



3rd Conference of the ICTM
National Committee
for Portugal

Connecting Ethnomusicology Data Collections Using Distributed Repositories and Linked Data Technology

DI Martin Weise¹

Peter Knees¹, PhD

Alex Hofmann², PhD

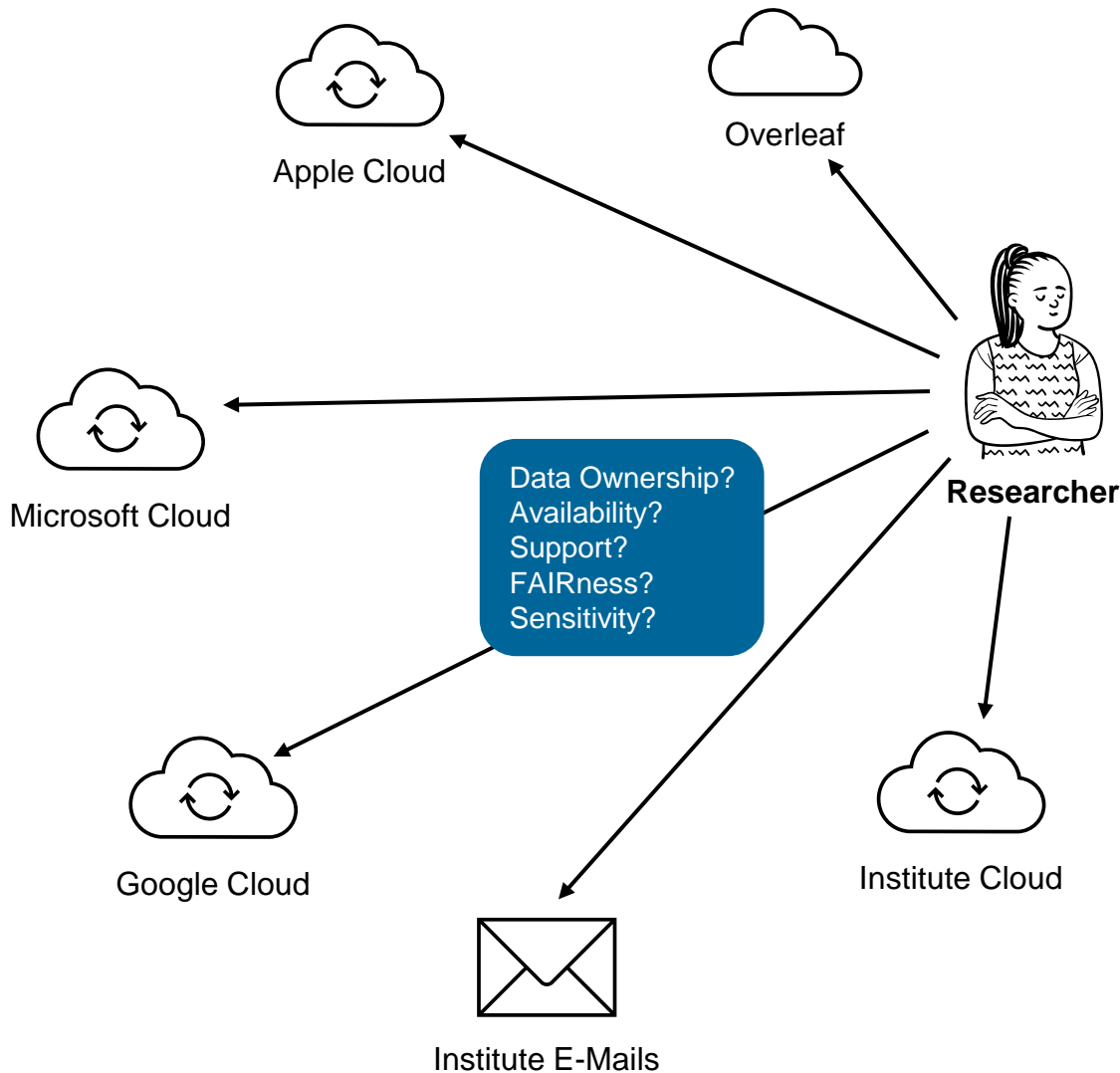
Ardian Ahmedaja², PhD

Prof. Anda Beitāne³, PhD

Prof. Andreas Rauber¹, PhD



Overview



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 identifiers⁴ within the repository.

It is important to preserve the integrity of research data and to comply with the FAIR principles⁶. 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.

