1st Conference on Research Data Infrastructure

Natural Sciences

https://doi.org/10.52825/CoRDI.v1i.277

Published: 07 Sept. 2023

Overarching Data Management Ecosystem at HZDR

From Small Experiments to Large-Scale Research Facilities

Oliver Knodel[1] [https://orcid.org/0000-0001-8174-7795], Thomas Gruber[1] [https://orcid.org/0000-0001-6940-2065], Jeffrey Kelling[1] [https://orcid.org/0000-0003-1761-2591], Mani Lokamani[1] [https://orcid.org/0000-0001-8679-5905], Stefan Müller[1] [https://orcid.org/0000-0001-6273-7102], David Pape[1] [https://orcid.org/0000-0002-3145-9880], Martin Voigt[1 2] [https://orcid.org/0000-0001-5556-838X], and Guido Juckeland[1] [https://orcid.org/0000-0002-9935-4428]

[1]Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Germany

[2]Technische Universität Dresden, Germany

Abstract: When dealing with research data management, researchers at Helmholtz-Zentrum Dresden – Rossendorf (HZDR) face a variety of systems and tools. These range from the project planning phase (proposal management, data management plans and policies), over documentation during the experiment or simulation campaign, to the publication (collaborative authoring tools, metadata catalogs, publication systems, data repositories). In addition, modern research projects usually are required to interact with a variety of software stacks and workflow management systems to allow comprehensible and FAIR science on the underlying IT infrastructure (HPC, data storage, network file systems, archival). This article first demonstrates the data management systems and services provided at HZDR, followed by an overview of a self-developed guidance system. It is concluded by a real-world example.

Keywords: research data management, data life cycle, workflows, metadata, FAIR, data provenance, HELIPORT

1 Data Management Ecosystem at HZDR

Over the last years, we have been developing a uniform data management ecosystem aiming to make the entire life cycle of a scientific project – from the submission of a beamtime proposal to the publication of produced datasets – comprehensible according to the FAIR principles [1]. Figure 1 shows this ecosystem consisting of the various systems and services in our IT landscape. The systems access the underlying hardware in the data centre without abstraction via the common interfaces.

Visiting scientists typically start their journey through our infrastructure by submitting a proposal to the proposal management system *Gate* [3] on the left side of Figure 1. After acceptance, the first basic set of metadata, such as researcher names and Open Researcher and Contributor IDs (ORCiDs) [4], title and abstract of the experiment, as well as experiment type or the facility used, are known. Using this initial data, a data management plan (DMP) can be generated with the help of the Research Data Management Organiser (RDMO) [5].

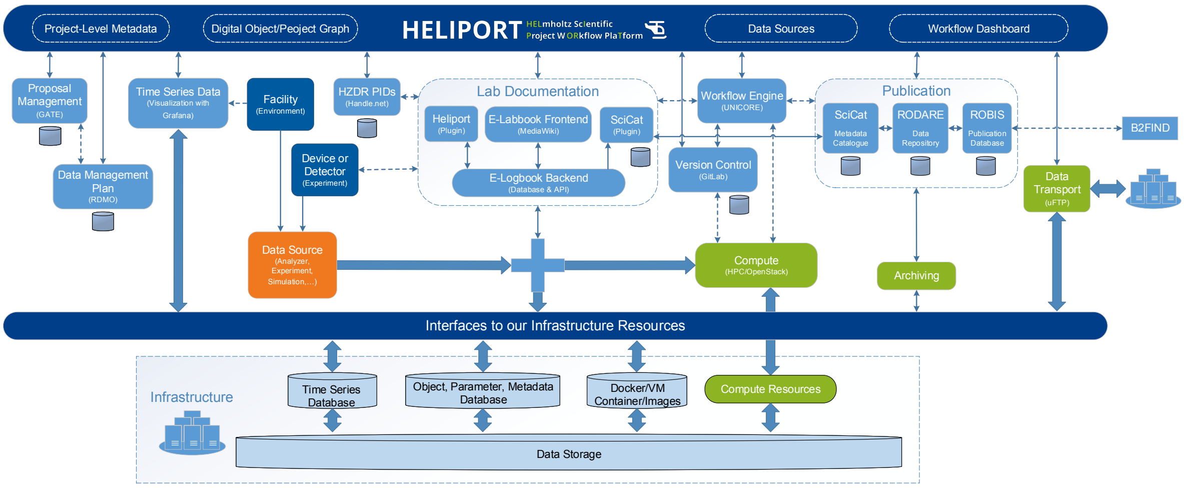
Project-Level Metadata
Digital Object/Peoject Graph
HELIPORT
HELmholtz ScIentific Project WORkflow PlaTform
Data Sources
Workflow Dashboard
Proposal Management (GATE)
Data Management Plan (RDMO)
Time Series Data (Visualization with Grafana)
Facility (Environment)
HZDR PIDs (Handle.net)
Device or Detector (Experiment)
Lab Documentation
Heliport (Plugin)
E-Labbook Frontend (MediaWiki)
SciCat (Plugin)
E-Logbook Backend (Database & API)
Workflow Engine (UNICORE)
Version Control (GitLab)
Publication
SciCat Metadata Catalogue
RODARE Data Repository
ROBIS Publication Database
B2FIND
Data Transport (uFTP)
Data Source (Analyzer, Experiment, Simulation,...)
Compute (HPC/OpenStack)
Archiving
Interfaces to our Infrastructure Resources
Infrastructure
Time Series Database
Object, Parameter, Metadata Database
Docker/VM Container/Images
Compute Resources
Data Storage

