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

¹TU Wien  ²University of Music and Performing Arts Vienna  ³Latvian Academy of Music
Overview
Common Data Management


Research Data Policy 2018

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. 6). Research data must be provided with persistent identifiers within the repository.

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

Research data and records are to be stored and made available in accordance with intellectual property laws and the requirements of third-party funders as well as applicable legal or contractual requirements (e.g. to 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.
Proposed Solution

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

- Analysis
- Hot Data
- Cold Data

- Code
- API
- Metadata
RDM Architecture at TU Wien

- **DSpace (ReposiTUm)**
- **GitLab**
- **INVENio (TUWRD)**
- **DBRepo**

**Project Start**
- Publications
- Sourcecode & Documentation
- Structured research data in databases (e.g., live data from data streams (IoT, continuous measurements))

**Dissemination**
- Publications
- Sourcecode & Documentation
- File-based data

**Project End**
- Publications
- Sourcecode & Documentation
- File-based data
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/
Title of the dataset: A QR-Code optical covert channel in an air-gapped secure data infrastructure


Persistent identifier (DOI): 10.34726/hss.2022.84700

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

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

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
The Sentinel-1 Global Backscatter Model (51GBM) - Mapping Earth’s Land Surface with C-Band Microwaves

Citation
Bauer, Matthias; Berendt, C.; Casas, Steinar; Nascimento, Claudia; Freeman, Halil; Heiß, Peter; Krieger, Carsten; Marín, Manuel; Vega, Francisco; Cella, D.; Socie, Paul; Altherr, Frank; Emery, Christian; Wagner, Wolfgang.

Description of the dataset
The Sentinel-1 Global Backscatter Model (51GBM) is a global dataset of backscatter coefficients derived from Sentinel-1A/B C-band SAR data. The model is based on a statistical approach that uses a large database of SAR images to estimate backscatter coefficients for different land cover types. The dataset is available for download in various formats and includes information on the spatial resolution, frequency, and polarization of the SAR images used in the model.

Version of the dataset
Version 1.0 - August 2023

Persistent identifier (DOI)
https://doi.org/10.48436/n2d1v-gqb91

License
CC-BY-NC-SA-3.0

Preview file
The Sentinel-1 Global Backscatter Model (51GBM)

Files for download
- 51GBM_VH_mean_model_v5_EU27_AFR10M.zip
- 51GBM_VH_mean_model_v5_EU27_AFR10M_zp.zip
- 51GBM_VH_mean_model_v5_EU27_AFR10M.zip
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://purl.org/ontology/mo/Genre

Assign Semantic Information
We recommend the following ontologies:

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

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

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

RDF
SPARQL
Jena
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/](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
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

Analysis Environment

```python
def suppress_stdout_stderr():
    """A context manager that redirects stdout and stderr to devnull""
    with open(os.devnull, 'w') as fnnull:
        with redirect_stder(fnnul) as err, redirect_stdout(fnnul) as out:
            yield out, err

def generate_mfcc_feature(filepath: Path, sr: int = DEFAULT_SAMPLING_RATE, number_mfccs: int = 40):
    x, sr = librosa.load(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)
    return x, sr

with suppress_stdout_stderr(), ThreadPoolExecutor(5) as executor:
    dataframes = list(executor.map(
        lambda args: generate_mfcc_feature(args[0], args[1]),
        files
    ))
```

```
# grid for C, gamma
C_grid = [0.005, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 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(accuracy_score(grid.predict(x_val), y_val))
```

```
meta_columns = ["sample", "filename", "label"]
meta = pd.DataFrame(meta_features["filename"]), last())
meta.columns = pd.MultiIndex.from_tuples(["[1, 1]", "]\]
# needed for merge
meta = pd.concat([meta, meta_features["filename"]])
meta = meta.dropna(subset=['features', 'right_index'])
# create multi index to single index
one_level_cols = ["[".join(["for el in col") for col in meta.columns.levels[1]]
one_level_cols.insert(0, "label")
```

```
files = meta["filename"]
meta_features = meta["meta_features"]
```

```
Reproducing Research Results

Link to Git repository

Branch name or Commit hash

Launch Jupyter notebook

Action log
Deposit structured data from start

Persistent Identifier

Title

Citation recommendation

Subset query

Download data

Related identifiers

Subset hash

Subset data

OAI-PMH Endpoint
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