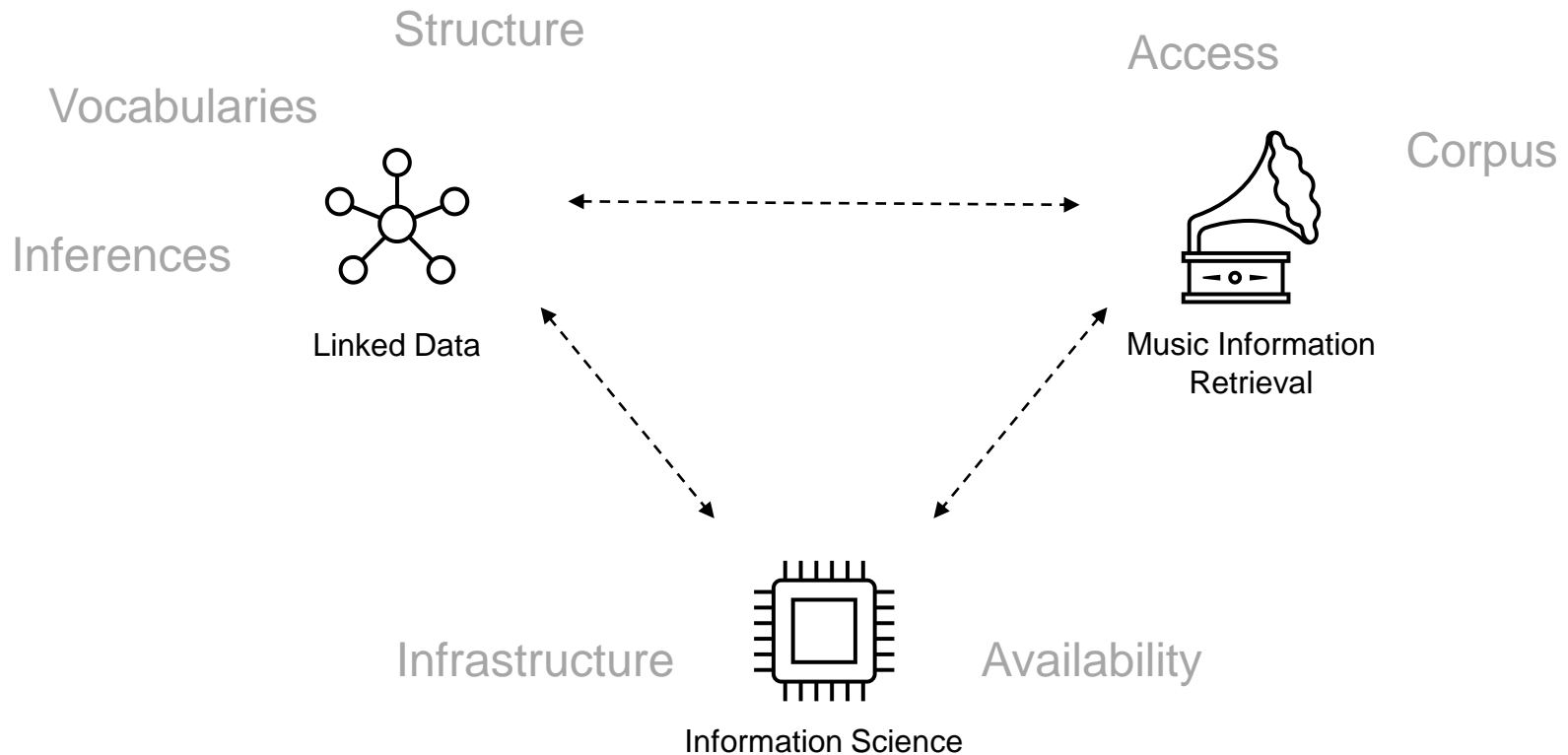
Research data and records are to be stored and made available in accordance with intellectual property laws or the requirements of third-party funders as well as applicable legal or contractual requirements (e.g. EU restrictions on where identifiable personal data may be stored). Research data that may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