Figure 1. Top-Level Architecture of the HZDR Data Management ecosystem with the various underlying systems and services. [2]

Experiment control, data acquisition and facility (meta)data are provided by different (experiment-specific) subsystems, and are typically stored in a time series database. For visualization, a central Grafana OSS [6] instance is available. Together with automated and user-defined documentation in our e-labbook (based on Semantic MediaWiki [7]), the actual data from an experiment or a simulation (digital twin), is used for pre- and post-analysis workflows on a high performance computing (HPC) cluster. The access to our HPC cluster is provided by the Uniform Interface to Computing Resources (UNICORE) [8]. A combination of metadata- and data-repository is used for the publication of results and the registration in our publication database Rossendorf Bibliography System (Robis). The experiment-specific metadata can be transferred or entered directly into a metadata catalog based on SciCat [9] with a direct link to the data available in our data repository Rossendorf Data Repository (Rodare) [10], which is based on Zenodo [11] and Invenio [12].

We plan to provide additional Digital Object Identifiers (DOIs) for our large scale facilities and beamlines. In the past, beamline scientists created descriptions of their facilities in the Journal of large-scale research facilities (JLSRF) as for instance [13]. Due to the fact that most of our facilities have not published a citable description, uniform landing pages are created, such as [14] for the NELBE facility. This information, as well as the proposal metadata, can be attached to the publication of the dataset on Rodare as a related identifier and the environment of an experiment where a dataset was created can be comprehended and reproduced.

2 Guidance System – HELIPORT

Over the last years, the overarching layer (or guidance system) of our data management ecosystem, introduced in section 1, received the name *HELIPORT*. This abstraction layer HELIPORT [15], [16] is an overall data management solution that aims at making the steps of the entire research experiment's life-cycle findable, accessible, interoperable and reusable according to the FAIR principles. In doing so, it makes the components involved in the project discoverable for new team members and provides valuable functionality to exchange data and metadata between systems.

Among other information, HELIPORT integrates documentation, scientific workflows and their results, and the final publication of the primary data and research results – all via already established solutions. Integration is accomplished by presenting the researchers with a high-level overview to keep all aspects of the experiment in mind.

Computational agents can interact with HELIPORT via a REST API that allows access to all components. Furthermore, all aspects of the experiment are registered as digital objects with landing pages that contain metadata in various standardized formats and schemas. Thus, the metadata is readable for both humans and machines. An overall digital object graph, combining the metadata harvested from all sources, provides the scientists with a visual representation of interactions and relations between their digital objects. Additionally, the project timeline lists all digital objects ordered by the time they were created. Through integrated computational workflow systems, HELIPORT can automate calculations using the collected metadata. We also created a HELIPORT team on WorkflowHub [17] to allow users to exchange their workflows.

By visualizing all aspects of large-scale research experiments, HELIPORT enables deeper insights into a comprehensible data provenance with the chance of raising awareness for data management.

