSDIP

(highlight)

1. Researcher sends **request** to Data Owner (*Person, question, required data*)
2. 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. Expose metadata of data subsets (FAIRness)
4. …
5. Provisioning of VNC and Compute VMs with dedicated software and data



Weise, M., et al, 2022. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting, in *Data Science Journal*, 21(1). DOI: 10.5334/dsj-2022-004

4.3. OSSDIP Further Reading

Material

- https://doi.org/10.5334/dsj-2022-004 (journal paper)
- https://doi.org/10.34726/hss.2022.84700 (master's thesis)
- https://doi.org/10.17605/OSF.IO/VKN4R (iPRES'21 paper)

Resources

- https://ossdip.at/ (documentation, getting started guide)
- https://gitlab.tuwien.ac.at/martin.weise/ossdip (source code)

5. Future Work

DBRepo

- Prepare for **test** phase Q1 2023
- Prepare for **rollout** phase Q1 2024
- Document all endpoints, methods, files, readme, changelog, etc.
- Implement OAI-PMH interface for metadata harvesting

OSSDIP

- Find (friendly) test-users that want to deploy it within their premises using **synthetic** data
- Implement DBRepo features to make sensitive data **findable** and **reusable** (FAIR)
- **Transform** into Virtual Research Environment (e.g. integrating *Jupyterhub, Collaborative chat*)
- HPC

6. Conclusion

Different Repositories

- Must support a **wide range** of research objects
- Provide **good visibility** to them
- Will be a place where research happens

Trusted Research Environments

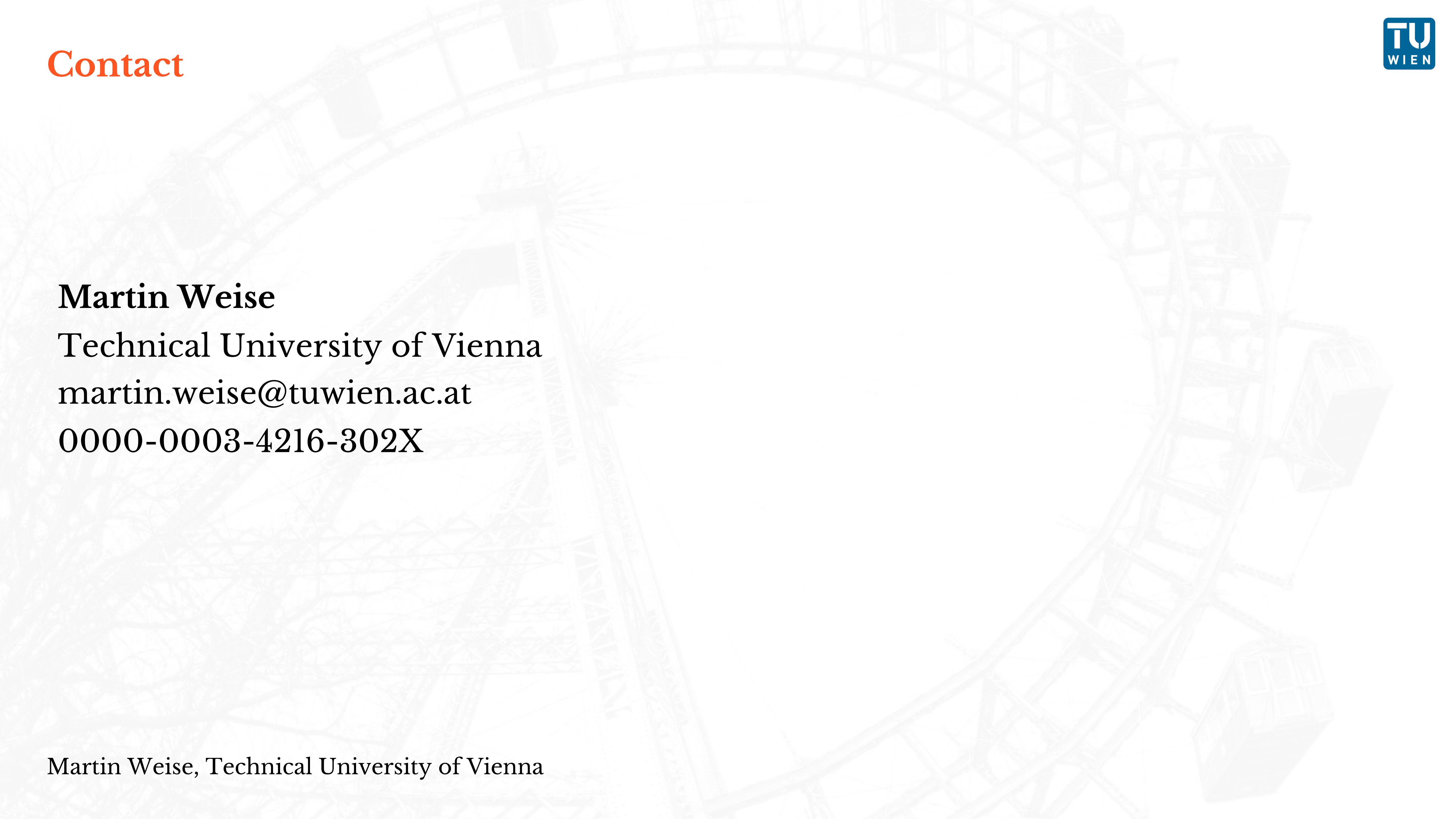
- Traditional privacy methods **not sufficient** for exploratory research
- **Virtual** meeting points to work with sensitive data and known tools
- Organizational / Technical / Legal

Composition of systems needed

- Gradual development with increasing complexity
- Integration and automation are key to facilitate adoption

Repositories and TREs are as **trustworthy** as their institutions

Contact



Martin Weise
Technical University of Vienna
martin.weise@tuwien.ac.at
0000-0003-4216-302X