

## Introduction

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.

## Repositories

The Center for Research Data Management in cooperation with the IT department (TU.it) started operating the **research data repository<sup>a</sup>** (TUWRD) based on InvenioRDM [1] to host non-structured data.

Algorithms and code can be deposited into the code repository **TUgitLab<sup>b</sup>** based on Gitlab. Publications, reports and presentations can be deposited into the publication repository **repositUm<sup>c</sup>**, based on DSpace.

To complement these, the **database repository<sup>d</sup>** (DBRepo) [2] based on Java (Spring Boot), Python (Flask), open source components RabbitMQ, Keycloak, MariaDB and OpenSearch was developed that manages structured research data.

- Work with evolving research data
- Versioning and timestamping all changes
- Enable citation of subsets, views and databases

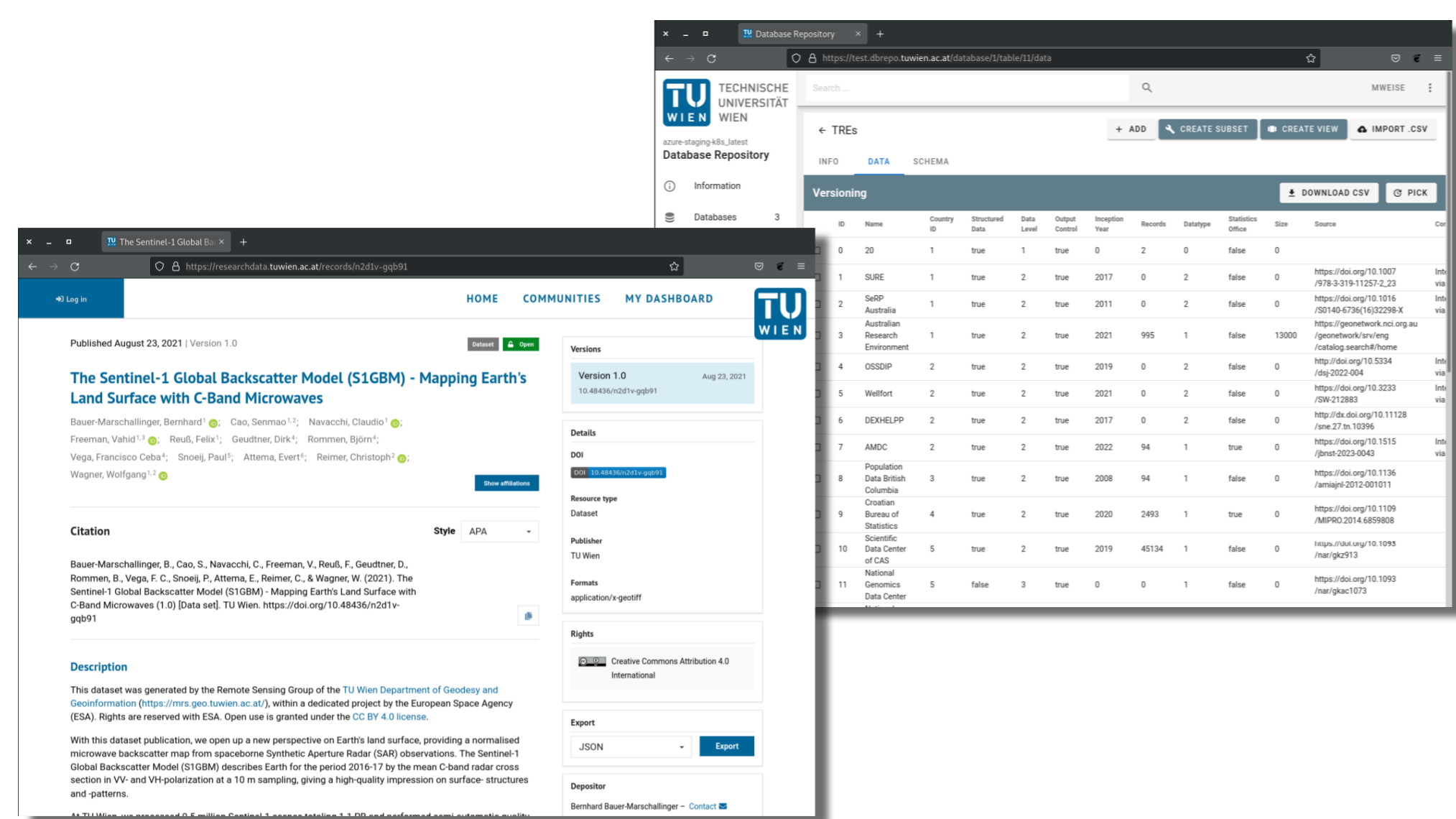


