3rd Conference of the ICTM
National Committee
for Portugal

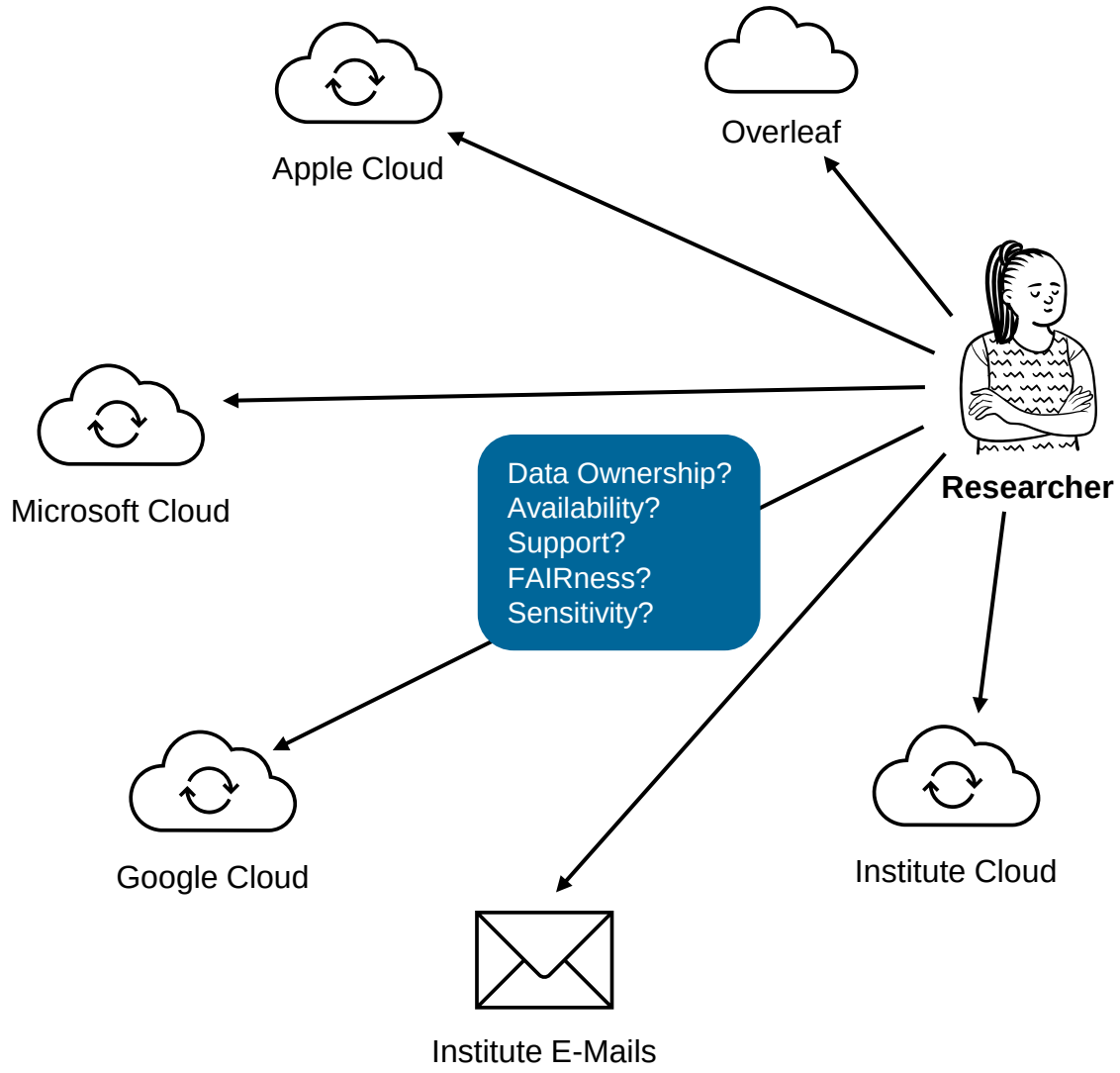
Connecting Ethnomusicology Data Collections Using Distributed Repositories and Linked Data Technology

DI Martin Weise[1]
Peter Knees[1], PhD
Alex Hofmann[2], PhD
Ardian Ahmedaja[2], PhD
Prof. Anda Beitāne[3], PhD
Prof. Andreas Rauber[1], PhD

[1]TU Wien *[2]University of Music and Performing Arts Vienna* *[3]Latvian Academy of Music*

Overview

Common Data Management



TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

5. HANDLING RESEARCH DATA

Research data should from the beginning be stored and maintained in appropriate systems and made available for use in a suitable repository (see 6.1. b). Research data must be provided with persistent identifiers[5] within the repository.

It is important to preserve the integrity of research data and to comply with the FAIR principles[6]. Research data must be stored in a correct, complete, unadulterated and reliable manner. They must be findable, identifiable, accessible, traceable, interoperable and whenever possible reusable and replicable.

In compliance with intellectual property rights, and unless third-party rights, legal requirements, Rectorate decisions, other reasonable interests or property laws prohibit it, research data should be assigned an open use license.[7]

Citation norms and requirements regarding publication and future research should be followed; data sources should be explicitly traceable in order for the original sources to be acknowledged.

Research data and records are to be stored and made available in accordance with intellectual property laws or the requirements of third-party funders as well as applicable legal or contractual requirements (e.g. EU restrictions on where identifiable personal data may be stored). Research data that may be of future historical interest and the records accompanying them should also be archived.