Combine solutions from different disciplines



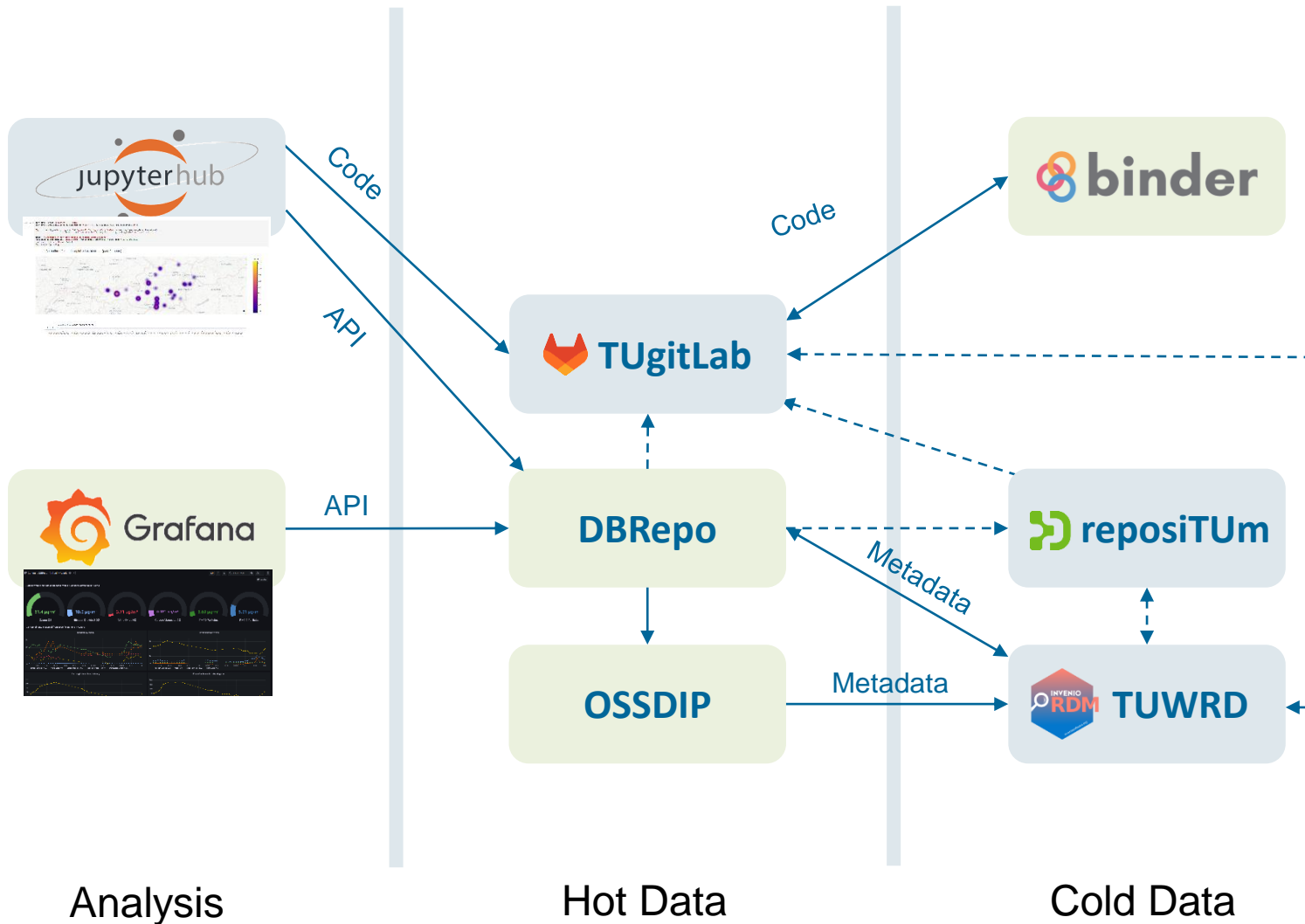
Requirements to musicology data FAIR

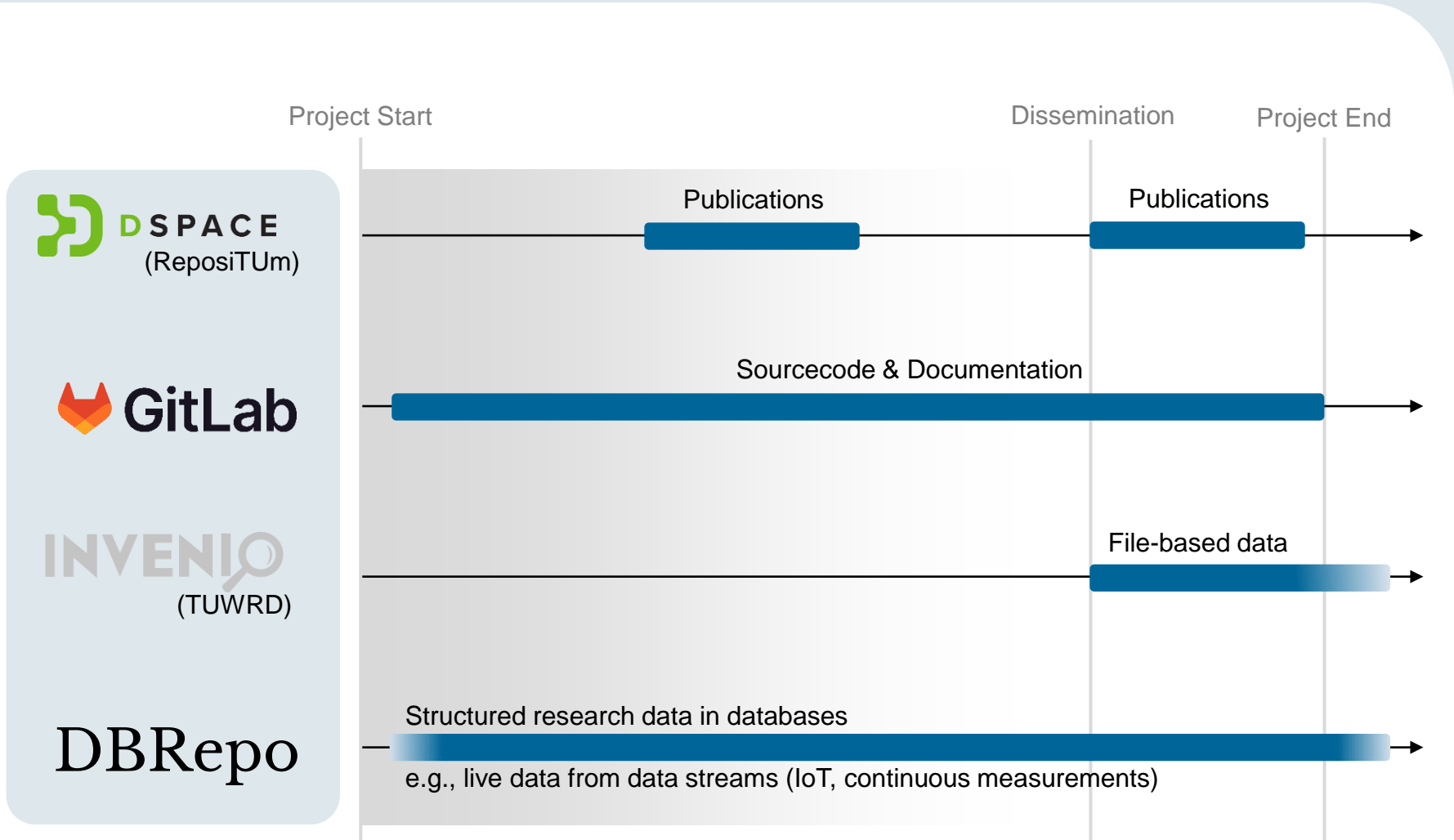
- R1: Secure storage and easy management of gathered research data.
- R2: Controlled data access and sharing with collaborators and contributors. Clarity on the data rights for sharing and reuse.
- R3: Importing existing collections
- R4: Description of data using a standardized vocabulary, to search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



Repository Infrastructure

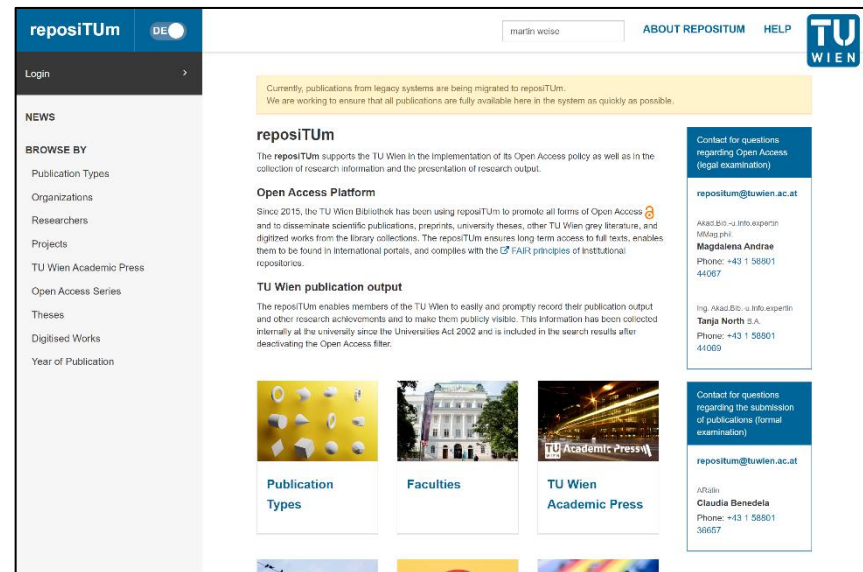
Repository Infrastructure





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 DE

Search ... ABOUT REPOSITUM HELP

Login >

NEWS

Publication Types

Open Access Series

Theses

Digitised Works

Year of Publication

Record link: <https://doi.org/10.34726/hss.2022.84700>
<http://hdl.handle.net/20.500.12708/19275>

Title: **A QR-Code optical covert channel in an air-gapped secure data infrastructure** en

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>

repositUm DOI: [10.34726/hss.2022.84700](https://doi.org/10.34726/hss.2022.84700)

CatalogPlus: [AC16417763](#)

Publication Type: Thesis - Diploma Thesis en
Hochschulschrift - Diplomarbeit de

Language: English

Authors: [Weise, Martin](#)

Advisor: [Rauber, Andreas](#)