Figure 1: TUWRD repository for files and DBRepo for databases

## Compute Platforms

Researchers at TU Wien that perform data science can use the managed **JupyterHub** environment, allowing for interactive processing on *high-performance computing* environments located also at TU Wien. Typically these notebooks are stored also in repositUm, from where the whole environment (Python version, packages) can be reproduced with **Binder**. The two data repositories TUWRD and DBRepo (or external sources) can be used as input for data streams visualized in **Grafana**.

<sup>a</sup><https://researchdata.tuwien.ac.at/>

<sup>b</sup><https://gitlab.tuwien.ac.at/>

<sup>c</sup><https://repositum.tuwien.at/>

<sup>d</sup><https://test.dbrepo.tuwien.ac.at/>

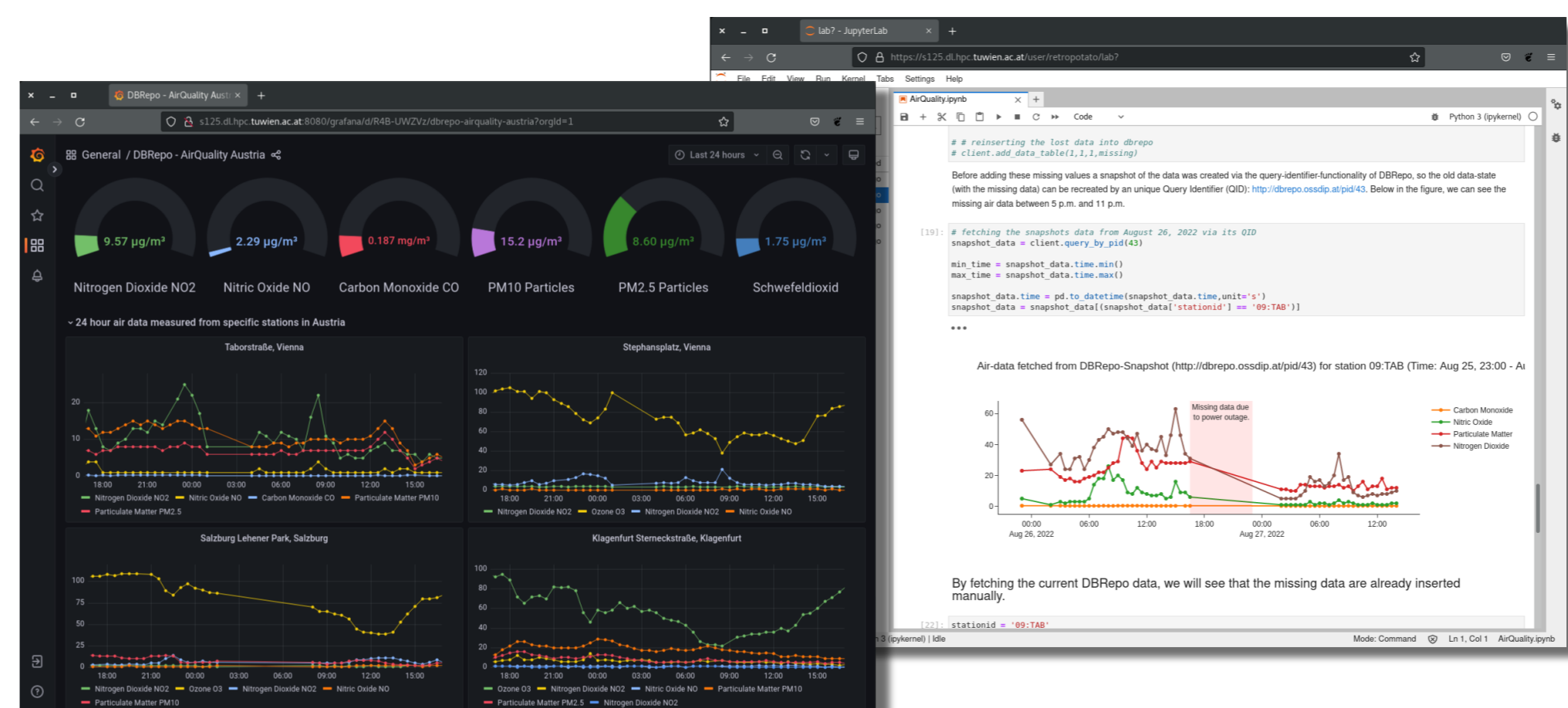


Figure 2: External analytic/monitoring tools connected to DBRepo

## Results

The key is to connect each of these solutions (c.f. Figure 3) so that 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.

