Research Data Infrastructure Landscape at TU Wien

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

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

In the era of big data, research has become increasingly 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 preserve research data. The TU Wien research data infrastructure landscape enables data citation as well as collection of provenance and allows research data to be findable, accessible, interoperable and reusable (FAIR).

The intellectual property of TU Wien is very diverse. Examples include the "Sentinel-1 Global Backscatter Model" datacube [1] that gives a high-quality impression on surface- structures and -patterns of the Earth, a paper self-archive for an "Open-Source River Basin Management System" needed for the yearly intellectual capital statement and Jupyter Notebooks of a recent training event of the Vienna Scientific Cluster. Since there is no single solution that fits our needs of storing, processing, sharing, and preserving this collected research data, a landscape of repositories (c.f. Fig. 1) based on open-source software was, where available, selected, or developed.

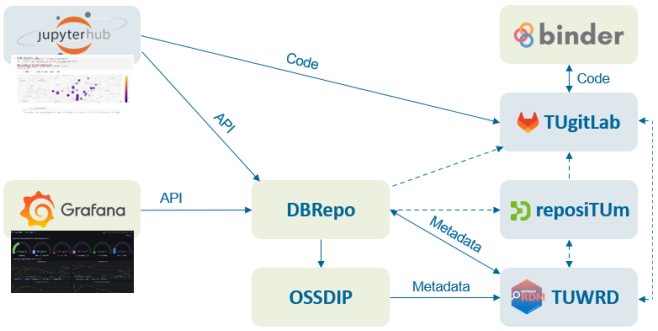

Fig. 1: Research data infrastructure at TU Wien. Nodes in green are currently in development.

File-based research data is managed by the research data repository TUWRD[1], text documents such as publications in the publication repository reposiTUm[2] and analysis steps, such as algorithms and software code should be deposited at the code repository TUgitLab[3] (and not in the research data repository although it is technically possible). These repositories are complemented by a compute platform (JupyterHub), forming a virtual research environment that allows seamless deployment of e.g., Jupyter Notebooks stored in TUgitLab via Binder, which in turn can read data files from TUWRD and/or DBRepo [3], perform some analysis and store results again directly into these repositories. All artifacts can be linked via persistent identifiers (PIDs, e.g. DOI) supporting provenance tracing, citation, and reproducibility of experiments. This virtual research environment (JupyterHub) and visualization tools (e.g., Grafana) can be conveniently accessed via a web browser. The key is to connect each of these solutions and offer seamless integration for the user to support the full lifecycle of research data while increasing external visibility of the data (for sensitive data we are currently developing a secure data infrastructure blueprint [2]).

References

[1] Bernhard Bauer-Marschallinger, Senmao Cao, et al. The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves, 2021. doi:10.48436/n2d1v-gqb91.

[2] Martin Weise, Filip Kovacevic, Nikolas Popper, et al. OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. *Data Science Journal*, 21. doi:10.5334/dsj-2022-004.

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

[1] TUWRD is based on InvenioRDM. [Online]. URL: https://researchdata.tuwien.ac.at/
[2] reposiTUm is based on DSpace. [Online]. https://repositum.tuwien.at/
[3] TUgitLab is based on GitLab. [Online]. https://gitlab.tuwien.ac.at/