Organisational Unit: E194 - Institut für Information Systems Engineering

Date (published): 2021

Number of Pages: 97

Keywords: Covert Channel; QR-Code; Secure Data Infrastructure; Steganography en

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 reformen erregene aufgesetzt und

Page view(s)
283
checked on Apr 16, 2023

Download(s)
91
checked on Apr 16, 2023

Google Scholar™
Check

Title of the dataset →

Citation →

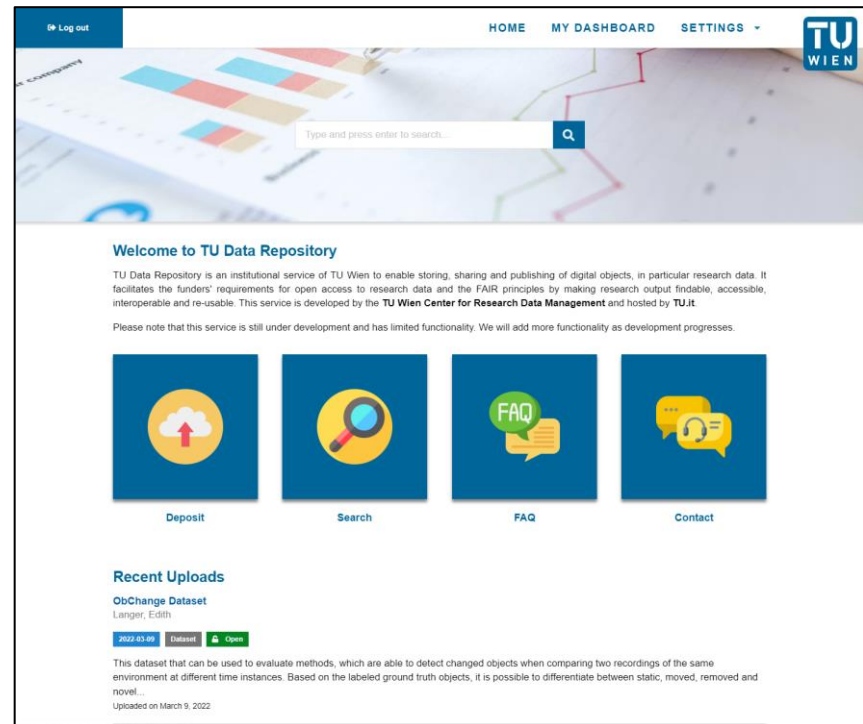
Persistent identifier (DOI) →

Download

External Systems

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>

Title of the dataset

Citation

Description of the dataset

Preview file

Files for download

The screenshot shows a dataset page on the TUWRD platform. The title is "The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves". The page includes a list of authors, a citation in APA style, a description of the dataset, and a section for files. The files section contains a table with columns for Name and Size, listing files like 'preview.png', 'S1GBM_VH_mean_mosaic_v1_EQU17_AF010M.zip', and 'S1GBM_VH_mean_mosaic_v1_EQU17_AS010M.zip'.

Version of the dataset

Persistent identifier (DOI)

