Repository Infrastructure Supporting Virtual Research Environments

Martin Weise, Tomasz Miksa, Tobias Grantner, Josef Taha, Maximilian Moser, Sotirios Tsapelakis, Barbara Sanchez Solis and Andreas Rauber

TU Wien, Research Unit Data Science
TU Wien, Center for Research Data Management
TU Wien, TU.it

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 (TUWRD) based on InvenioRDM [1] to host non-structured data.

Algorithms and code can be reposted into the code repository TUgitLab based on Gitlab. Publications, reports and presentations can be deposited into the publication repository reposiTUm, based on DSpace.

To complement these, the database repository 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

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.

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.

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


Contact: Martin Weise martin.weise@tuwien.ac.at