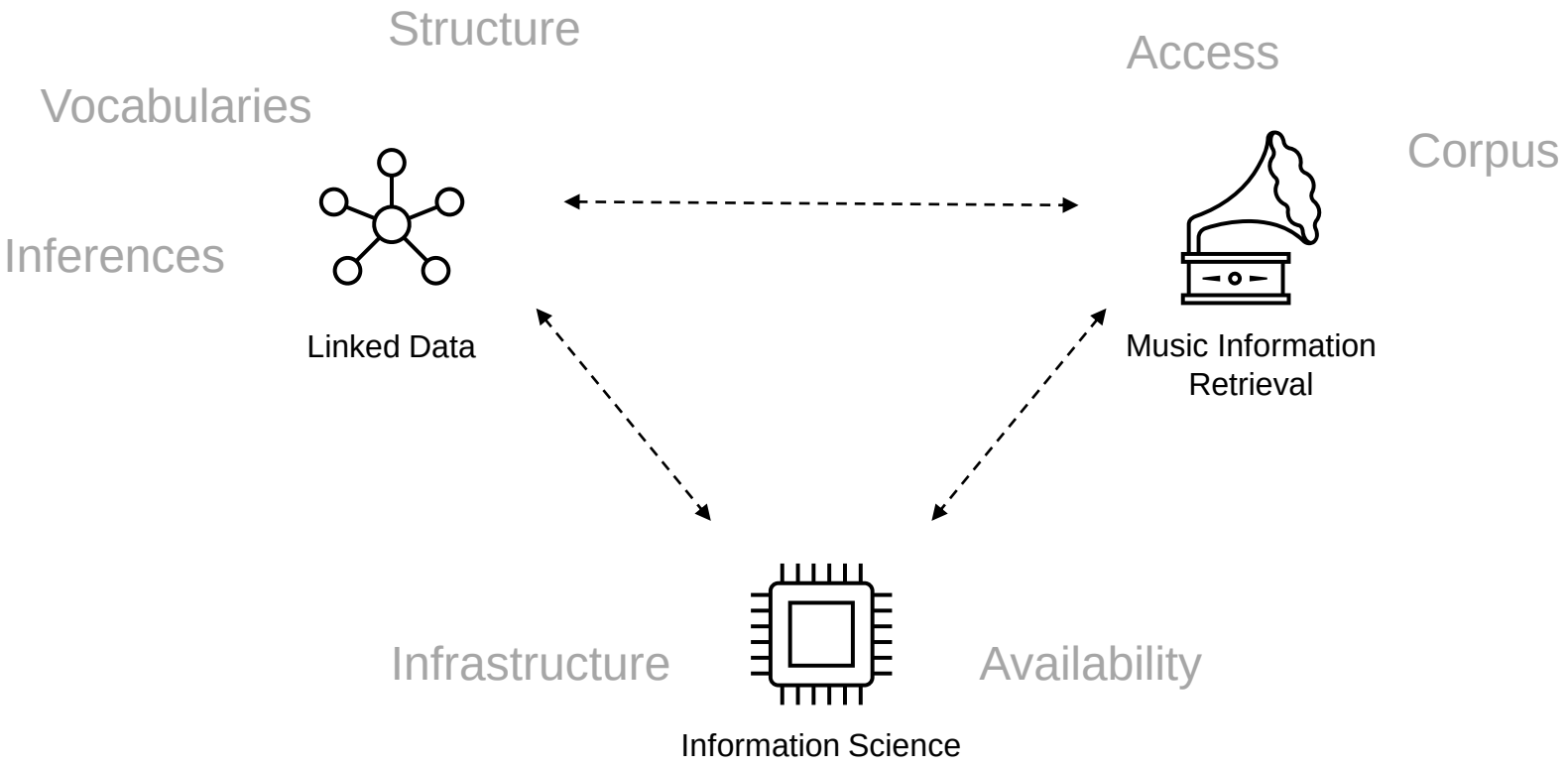
The minimum retention period for research data and records is 10 years after either the assignment of a persistent identifier or the publication of a related work following research completion, whichever is later.

In the event that research data and records are to be deleted or destroyed, either after expiration of the required retention period or for legal or ethical reasons, such action is to be carried out only after consideration of all legal and ethical perspectives. The following aspects must be taken into consideration when decisions are made about the retention or destruction of research data: interests and contractual provisions of third-party funders and other stakeholders, employees and partner participants in particular, as well as confidentiality and security. Any decision taken must be documented.

Research Data Policy 2018

Proposed Solution

Combine solutions from different disciplines



Hofmann, A., et al. (2021). Enabling FAIR Use of Ethnomusicology Data - Through Distributed Repositories, Linked Data and Music Information Retrieval. *Empirical Musicology Review*, 16(1), p.47-64. DOI: https://doi.org/10.18061/emr.v16i1.7632

FAIR Musicology Data

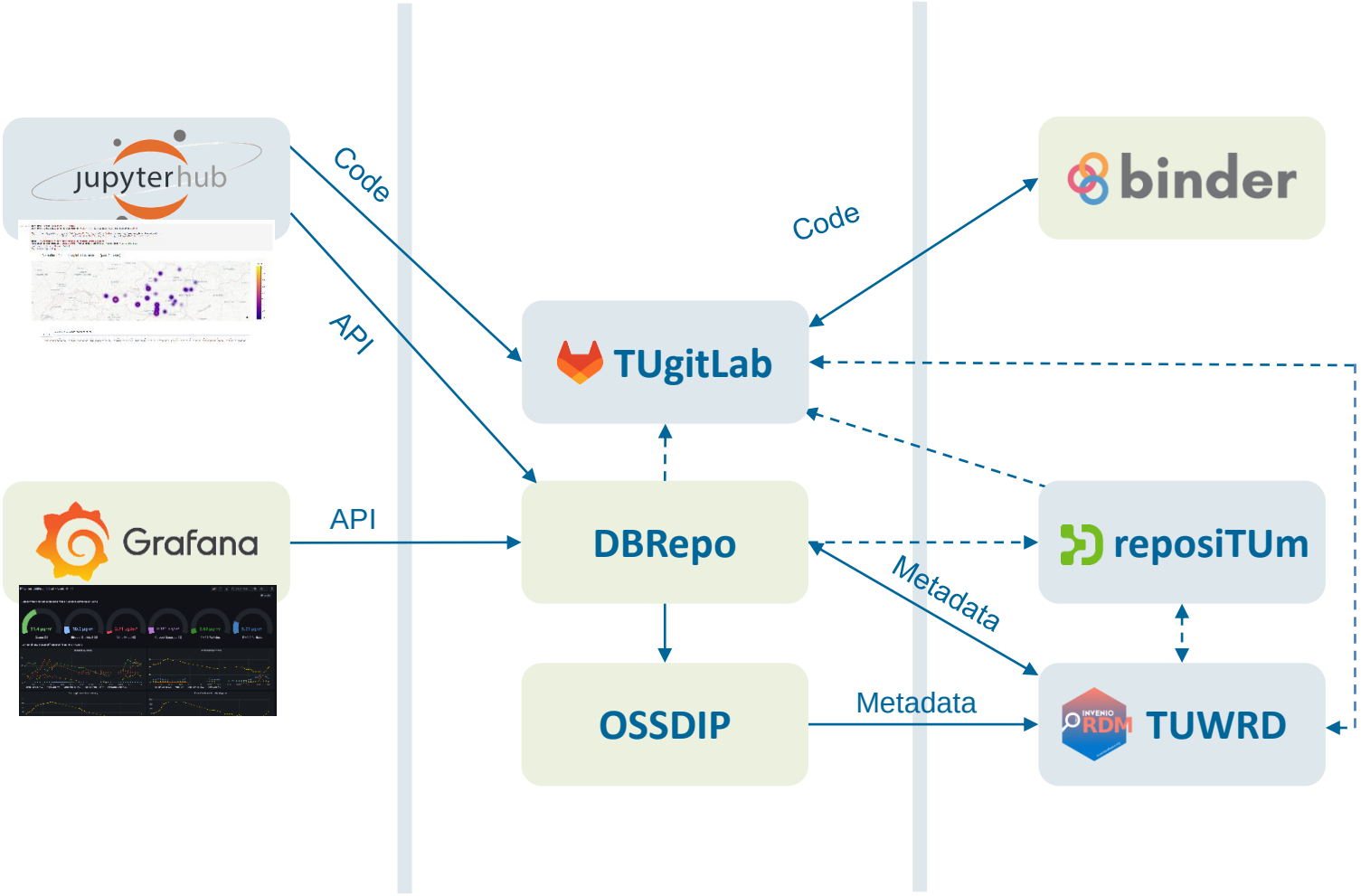
Requirements to musicology data FAIR

- R1: Secure storage and easy management of gathered research data.
- R2: Controlled data access and sharing with collaborators and contributors. Clarity on the data rights for sharing and reuse.
- R3: Importing existing collections
- R4: Description of data using a standardized vocabulary, to search across distributed data collections
- R5: Automatic (audio) data analysis for metadata generation