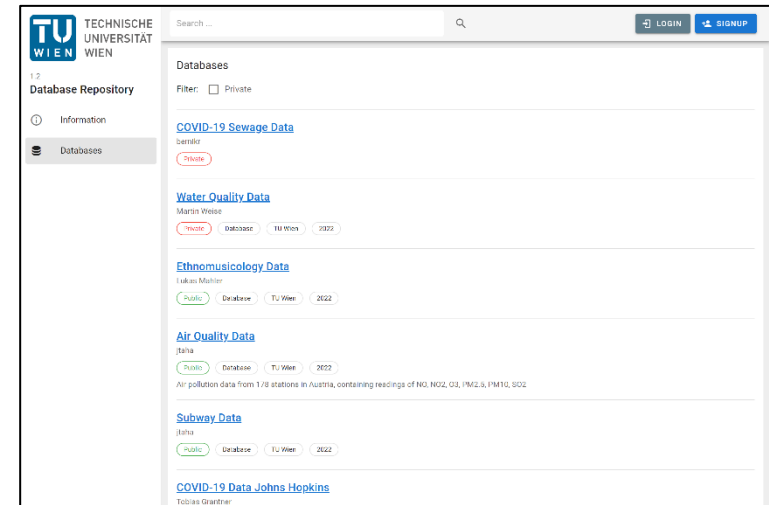
License
CC-BY-NC-SA-3.0

TU Wien Research Data Repository

- Handled ab-initio, 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**

Supporting FAIR principles

Supporting RDA WGDC principles on data citation



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

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					

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>

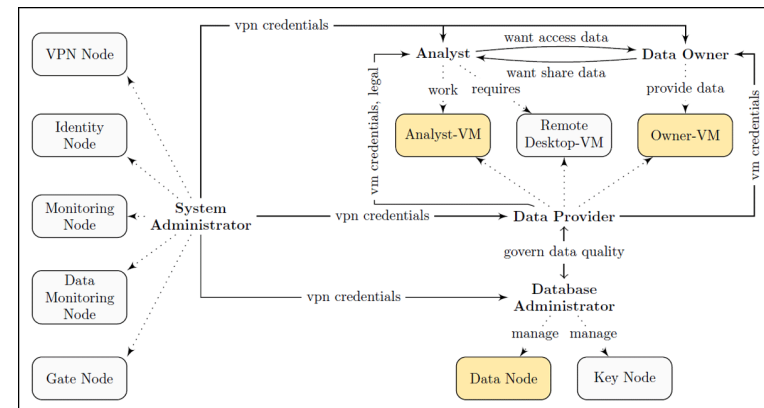
CANCEL SAVE

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



Secure analysis environment

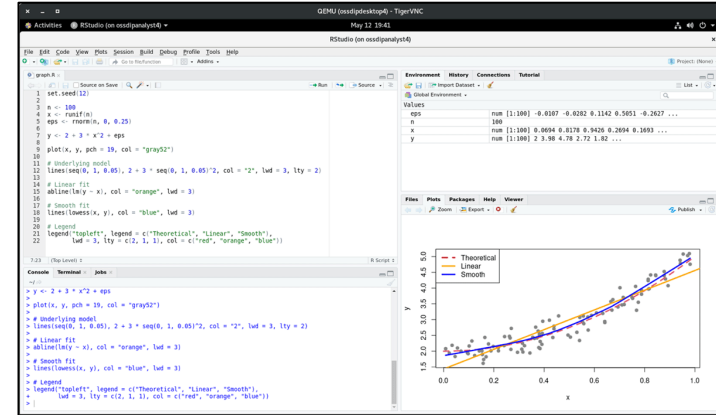
- TRE model and reference implementation
- Based on best-practice & open-source software
- Sensitive data (privacy issues, commercial interest), provide **access for analysis**, but ensure data is **not leaked** or misused
- Standard processes for involved roles



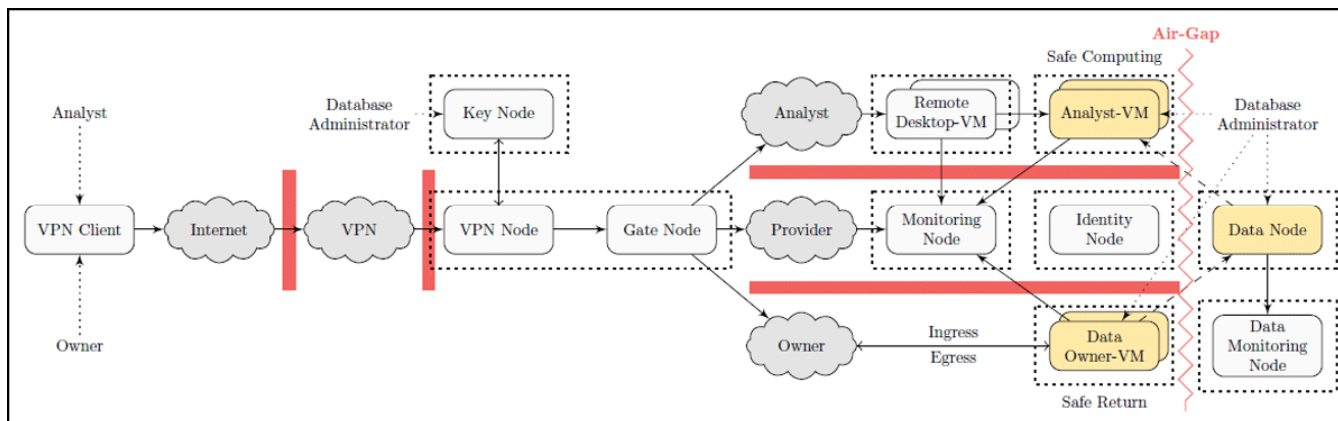
<https://ossdip.at/>

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/>



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

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

DATA SCIENCE

Files
Running
Clusters

Select items to perform actions on them.

0

/ musicology

..