3 Data Management View of the TELBE Experiment in HELIPORT

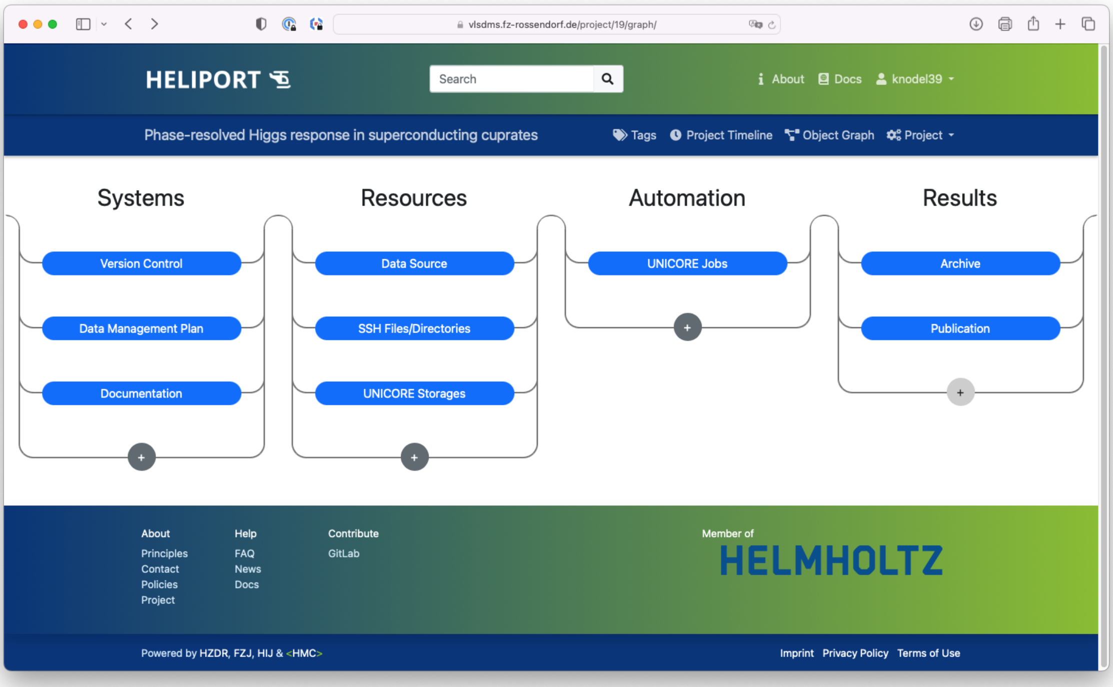

Figure 2. Overview page with all services and systems used in an exemplary High-Field High-Repetition-Rate Terahertz facility (TELBE) project.

As a first example to demonstrate the benefits of HELIPORT (see section 2) we used an experiment from the High-Field High-Repetition-Rate Terahertz facility (TELBE) [18], which is part of ELBE [13]. The HELIPORT project overview for this experiment is shown in Figure 2. A variety of the systems mentioned in section 1 are present here, e. g.:

Version Control gathers the source code for post-processing scripts stored in Helmholtz Codebase (a GitLab for the entire Helmholtz association).

Documentation: The TELBE group documents their experiment and record scientific metadata within the e-labbook. It is linked here to be easily findable and updateable.

UNICORE Jobs: This section provides access to the output logs of the UNICORE jobs that run post-processing, as well as some management functionality for these jobs implemented through the UNICORE API.

Publication: Here, the data publications of the post-processed primary data can be found. Datasets registered under **Resources** can be tagged and automatically published on Rodare.

This project also makes use of the HELIPORT REST API: The experiment control is implemented in LabView and makes API requests to HELIPORT for current experiment metadata. The metadata is then used to start new post-processing runs on the HPC cluster via UNICORE after each measurement. A callback URL provided by HELIPORT even allows registration of new jobs and datasets with the project without any user intervention. UNICORE posts status updates about running jobs to this URL and HELIPORT reacts accordingly.

In cooperation with the TELBE experiment we created a proof-of-concept to demonstrate the potential of our guidance system HELIPORT to abstract from the complex data management ecosystem it builds upon.

Author contributions

The authors contributed equally to this work.

Competing interests

The authors declare that they have no competing interests.