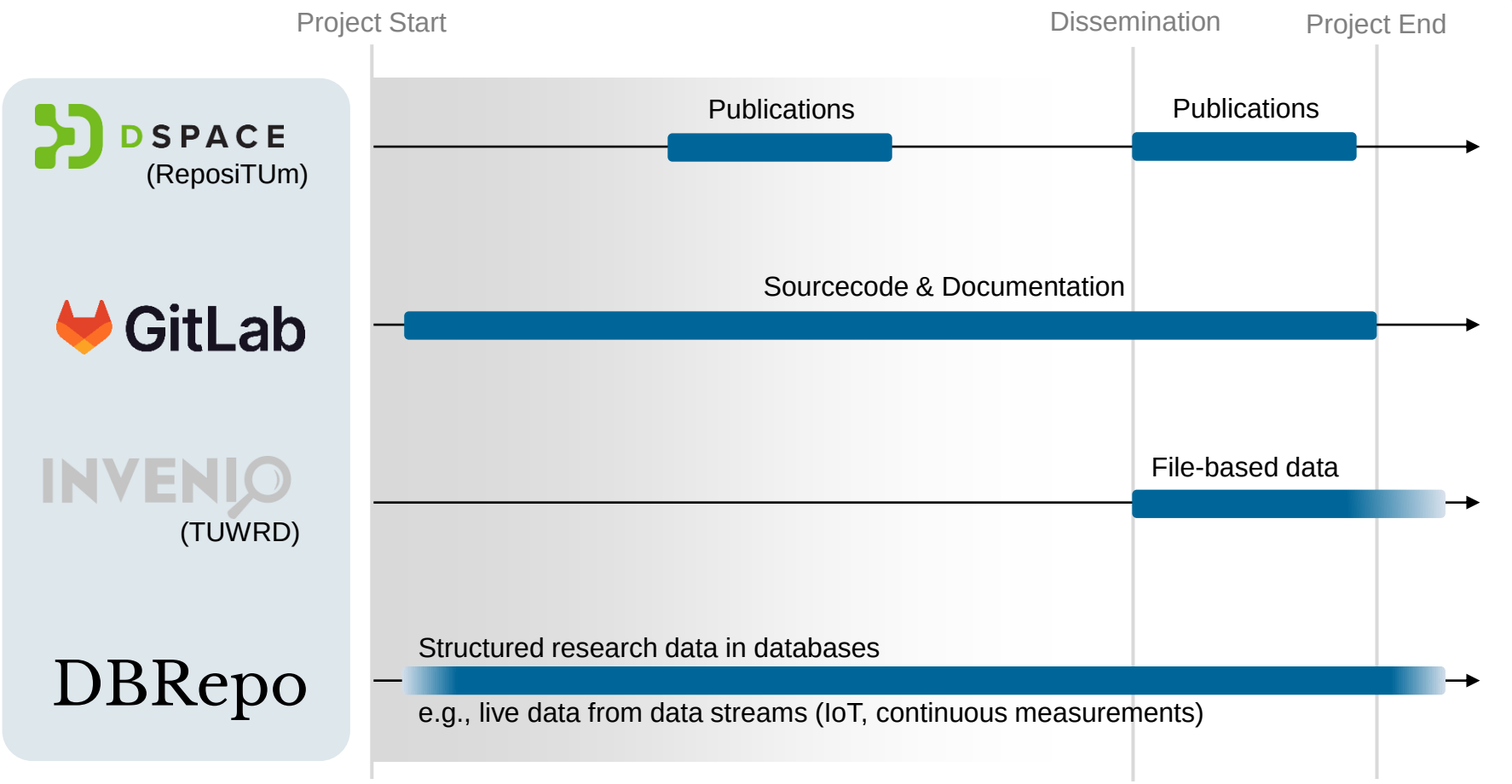
Hofmann, A., et al. (2021). Enabling FAIR Use of Ethnomusicology Data - Through Distributed Repositories, Linked Data and Music Information Retrieval. *Empirical Musicology Review*, 16(1), p.47-64. DOI: https://doi.org/10.18061/emr.v16i1.7632

Repository Infrastructure

Repository Infrastructure



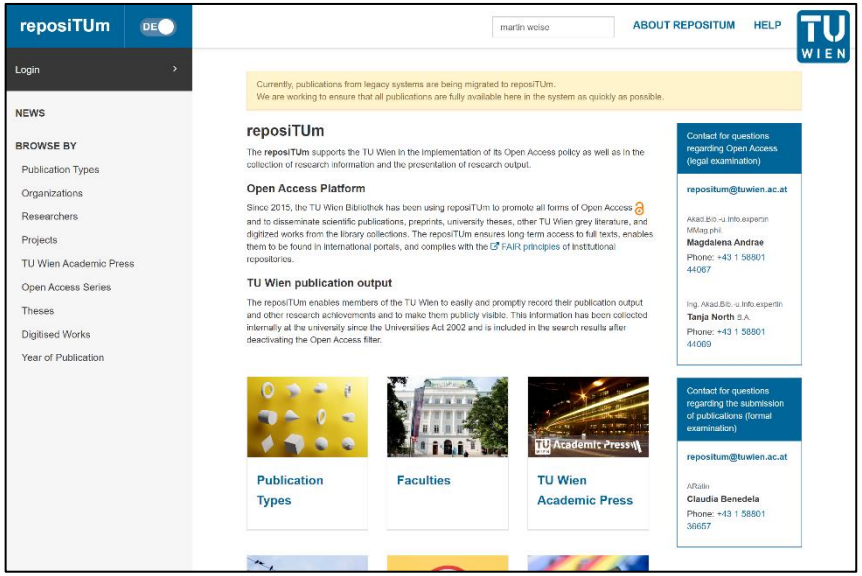
RDM Architecture at TU Wien



reposiTUm (publications)

