1st Conference on Research Data Infrastructure
Natural Sciences
https://doi.org/10.52825/CoRDI.v1i.377

Published: 07 Sept. 2023

Two-Step Approach in Metadata Management for Data Publications at Research Centres

Thomas Gruber[1] [https://orcid.org/0000-0001-6940-2065], Hans-Peter Schlenvoigt[1] [https://orcid.org/0000-0003-4400-1315], Oliver Knodel[1] [https://orcid.org/0000-0001-8174-7795], Kristin Tippey[1] [https://orcid.org/0000-0002-9261-7643], and Guido Juckeland[1] [https://orcid.org/0000-0002-9935-4428]

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

Abstract: Data repositories like Zenodo have a limited list of metadata to search for. Metadata catalogues are designed to provide a community-specific parameters search, but their deployment has just started. These catalogues require metadata standards for interoperability, which in turn are often in development for many communities. To support publications with a metadata standard in the future, a two-step concept is presented in this article. It discusses how the electronic documentation should be constructed, in order to convert this later into a standardised schema for publication. We will present examples from the laser-plasma community for both steps, firstly how we deal with the complex challenges of metadata management and secondly for methods for developing metadata schemas.

Keywords: Data Management, Workflows, Metadata, Data Provenance

1 General Concept of metadata handling

Data publications via repositories like Zenodo [1] are becoming increasingly popular. Many centres around the world are setting up their own instances. Within these repositories one can search for metadata according to DataCite [2]. However, queries on domain-specific metadata require other repositories like metadata catalogues. A key challenge in their development is to deal with the different, domain-specific metadata schemata, not only for metadata aggregation, but also to construct a usable search engine.

The repositories need to be generic in the usage of metadata schemas since those often do not yet exist or are in development, and are likely to evolve later on as knowledge and techniques advance. On the other hand, scientists want to document what they have now – waiting for an established metadata schema is no choice! Therefore we propose an iterative approach. First we collect all metadata currently known in a structured manner. The second step, for data publication, would be restructuring the metadata into a new, documented schema, such that a third party can use the metadata. This comes with several advantages. The documentation of experiments and simulations remains independent from any schema development. Later reformatting would be possible if the schema changes. What it needs is a mapping between local structured metadata and public metadata schema. A curation and selection process can be implemented between data taking and publication. This comes at the cost of additional work of transformation and managing the mapping configuration.

The diversity of experiments at a research centre demands a high flexibility on the electronic documentation of experiments and simulations. Usually, the starting point is an electronic lab notebook (ELN), but additional databases can be involved to provide full documentation. All sources of electronic documentations will further be called electronic lab documentation and include the interconnections via IDs or links, which brings all metadata and data together. Each element of the e-lab documentation can store and provide metadata and data in a structured fashion like key-value pairs. An ELN is basically a database with a user friendly web front end to manually access and enter metadata. The ELN comes along with special features, which are the reason for the wide variety of ELNs.

At the Helmholtz-Zentrum Dresden – Rossendorf (HZDR), we focus on Mediawiki (based on Semantic MediaWiki [3]) as an ELN and integrate it into many systems which generate or accept metadata (Figure 1). Page templates allow to define structured data types inside the pages as key-value pairs. Input forms can be specifically designed to help and guide the scientist for manual entry of structured metadata. The web front end provides also a html editor to enter unstructured (meta)data, figures, screenshots etc. It is also possible to include or link information from external sources like databases, depending on the specific lab. If an experiment is performed by a device control software like Labview, that can be deployed to automatically send all data and metadata to the e-lab documentation. Likewise, computer simulations can be amended similarly to automatically submit the input parameters and metadata to the e-lab documentation.

