Repository Infrastructure Supporting Virtual Research Environments

Martin Weise[a,b], Tomasz Miksa[a,b], Tobias Grantner[a], Josef Taha[c], Maximilian Moser[c], Sotirios Tsepelakis[c], Barbara Sanchez Solis[b] and Andreas Rauber[a]

[a] *TU Wien, Research Unit Data Science*
[b] *TU Wien, Center for Research Data Management*
[c] *TU Wien, TU.it*

Research increasingly becomes data-driven, with vast amounts of information being generated and analyzed to produce new insights and discoveries. This data deluge requires a combination of methods and technologies to store, process, share and reuse research data. Since the inception of the guiding principles for scientific data management (FAIR) [4], many projects addressed these problems for datasets available as files, dumps or compressed archives. Academic institutions subsequently developed research data policies that recommend putting research data in adequate repositories.

Our Center for Research Data Management in cooperation with our IT department (TU.it) started operating the research data repository[1] (TUWRD). This works well for non-structured data and data deposited for archival purposes. It is less suitable for structured or evolving data stored in databases. For example: a recently added Danube water pollution measurement dataset [2] is available as PostgreSQL dump. To use this dataset, researchers have to: (i) set-up a compute environment that has the necessary resources, (ii) know how to deploy a database and import the PostreSQL dump, (iii) use a compatible version of the Postgres interactive terminal `psql` (v15.2) installed for the import. Only after performing these steps, the complete dataset can be inspected and used by the researcher and machines. Further, important information such as the table schema is not available in machine-readable format, but in a human-readable `PDF` text report. Last, but not least, updated measurements that arrive as a continuous data stream cannot be immediately used for research, but only become available after batch updates of deposited database dumps.

To mitigate these problems, we developed a database repository infrastructure DBRepo [3] that manages research data in databases, from the beginning of a project with the capability of providing access to time-versioned subsets (instead of periodic snapshots) that can be generated at arbitrary points in time and cited in e.g. publications using persistent identifiers. A database in DBRepo is managed by IT-professionals and curated by data stewards who annotate the dataset with metadata such as controlled vocabulary and ontology-mappings.

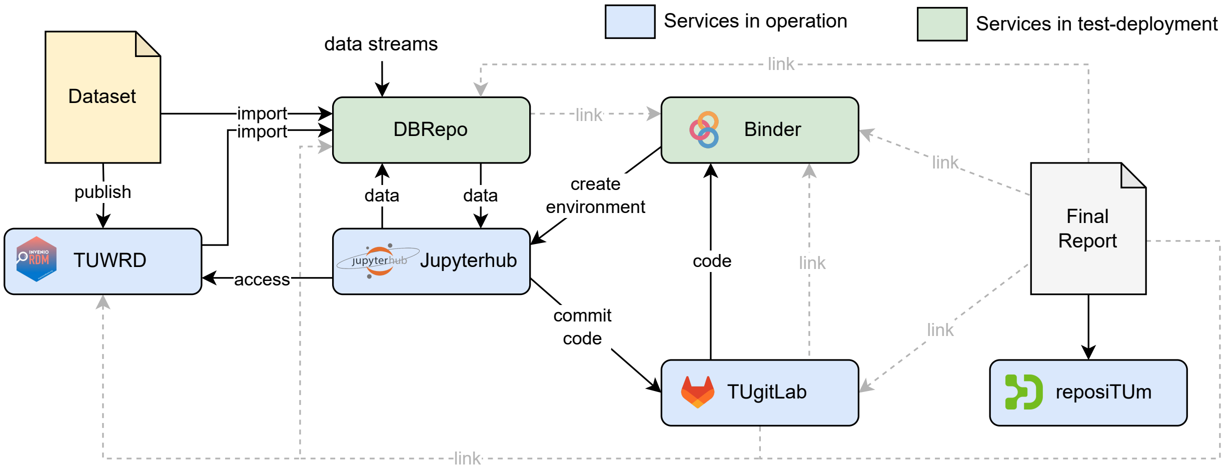

Figure 1: Research data infrastructure at TU Wien. Nodes in green are currently in development

We present a virtual research environment (c.f. Figure 1) including a set of four repository systems for files,

[1] `https://researchdata.tuwien.ac.at/`, based on InvenioRDM

databases, code and publications, connected to a Jupyterhub compute platform supporting automatic deployoment of code via Binder. The technical framework consists entirely of open-source software: a repository for databases *DBRepo*[2] [3] and continious data streams such as from IoT sensors, a compute environment *Jupyterhub*[3] closely located to this repository, the TUWRD repository to preserve database snapshots and unstructured research data suck as images, publication repository *reposiTUm*[4], code repository *TUgitLab*[5] as well as a visualization environment *Grafana*[6].

The key is to connect each of these solutions to a degree where each record points to another with a persistent identifier and to offer data import and access capabilities for the user, to support the full lifecycle of research data while increasing external visibility of the data. We find that a separation of concerns is crucial: researchers work with their data, IT-experts take care of the operation and maintenance, while data stewards curate the data and handle persistent identifier registration of the repositories. Our repository infrastructure fulfills the requirements of virtual research environments [1] by providing a web-based working environment, being tailored to the needs of researchers within the data science domain, providing this community with data ingest/compute/egress tools, being open and flexible and allowing access controls to share both intermediate and final research outputs while asserting ownership, provenance and attribution.

The publication repository, code repository, and research data repository are already operational as core services and open to scientific employees, whereas DBRepo is currently being operated on a friendly-user basis, together with the integration of the JupyterHub compute environments and Binder. Annotation of this metadata in DBRepo is done by a data steward for each table column, by providing an URI link to the semantic concept of this column, as well as the URI link to the unit of measurement, where applicable. Interlinking data with semantic web technology is necessary to provide machine-readable interfaces and allow the retrieval of resources based on their semantic concepts and unit of measurement. Future work includes connecting this infrastructure further by implementing a cross-repository search, listing available datasets regardless of their physical location, creation of databases in DBRepo based on their archived snapshot in TUWRD and automated reporting of available intellectual property in this infrastructure for academic intellectual property reports and the generation of a scientific knowledge graph.

We summarize our contributions towards "policy and practice of data in research" as follows:

- providing a virtual research environment that can host research data from the start of a project,
- separating concerns between researchers, data stewards curating the data and IT-experts maintaining the environment,
- improving FAIRness of the research data by curation of metadata by data stewards in conjunction with the researchers
- linking metadata to semantic concepts and controlled vocabulary

References

[1] Leonardo Candela et al. Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal*, 12(0):GRDI75–GRDI81, 2013. doi:10.2481/dsj.grdi-013.

[2] Steffen Kittlaus, Adrienne Clement, Máté Krisztián Kardos, Katalin Mária Dudás, Nikolaus Weber, Ottavia Zoboli, and Matthias Zessner. Inventory of hazardous substance concentrations in different environmental compartments in the danube river basin, 2023. doi:10.48436/xwve4-h7v43.

[3] Martin Weise et al. DBRepo: a Semantic Digital Repository for Relational Databases. *International Journal of Digital Curation*, 17, 2022. doi:10.2218/ijdc.v17i1.825.

[2] https://dbrepo1.ec.tuwien.ac.at

[3] https://dbrepo1.ec.tuwien.ac.at/jupyterhub, based on Jupyterhub

[4] https://repositum.tuwien.ac.at, based on DSpace

[5] https://gitlab.tuwien.ac.at/, based on Gitlab

[6] https://dbrepo1.ec.tuwien.ac.at/grafana

[4] Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, Mar 2016. `doi:10.1038/sdata.2016.18`.