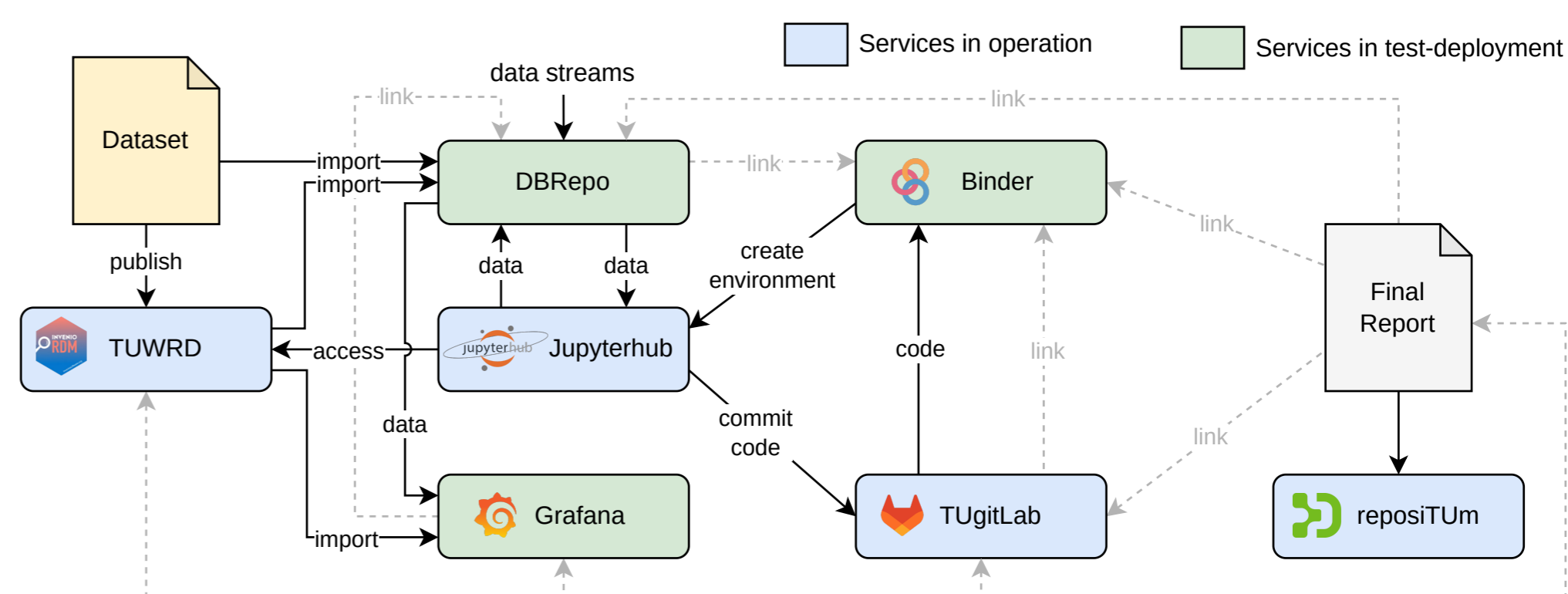


Figure 3: Repository infrastructure (TUWRD, DBRepo, TUgitLab, repositUm) supporting virtual research environments (Jupyterhub, Grafana, Binder).

We present a virtual research environment:

- 4 repository systems for files, databases, code and publications; connected to
- 3 compute platforms for analysis

## Conclusions & Future Work

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. Future work includes connecting this infrastructure further by implementing a cross-repository search regardless of the physical location of the data and automated reporting of available intellectual property in this infrastructure and the generation of a scientific knowledge graph.

## References

[1] InvenioRDM. <https://inveniordm.web.cern.ch/>.  
 [2] M. Weise et al. DBRepo: a Semantic Digital Repository for Relational Databases. *International Journal of Digital Curation*, 17(1), 2022. doi:10.2218/ijdc.v17i1.825.