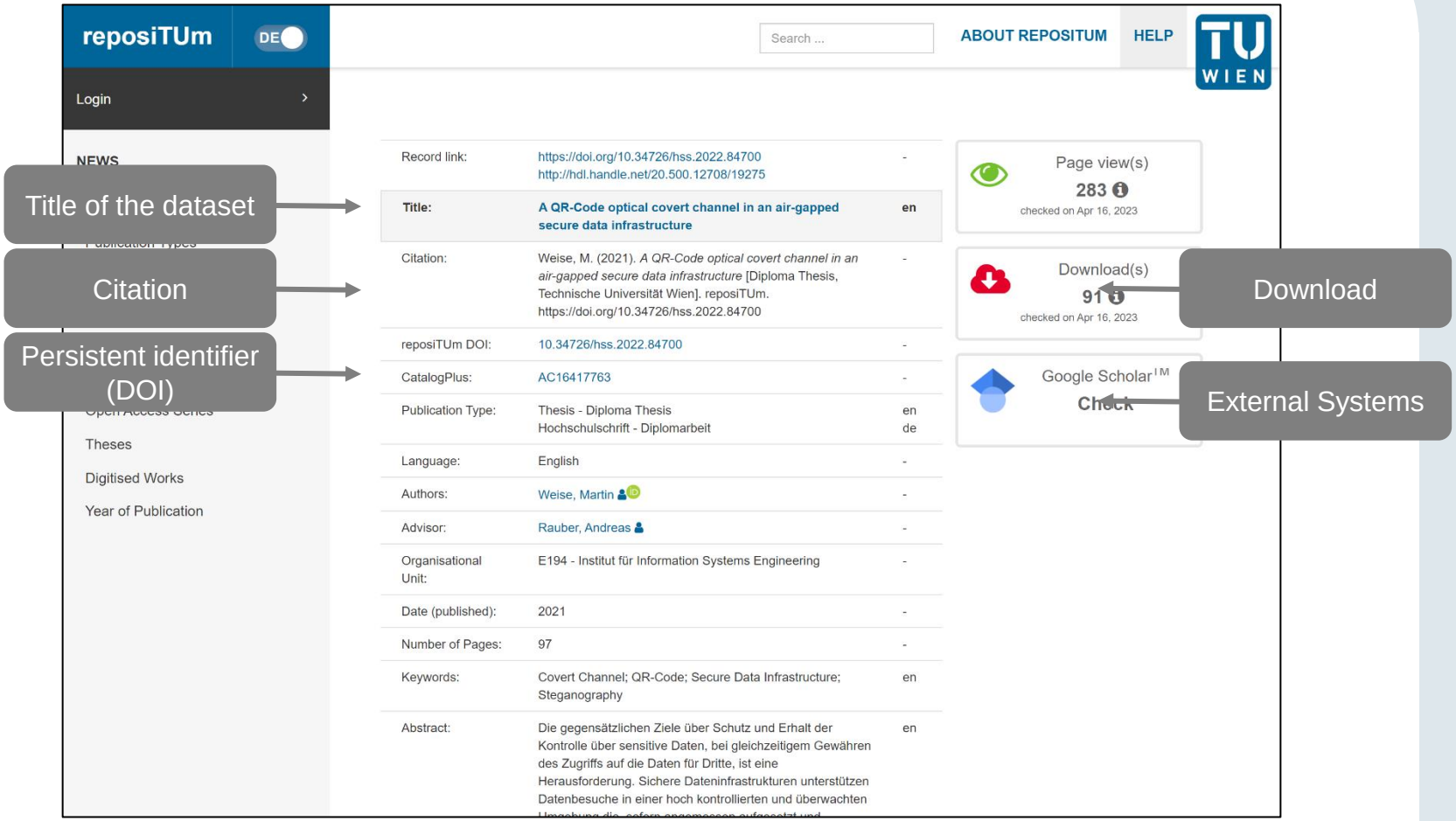
TU Wien Publication Repository

- Document-based research outputs
- Preservation
- Intellectual property
  - University ranking
  - Performance agreements
- Findability, Reusability
  - Papers
  - Presentations
  - Posters
  - Thesis’



https://repositum.tuwien.at/

reposiTUm (publications)

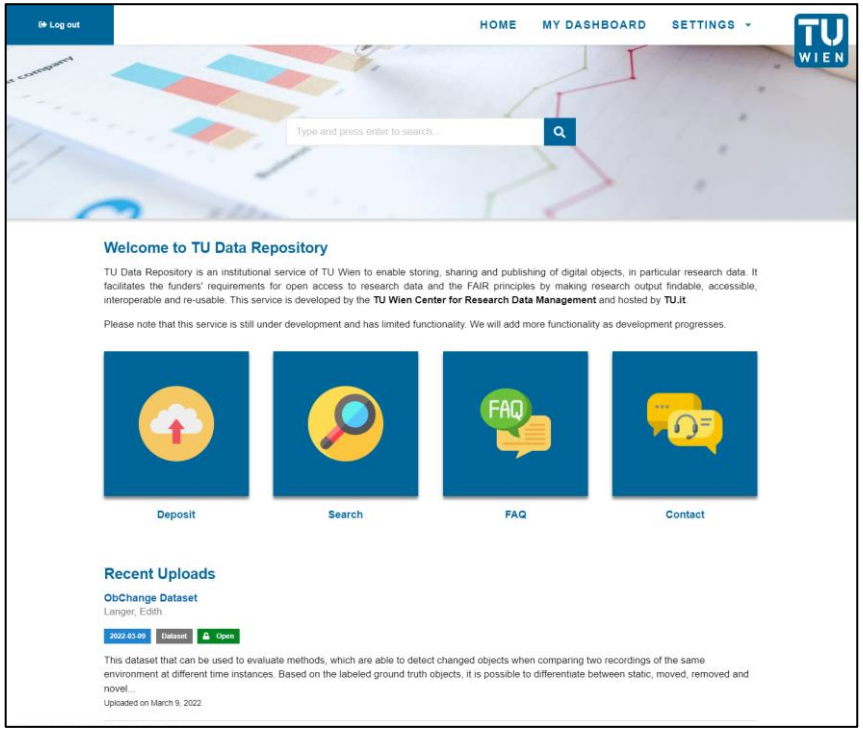


https://doi.org/10.34726/hss.2022.84700

TUWRD (data sets)

TU Wien Research Data Repository

- File-based research data
- Individual, collections
- Extensive metadata
  - DOIs
- Not for publications
  - Other system exists
- Operational since 2022
- CEPH storage, backups
- 66 datasets
- 9 TiB used currently



https://researchdata.tuwien.ac.at

TUWRD (data sets)