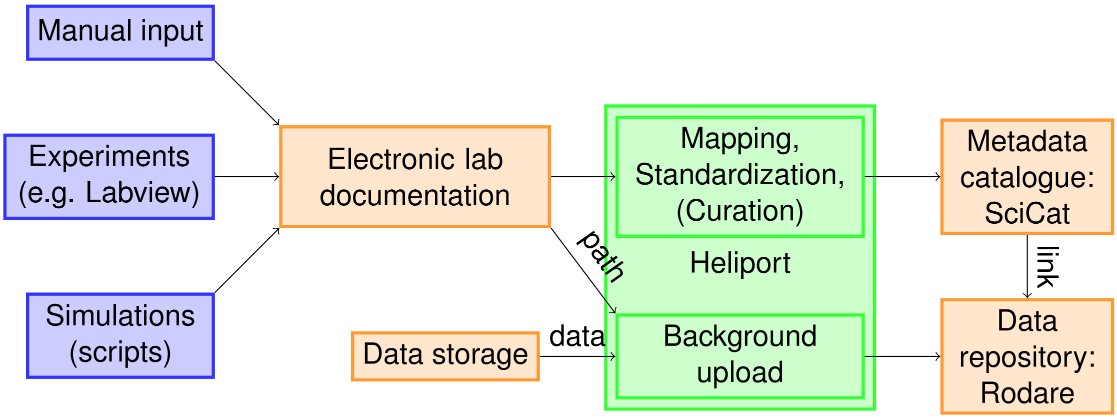
Manual input
Experiments (e.g. Labview)
Simulations (scripts)
Electronic lab documentation
path
Data storage
data
Mapping, Standardization, (Curation)
Heliport
Background upload
Metadata catalogue: SciCat
link
Data repository: Rodare

Figure 1. Metadata flow at HZDR. Data and metadata sources (blue) send the (meta)data to the storages (orange), initially to those on the left side. The workflow from primary to publicly available storage is depicted in green.

Once the information is collected in a structured manner, database-like queries can be used to retrieve specific information, not only within the e-lab documentation but also through the API. This is essential to forward the stored information to other services like the metadata catalogue SciCat [4]. A publication workflow could extract the metadata of the selected datasets, transform it into a SciCat [4] publication including all the metadata of the linked sample, project and instrument, by using the mapping configuration to create a certain metadata schema. In parallel the linked data storage path in the e-lab documentation is used to upload the data from the storage to our Rossendorf Data Repository (Rodare) [5], which is based on Zenodo [1] and Invenio [6]. Within this step the data could be combined with the metadata to create a HDF5 or Nexus file and upload this instead. The data is then referenced within SciCat as a direct link. To manage the workflows and documentation, HELIPORT [7] is used at HZDR.

In the following, we highlight two example activities of metadata management at HZDR where we touch different communities.

2 Example for (meta)data aggregation: DAPHNE4NFDI

The DAta from PHoton and Neutron Experiments for NFDI (DAPHNE4NFDI) project is centred at large-scale research facilities with photons and neutrons and aims for improved metadata management, commonalities in repositories and databases as well as establishing a software ecosystem for analysis software. HZDR's high-intensity laser group participates in DAPHNE4NFDI due to their research of laser-driven plasmas with XFELs. That research of laser-plasma interaction always employs both experiments and numerical simulations. Only the latter allow for insights into the micro-physics (on nanometer spatial and femtosecond temporal scales) whereas the former includes all effects without assumptions, approximations or models.

Within DAPHNE4NFDI, HZDR is working on improving metadata generation and capturing for both simulations and experiments. For simulations, input parameters, based on the Particle-In-Cell Modeling Interface (PICMI) [8], are filed as metadata into a database where simulation output is the data. Thereby, a searchable collection of already conducted simulations can be generated.