1_audio_files.ipynb

2_generate_features.ipynb

3_aggregate_features.ipynb

4_split.ipynb

5_ml_model.ipynb

6_report.ipynb

ismir_presentation.ipynb

main.ipynb

miri

misc

ml_a

testi

```
In [13]:
# grid for C, gamma
C_grid = [0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
gamma_grid = [0.001, 0.01, 0.1, 1, 10]
param_grid = {'C': C_grid, 'gamma': gamma_grid}

grid = GridSearchCV(SVC(kernel='rbf'), param_grid, cv=5, scoring="accuracy")
grid.fit(X_train, y_train)

# Find the best model
print(grid.best_score_)
print(grid.best_params_)
print(grid.best_estimator_)
print(accuracy_score(grid.predict(X_val), y_val))

0.702865761689291
{'C': 2, 'gamma': 0.01}
SVC(C=2, gamma=0.01)
0.84375
```

```
In [39]:
@contextmanager
def suppress_stdout_stderr():
    """A context manager that redirects stdout and stderr to devnull"""
    with open(os.devnull, 'w') as fnull:
        with redirect_stderr(fnull) as err, redirect_stdout(fnull) as out:
            yield err, out
```

```
In [40]:
def generate_mfcc_feature(filepath: Path, sr: int = DEFAULT_SAMPLING_RATE, number_mfccs: int = 40):
    x, _ = load_mp3(filepath, sr=sr)
    assert sr == _
    mfcc = librosa.feature.mfcc(x, sr=sr, n_mfcc=number_mfccs)

    # transpose to use mfcc bands as columns instead of rows
    return pd.DataFrame(mfcc).transpose()

def load_mp3(filepath: Path, sr: int = DEFAULT_SAMPLING_RATE):
    x, sr = librosa.load(filepath, sr=sr) # extract wave (x) with sample rate (sr)
    return x, sr

with suppress_stdout_stderr(), ThreadPoolExecutor(6) as executor:
    dataframes = list(executor.map(
        lambda args: generate_mfcc_feature(args), files
    ))
```

Running a day ago 12.6 kB

```
In [24]:
meta_columns = ["sample", "filename", "label"]
mfcc_aggregated = raw_features\
    .drop(meta_columns, axis=1, errors='ignore')\
    .groupby(raw_features.filename).agg(['min', 'max', 'mean', 'std', 'skew'])

mfcc_meta = pd.DataFrame(raw_features['label']).groupby(raw_features.filename).last()
mfcc_meta.columns = pd.MultiIndex.from_arrays(['label', ['']]) # needed for merge
mfcc_merged = pd.merge(mfcc_meta, mfcc_aggregated, left_index=True, right_index=True)

# reduce multi index to single index
one_level_cols = ['.'.join([str(e1) for e1 in col]) for col in mfcc_merged.columns[1:]]
one_level_cols.insert(0, "label")

mfcc_merged.columns = pd.Index(one_level_cols)
mfcc_merged = mfcc_merged.reset_index()
mfcc_merged
```

	filename	label	0_min	0_max	0_mean	0_std	0_skew	1_min	1_max	1_mean	...	38_min	38_max	38_me
0	classical_1.mp3	classical	-530.78436	-163.308350	-302.203167	51.142183	-0.468374	0.000000	178.75162	111.332342	...	-44.098070	47.308060	-3.7135
1	classical_10.mp3	classical	-562.85785	-96.164795	-219.259016	53.561838	-0.772320	0.029956	259.63270	215.094182	...	-27.458416	29.811110	0.484
2	classical_100.mp3	classical	-536.23737	-61.608826	-177.804114	83.381622	-2.587179	0.000000	190.47589	112.471713	...	-27.335668	27.610388	-0.333
3	classical_11.mp3	classical	-536.45746	-120.429665	-222.126303	76.246992	-2.402418	0.000000	159.42575	99.853645	...	-31.774948	31.500881	-3.7811
4	classical_12.mp3	classical	-562.67523	-148.133560	-270.975406	52.191182	-0.366586	0.000000	194.26416	148.226647	...	-44.843810	28.490644	-6.2421
...
395	rock_95.mp3	rock	-553.11010	-5.218835	-193.506047	76.869437	-0.201055	-89.948746	201.18045	111.724191	...	-27.043941	22.451445	-7.234
396	rock_96.mp3	rock	-541.23600	27.163334	-119.113996	58.420684	-0.957699	-7.415961	210.49246	125.453699	...	-37.584858	28.087936	-9.704
397	rock_97.mp3	rock	-518.49500	58.526745	-66.267744	65.635619	-0.898026	-58.824410	175.20135	99.288265	...	-29.620445	26.325895	-5.7221
398	rock_98.mp3	rock	-518.64307	53.555115	-45.734517	52.444200	-1.705641	0.000000	187.04274	96.440874	...	-26.967848	8.714737	-9.5111
399	rock_99.mp3	rock	-544.70310	75.612130	-49.380943	54.045627	-0.863093	-32.930653	191.73538	93.971242	...	-21.929403	17.050608	-5.296

400 rows x 202 columns

The screenshot shows the GitLab interface for the 'dbrepo-ismir' repository. Callout boxes on the left and right point to specific features:

- Project name:** dbrepo-ismir
- Project statistics:** 85 Commits, 2 Branches, 0 Tags, 368.6 MB Project Storage
- Reproduce Jupyter Environment:** launch binder
- Branch:** master
- List of source code files:** A table listing files and their last commit details.
- Make changes:** Fork button
- Recent events:** Merge branch 'dev' into 'master'
- Download:** Clone button

Name	Last commit	Last update
config	add python-git and start working refere...	4 months ago
dbrepo_ismir	add platform to artifact and move .existi...	2 months ago
notebooks	add working invenio uploads	2 months ago
notes	add python-git and start working refere...	4 months ago
resource	add working invenio uploads	2 months ago
scripts	add file flattening nb in scripts	3 months ago
test	improve modular notebook calling soluti...	3 months ago

Reproducing Research Results

Link to Git repository

Branch name or Commit hash

Launch Jupyter notebook

Action log

Build and launch a repository

Arbitrary git repository URL (<http://git.example.com/repo>)

Git repository

Git ref (branch, tag, or commit) Path to a notebook file (optional)

Copy the URL below and share your Binder with others:

Expand to see the text below, paste it into your README to show a binder badge:

Waiting Building Pushing

Build logs [view raw](#) [hide](#)

```

---> 788f8016ed98
Step 44/47 : COPY /python3-login /usr/local/bin/python3-login
---> 448a29520f62
Step 45/47 : COPY /repo2docker-entrypoint /usr/local/bin/repo2docker-entrypoint
---> 2ce35a5d6e31
Step 46/47 : ENTRYPOINT ["/usr/local/bin/repo2docker-entrypoint"]
---> Running in 6e81ec8043f8
Removing intermediate container 6e81ec8043f8
---> 6be81b1b9d77
Step 47/47 : CMD ["jupyter", "notebook", "--ip", "0.0.0.0"]
---> Running in 5d8a09dfbbb6
Removing intermediate container 5d8a09dfbbb6
---> 27adc5b98621
{"aux": {"ID": "sha256:27adc5b986217b2bd93c9d699f5ab024149a58913c16a3c5502ec5e3f60b865"}}
Successfully built 27adc5b98621
Successfully tagged 2lmrrh8f.gra7.container-registry.ovh.net/mybinder-bullds/r2d-g5b5b759h
ttps-3a-2f-2fgitlab-2etuwien-2eat-2fmartin-2eweise-2fdbrepo-2dismir-db7e16:406575367f
e222fle20229e551ad350926be5816
Pushing image
Pushing image
Pushing image
Pushing image

```

Deposit structured data from start

Download data

Persistent Identifier

Genre classification of 400 song excerpts

DATA .CSV IDENTIFIER .XML

Persistent Identifier
DOI: [10.82556/yjen-c230](https://doi.org/10.82556/yjen-c230)

Title
Genre classification of 400 song excerpts

Description
Machine-learning classification task using SVM to classify the correct genre from 4

Publisher
Technische Universität Wien

Publication Date
2023-05-20

Related Identifiers
URL: <http://www.projects.science.uu.nl/memation/emotifydata/> (IsDerivedFrom)
URL: <https://gitlab.tuwien.ac.at/martin.weise/dbrepo-ismir> (IsSupplementTo)

Citation
L., Mahler. (2023). Genre classification of 400 song excerpts. Technische Universität Wien. <https://doi.org/10.82556/yjen-c230>

Subset Information

Database Visibility
Public

Database Name
Ethnomusicology Data

Query Statement
`select 'track_id', 'genre', 'amazement', 'solemnity', 'tenderness', 'calmness', 'power', 'joyful_activation', 'tension', 'sadness', 'mother_tongue', 'disliked', 'age', 'gender', 'mother_tongue' from 'classification'`

Description

Subset hash

Title

Citation recommendation

Related identifiers

Subset Hash
sha256:e538fde57cc1f10fcefe243e2c4cc406b0e225232fc325ae1a65613a8f72f18c

Subset Creation
2023-05-21 11:42:47 (UTC)

Result Visibility
Public

Result Hash
3f375ccc7a65348ce5d26d07bab54e1a63

Result Number
8407

Subset data

nostalgia	tenderness	amazement	tension	mood	gender	joyful_activation	sadness	liked	track_id	solemnity
false	false	false	true	3	true	true	false	true	1	true
true	false	false	false	3	true	false	false	false	1	false
true	false	false	false	3	true	false	true	false	1	false
false	false	false	false	3	false	false	false	false	1	false
true	false	false	false	4	false	false	false	false	1	false
true	false	false	false	3	true	false	true	false	1	true
false	false	false	false	4	false	false	false	false	1	false
false	true	false	false	4	true	false	false	true	1	true
false	false	false	false	2	false	true	false	true	1	false
false	false	false	false	5	false	false	false	true	1	false

Subset query

OAI-PMH Endpoint

Rows per page: 10 1-10 of 8407

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