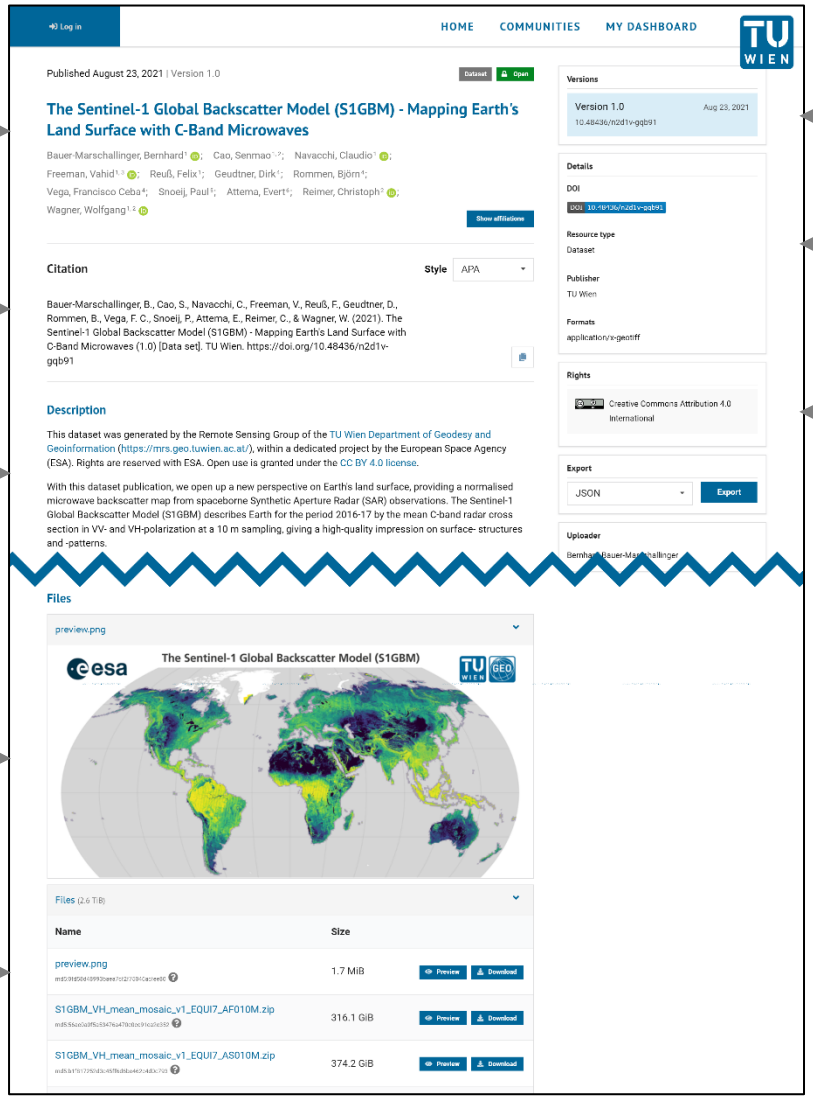
Title of the dataset

Citation

Description of the dataset

Preview file

Files for download



Version of the dataset

Persistent identifier (DOI)

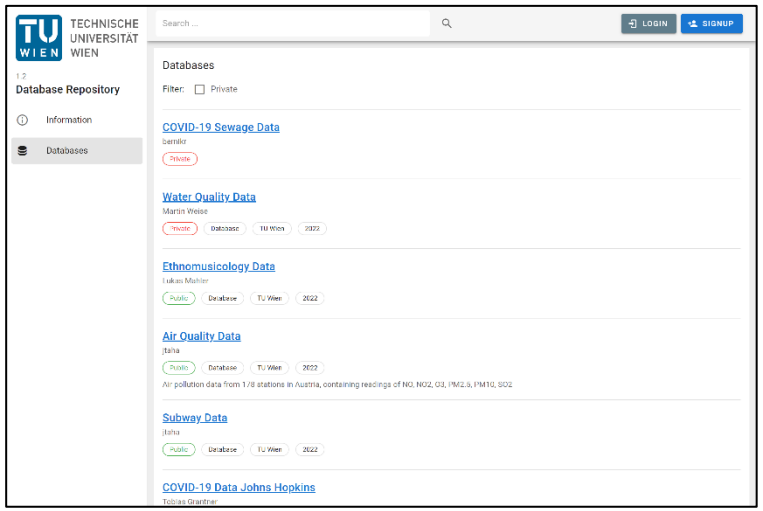
License CC-BY-NC-SA-3.0

https://doi.org/10.48436/n2d1v-gqb91

DBRepo (databases)

TU Wien Research Data Repository

- Handled ab-inition, no ex-post submission after project (no dumps)
- Handling **live data from data streams** (IoT, continuous measurements, …)
- Upload/download, **continuous feeding**, permissions, ownership
- Updates for corrections and versioning for reproducibility
- Web interface & **APIs for machine access**



https://dbrepo1.ec.tuwien.ac.at/

Supporting FAIR principles

Supporting RDA WGDC principles on data citation

Weise, M., Staudinger, M., Michlits, C., Gergely, E., Stytsenko, K., Ganguly, R., & Rauber, A. (2022). DBRepo: a Semantic Digital Repository for Relational Databases. *International Journal of Digital Curation*, 17(1), 11. https://doi.org/10.2218/ijdc.v17i1.825

DBRepo (databases)

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

http://purl.org/ontology/mo/Genre

Column Name	Type	Date Format	Concept	Unit	Primary Key	Unique	Nullable	Sequence
id	Number		ASSIGN	ASSIGN	• true	• true	false	false
start	Number		TIME	SECOND (TIME)	false	false	• true	false
stop	Number		TIME	SECOND (TIME)	false	false	• true	false
genre	Number		GENRE					
accuracy	Floating Number		ACCURACY					

https://www.wikidata.org/entity/Q54988221

Assign Semantic Information

We recommend the following ontologies

om2: http://www.ontology-of-units-of-measure.org/resource/om-2/
wd: https://www.wikidata.org/
mo: http://purl.org/ontology/mo/
dc: http://purl.org/dc/elements/1.1/
xsd: http://www.w3.org/2001/XMLSchema#
tl: http://purl.org/NET/c4dm/timeline.owl#
foaf: http://xmlns.com/foaf/0.1/
db: http://dbpedia.org

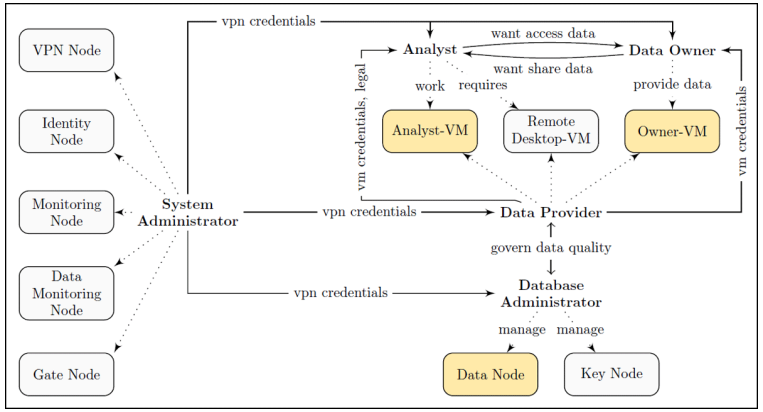
http://www.wikidata.org/entity/Q468777

CANCEL SAVE

OSSDIP (sensitive data)

Secure analysis environment

- TRE **model** and **reference implementation**
- Based on best-practice & open-source software
- Sensitive data (privacy issues, commercial interest), provide access for analysis, but ensure data is **not leaked** or misused
- Standard processes for involved roles