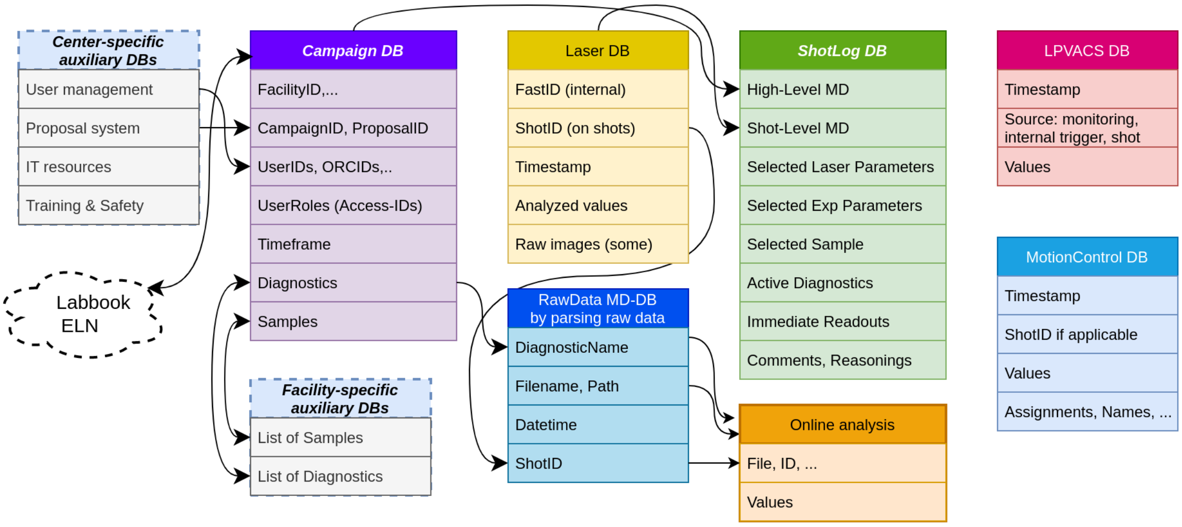

Figure 2. Current concept of a database system specific to laser-particle acceleration experiments at HZDR. This layout provides a concrete example of aggregating all currently available data and metadata as an electronic lab documentation as the first step in our approach.

For experiments, a database system for data and metadata is planned, see Figure 2. So far, raw data - mostly images - is initially stored on local file systems and subsequently aggregated into a file repository. There, structural metadata is encoded in the path structure but not independently searchable. In a first step, the path structure can be parsed into a database (dark blue). In parallel exists a database of the drive laser parameters (yellow). The sequence of laser shots is manually entered into the ShotLog database (green), alongside with comments on observations and decisions. This database is the cornerstone for later analysis. Further metadata and data sources, e.g. campaign IDs for metadata or instrument configuration for data, will be added incrementally.

3 Example for metadata schema development: HELPMI

The HElmholtz Laser Plasma Metadata Initiative (HELPMI) project is enabled by the "Helmholtz Metadata Collaboration" and aims at developing a metadata schema for laser-plasma experimental data, starting from the existing data and metadata standards openPMD [9] and NeXus [10]. So far, openPMD is widely used for laser-plasma simulation data, while NeXus is for experimental data from photon and neutron facilities. However, there is no data and metadata standard for experimental data from laser-plasma research.

HELPMI will develop - together with the global community - a glossary for laser-plasma experimental data. This is an important step towards data publications following the F.A.I.R. principles [11], such that there are commonly accepted definitions of terms in conjunction with domain-specific hierarchies.

HELPMI will make openPMD substantially extensible for custom hierarchies and will furthermore adopt the NeXus format, in particular in regard of geometry description and raw data and processed data handling. In combination with the glossary, openPMD can become a metadata standard for laser-plasma experiments and simulations. Furthermore there is potential that NeXus and openPMD - standards of two different scientific communities - can become interoperable.

Ultimately, once a metadata standard exists, the aggregated data and metadata can be mapped to that standard for F.A.I.R. data publications.

Author contributions

These authors contributed equally to this work.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the consortium DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortium is funded by the DFG - project number 460248799. The HELPMI project (ZT-I-PF-3-066) was funded by the "Initiative and Networking Fund" of the Helmholtz Association in the framework of the "Helmholtz Metadata Collaboration" project call 2022.

References

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

[2] D. M. W. Group *et al.*, "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs," 2021. DOI: https://doi.org/10.14454/3w3z-sa82.

[3] 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.

[4] *SciCat Metadata Catalogue*. [Online]. Available: https://scicatproject.github.io.

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

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

[7] 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.

[8] *Particle-In-Cell Modeling Interface*. [Online]. Available: https://picmi-standard.github.io/.

[9] *openPMD: A meta-data standard*. [Online]. Available: https://www.openpmd.org/.

[10] *the NeXus Data Format*. [Online]. Available: https://www.nexusformat.org/.

[11] 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.