Funding

Our overall data management strategy and our ecosystem was funded by HZDR and the Helmholtz Association.

The HELIPORT project (ZT-I-PF-3-021) was funded by the “Initiating and Networking Fund of the Helmholtz Association” in the framework of the first “Helmholtz Metadata Collaboration” project call 2021.

Acknowledgements

We especially thank Uwe Konrad, Maik Fiedler, Jürgen Grzondziel and Jan-Christoph Deinert from HZDR in supporting our initiative, as well as the HMC colleagues responsible for the hub *Matter* for the enriching discussions. Last but not least, we would like to thank the entire HELIPORT team from Forschungszentrum Jülich, Helmholtz Institute Jena and HZDR.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, pp. 1–9, 2016, ISSN: 20524463. DOI: 10.1038/sdata.2016.18.

[2] O. Knodel, T. Gruber, J. Kelling, *et al.*, *HZDR Data Management Strategy — Top-Level Architecture*, Feb. 2023. DOI: 10.14278/rodare.2162. [Online]. Available: https://doi.org/10.14278/rodare.2162.

[3] User Office, Helmholtz-Zentrum Dresden-Rossendorf, *HZDR Proposal Management System*. [Online]. Available: https://gate.hzdr.de/user/.

[4] *Open Researcher and Contributor ID (ORCiD)*. [Online]. Available: https://orcid.org.

[5] J. Klar, O. Michaelis, C. Engelhardt, *et al.*, *Research Data Management Organizer (RDMO)*, version 1.3, Oct. 2020. DOI: 10.5281/zenodo.596581. [Online]. Available: https://github.com/rdmorganiser/rdmo.

[6] Grafana Labs, *Grafana OSS | Metrics, logs, traces, and more*. [Online]. Available: https://grafana.com/oss/grafana/.

[7] M. Krötzsch, D. Vrandečić, and M. Völkel, "Semantic mediawiki," in *International semantic web conference*, Springer, 2006, pp. 935–942. DOI: https://doi.org/10.1007/978-3-642-19797-0_16.

[8] K. Benedyczak, B. Schuller, M. Petrova-El Sayed, J. Rybicki, and R. Grunzke, "Unicore 7—middleware services for distributed and federated computing," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*, IEEE, 2016, pp. 613–620. [Online]. Available: http://hdl.handle.net/2128/12214.

[9] *Scicat metadata catalogue*. [Online]. Available: https://scicatproject.github.io.

[10] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *RODARE - Rossendorf Data Repository*. DOI: http://doi.org/10.17616/R3BR40.

[11] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: 10.25495/7GXK-RD71. [Online]. Available: https://www.zenodo.org/.

[12] *Invenio - Powering Open Science*. [Online]. Available: https://inveniosoftware.org.

[13] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *ELBE Center for High-Power Radiation Sources*. DOI: https://doi.org/10.17815/jlsrf-2-58.

[14] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), *Desctiption of the NELBE facility*. DOI: https://doi.org/10.58065/24017.

[15] O. Knodel, M. Voigt, R. Ufer, *et al.*, "Heliport: A portable platform for FAIR Workflow | Metadata | Scientific Project Lifecycle management and everything," in *Proceedings of the 4th International Workshop on Practical Reproducible Evaluation of Computer Systems*, ser. P-RECS '21, Virtual Event, Sweden: Association for Computing Machinery, 2021, pp. 9–14, ISBN: 9781450383950. DOI: 10.1145/3456287.3465477. [Online]. Available: https://doi.org/10.1145/3456287.3465477.

[16] M. Voigt, R. Ufer, W. Schacht, *et al.*, *HELIPORT (HELmholtz ScIentific Project WORkflow PlaTform)*, Nov. 2022. DOI: 10.14278/rodare.1970. [Online]. Available: https://doi.org/10.14278/rodare.1970.

[17] University of Manchester and HITS gGmbH, *HELIPORT team on workflowhub.eu*. [Online]. Available: https://workflowhub.eu/projects/156.

[18] M. Helm, S. Winnerl, A. Pashkin, *et al.*, "The elbe infrared and thz facility at helmholtz-zentrum dresden-rossendorf," *The European Physical Journal Plus*, vol. 138, no. 2, p. 158, 2023. DOI: https://doi.org/10.1140/epjp/s13360-023-03720-z.