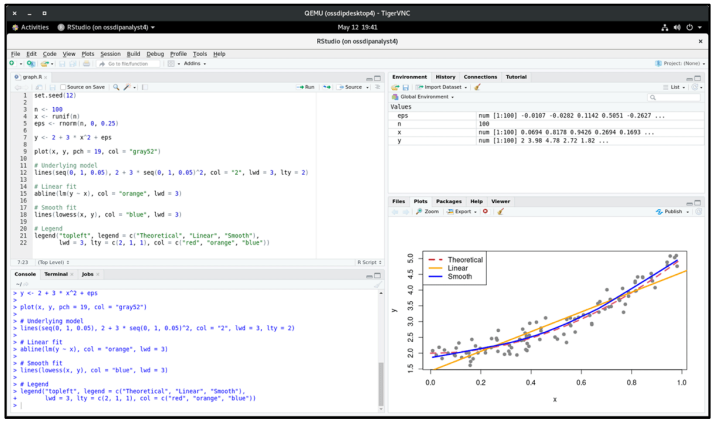
https://ossdip.at/

Weise, M., Kovacevic, F., Popper, N., & Rauber, A. (2022). OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. Data Science Journal, 21. https://doi.org/10.5334/dsj-2022-004

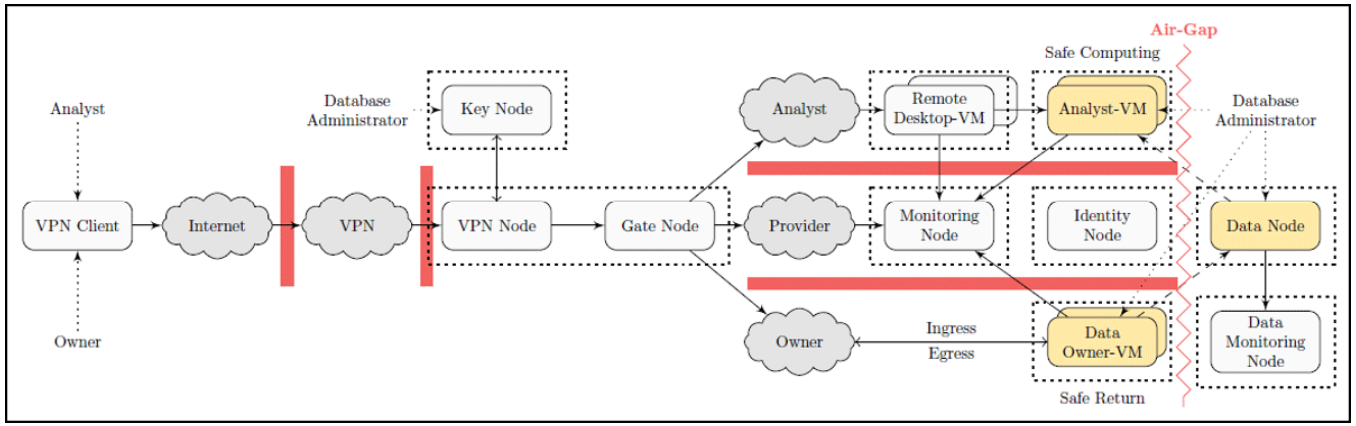
OSSDIP (sensitive data)

Data protection

- Air-gapped Data Node
- Only **brief** connections by trusted database admin
- Copy (fingerprinted, …) subset dataset from access request
- Analysis only via **multiple secure layers** & media breaks



https://ossdip.at/



Weise, M., Kovacevic, F., Popper, N., & Rauber, A. (2022). OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. Data Science Journal, 21. https://doi.org/10.5334/dsj-2022-004

Work Summary

Since 2019

- Started to operate **three repositories**
  - TUWRD for data sets
  - reposiTUm for publications
  - TUgitLab for code
- Started development of a **new repository** as none existed before
  - DBRepo for databases
- Started development of a secure data infrastructure
  - OSSDIP, blueprint and technical reference implementation

Musicology Use-Case

Musicology Use-Case

Emotify Dataset on Induced Musical Emotion

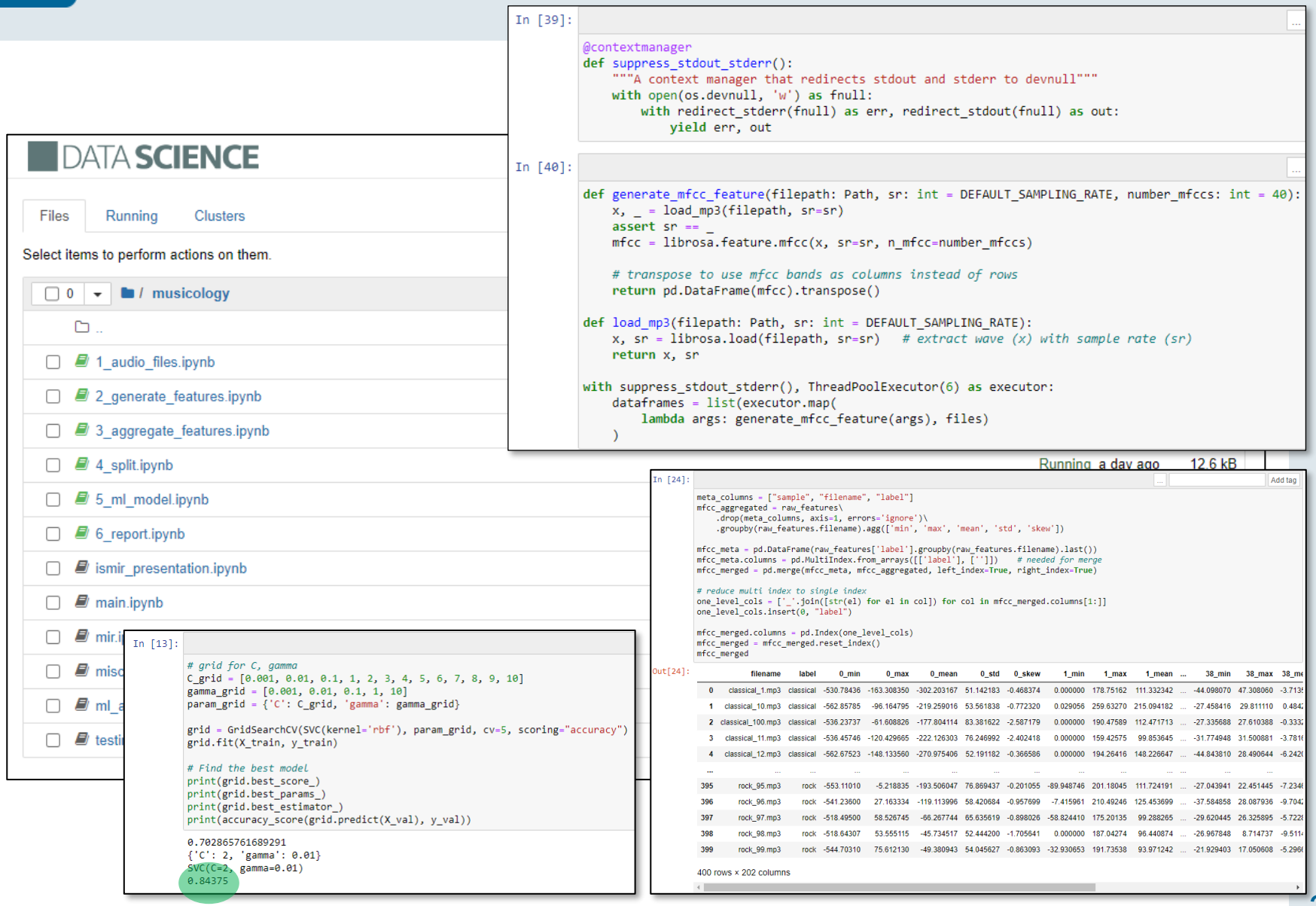
- **400 song excerpts** (each 1 minute long) in **4 genres** (rock, classical, pop, electronic)
- Annotated with max. 3 items from the **GEMS** scale

