

# **Repository Infrastructure Supporting Virtual Research Environments**

Martin Weise<sup>a,b</sup>, Tomasz Miksa<sup>a,b</sup>, Tobias Grantner<sup>a</sup>, Josef Taha<sup>c</sup>, Maximilian Moser<sup>b</sup>, Sotirios Tsepelakis<sup>b</sup>, Barbara Sanchez Solis<sup>b</sup> and Andreas Rauber<sup>a</sup>

<sup>a</sup>TU Wien, Research Unit Data Science <sup>b</sup>TU Wien, Center for Research Data Management <sup>c</sup>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**<sup>*a*</sup> (TUWRD) based on InvenioRDM [1] to host non-structured data.

Algorithms and code can be reposited into the code repository  $\mathbf{TUgitLab}^b$  based on Gitlab. Publications, reports and presentations can be deposited into the publication repos-



DSPACE &binde



# 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.

itory **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

	× _																		
		← → C O A https://test.dbrepo.tuwien.ac.at/database/1/table/11/data									a						☆ ⊽ €	=	
			F		NISCHE	Search	Search						Q				MWEISE		
						← TREs					+ ADD 🔧 CREATE S			SUBSET	CREA	TE VIEW & IMPORT .CS	A IMPORT .CSV		
			Dat	abase Repos	Repository INFO DATA SCHEMA														
			()	Information		Versio	ning									± 0	OWNLOAD CSV @ PIC	к	
× _ □ III The Sentinel-1 Global Bar × +			(()	Databases	3	D	Name	Country ID	Structured Data	Data Level	Output Control	Inception Year	Records	Datatype	Statistics Office	Size	Source	C	
← → C O A https://researchdata.tuwien.ac.at/records/n2d1v-gqb91			☆				20 SURE	1	true	1	true	0	2	2	false	0	https://doi.org/10.1007	Ir	
θð Log in	номе сомм		DASHB	BOARD		2 2	SeRP	1	true	2	true	2011	0	2	false	0	/978-3-319-11257-2_23 https://doi.org/10.1016 /S0140-6736(16)32298-X	vi Ir	
Published August 23, 2021   Version 1.0	Dataset 🔓 Open	Versions			WIEN	3	Australian Research Environment	1	true	2	true	2021	995	1	false	13000	https://geonetwork.nci.org.au /geonetwork/srv/eng /catalog.search#/home		
The Sentinel-1 Global Backscatter Model (S1GBM) - Mapp	ing Earth's	Version 1.0		Aug 23, 2021		3 4	OSSDIP	2	true	2	true	2019	0	2	false	0	http://doi.org/10.5334 /dsj-2022-004	lr vi	
Land Surface with C-Band Microwaves		10.48436/n2d1v-gqb	91		5	Wellfort	2	true	2	true	2021	0	2	false	0	https://doi.org/10.3233 /SW-212883	lr vi		
Bauer-Marschallinger, Bernhard <sup>1</sup> ; Cao, Senmao <sup>1,2</sup> ; Navacchi, Claudio <sup>1</sup> ;	Details				3 6	DEXHELPP	2	true	2	true	2017	0	2	false	0	http://dx.doi.org/10.11128 /sne.27.tn.10396			
Vega, Francisco Ceba <sup>4</sup> ; Snoeij, Paul <sup>5</sup> ; Attema, Evert <sup>6</sup> ; Reimer, Christoph <sup>2</sup> (6);		DOI				7	AMDC Population	2	true	2	true	2022	94	1	true	0	/jbnst-2023-0043	v	
Wagner, Wolfgang 🗠 👩	Show affiliations	Resource type	lpar .			8	Data British Columbia Croatian	3	true	2	true	2008	94	1	false	0	/amiajnl-2012-001011		
Citation Style	APA -	Dataset				9	Bureau of Statistics	4	true	2	true	2020	2493	1	true	0	https://doi.org/10.1109 /MIPRO.2014.6859808		
Bauer-Marschallinger, B., Cao, S., Navacchi, C., Freeman, V., Reuß, F., Geudtner, D., Rommen, B., Vega, F. C., Snoeij, P., Attema, E., Reimer, C., & Wagner, W. (2021). The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves (1.0) [Data set]. TU Wien. https://doi.org/10.48436/n2d1v- gqb91		Publisher TU Wien				3 10	Data Center of CAS	5	true	2	true	2019	45134	1	false	0	Intips://dui.org/10.1093 /nar/gkz913		
		Formats application/x-geotiff	Formats application/x-geotiff			11	National Genomics Data Center	5	false	3	true	0	0	1	false	0	https://doi.org/10.1093 /nar/gkac1073		
	8	Rights				г													
Description		Creative Co	ommons / al	Attribution 4.0		L.													
This dataset was generated by the Remote Sensing Group of the TU Wien Department of Geo Geoinformation (https://mrs.geo.tuwien.ac.at/), within a dedicated project by the European S																			
(ESA). Rights are reserved with ESA. Open use is granted under the CC BY 4.0 license. With this dataset publication, we open up a new perspective on Earth's land surface, providing a normalised		Export JSON		✓ Export		L.													
microwave backscatter map from spaceborne synthetic Aperture Radar (SAR) observations. Global Backscatter Model (S1GBM) describes Earth for the period 2016-17 by the mean C-ba section in VV- and VH-polarization at a 10 m sampling, diving a high-multiv impression on su	The Sentinel-1 Ind radar cross Inface- structures				-														
and -patterns.		Depositor Bernhard Bauer-Marsch	hallinger ·	- Contact 💌															

Figure 1: TUWRD repository for files and DBRepo for databases



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.

#### **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**.

## References

 InvenioRDM. https://inveniordm.web.cern.ch/.
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.

ahttps://researchdata.tuwien.ac.at/

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

<sup>&</sup>lt;sup>c</sup>https://repositum.tuwien.at/

dhttps://test.dbrepo.tuwien.ac.at/