Classification

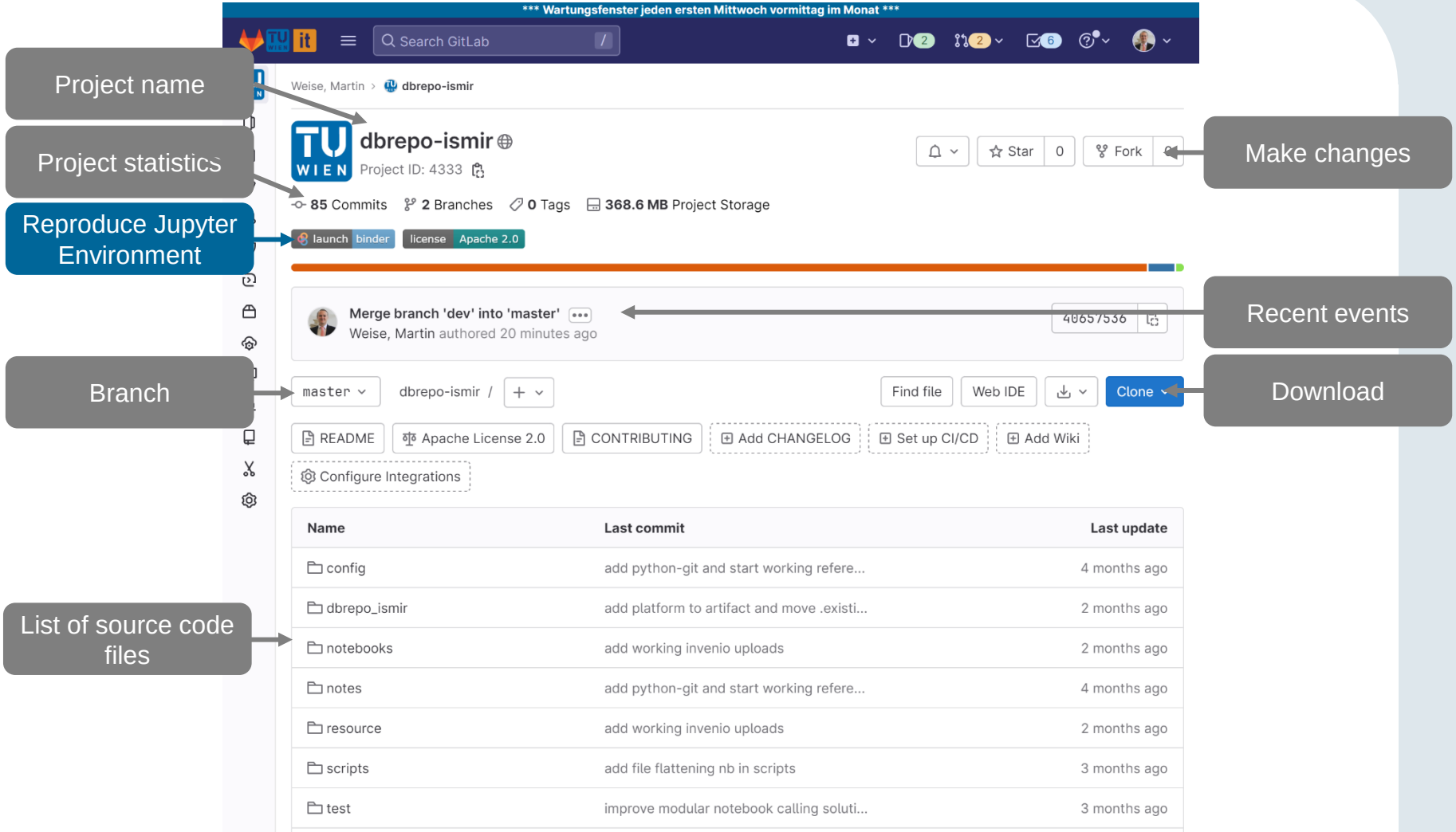
- Machine-learning task for Bachelor-thesis
- Generate 40 MFCC features per song excerpt
- Reduce dimensions with PCA
- Fit SVM
- Predict Genre from MFCCs

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying Emotion Induced by Music Through a Crowdsourcing Game. *Information Processing & Management*, 52(1), p.115-128. https://doi.org/10.1016/j.ipm.2015.03.004

Analysis Environment



Code Repository



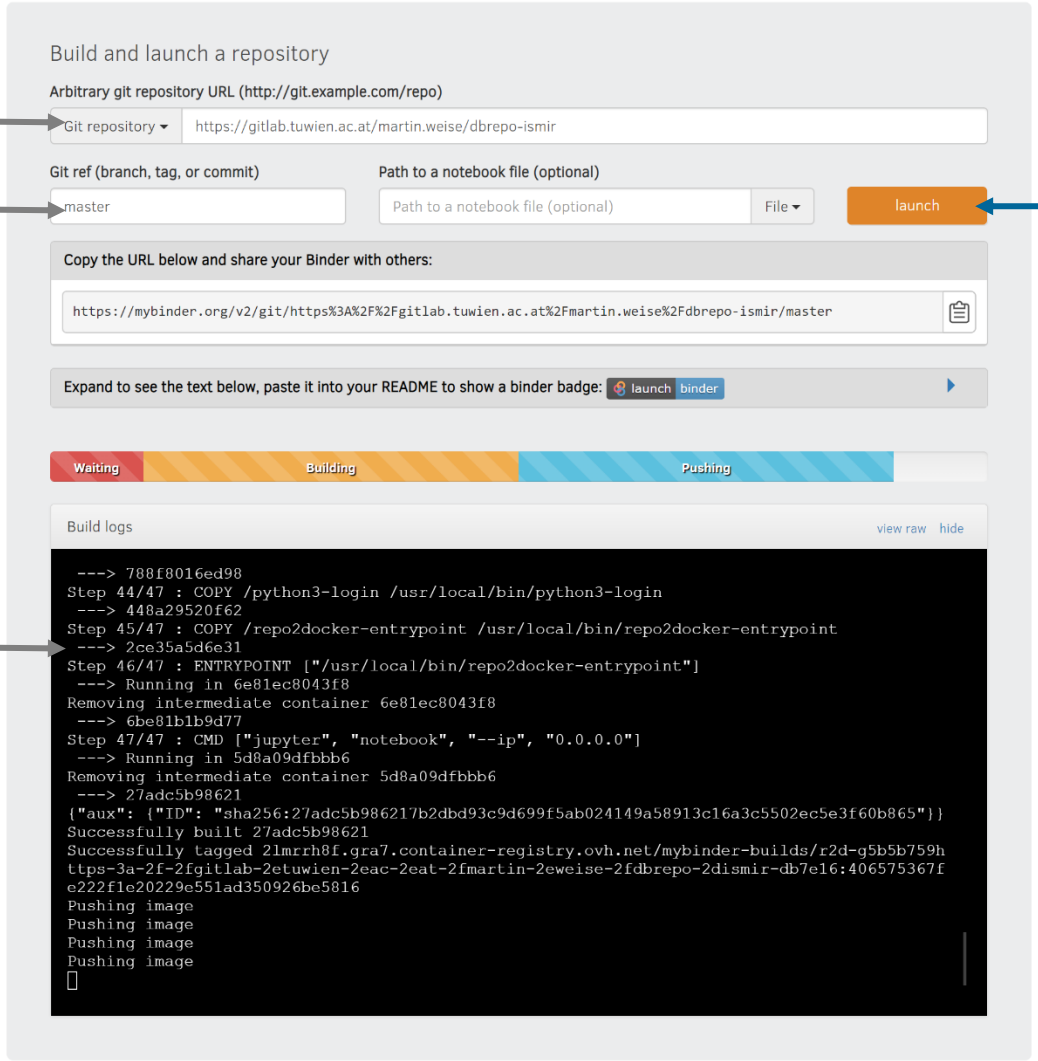
Mahler, L. (2023). DBRepo ISMIR. [Online] URL: https://gitlab.tuwien.ac.at/martin.weise/dbrepo-ismir, accessed 2023-05-21

Reproducing Research Results

Link to Git repository

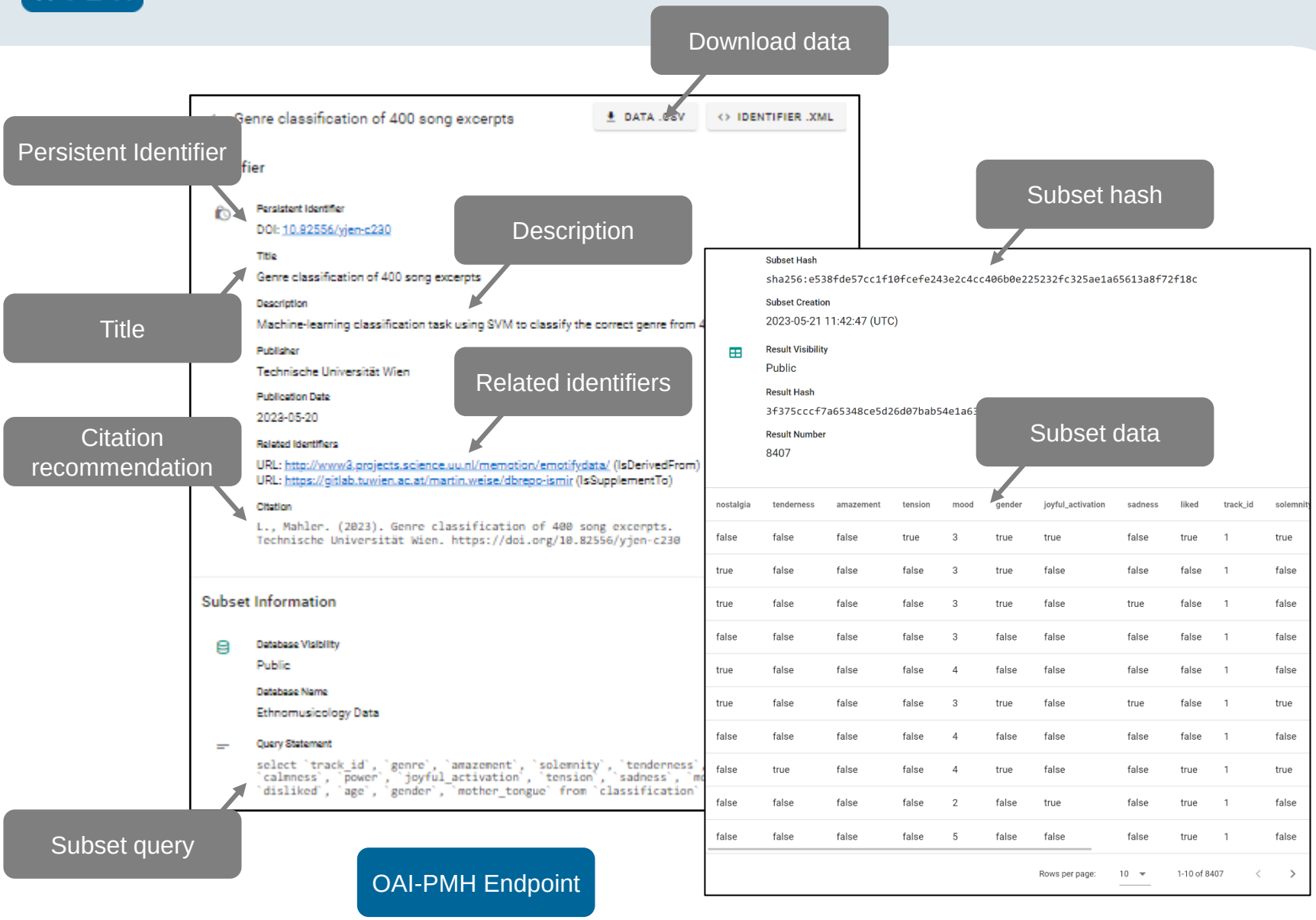
Branch name or Commit hash

Action log



Launch Jupyter notebook

Deposit structured data from start



Conclusions & Future Work

Future Work

- Proposed **operational** repositories and services at TU Wien
- Proposed two repositories that are in **development**
- Showed how **musicology data can be linked** using PIDs and controlled vocabulary
- Showed **reproducibility** of research results

Future Work

- Suggesting of semantic concepts based on table schema
- Suggesting of semantic concepts based on table contents

DI Martin Weise

E194-04 Data Science
Favoritenstraße 9-11
A-1040 Wien

martin.weise@tuwien.ac.at
0000-0003-4216-302X