1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.362

Published: 07 Sept. 2023

Exchanging Research Data with SampleDB

Malte Deckers[1] [https://orcid.org/0000-0001-9032-3062], and
Florian Rhiem[1] [https://orcid.org/0000-0001-6461-9433]

[1]Forschungszentrum Jülich GmbH, Germany

Abstract: Exchanging research data between systems is a vital task in research data management, especially in the context of collaborations. Therefore mapping data into standardized data descriptions and offering interfaces for efficient data exchange are required. The metadata database and electronic lab notebook SampleDB supports various ways of exporting data to other services, such as Dataverse, SciCat, and JupyterHub, as well as file exports in open formats. The concept of SampleDB federations allows data exchange in loosely coupled federations of SampleDB instances.

Keywords: Data Exchange, Electronic Lab Notebook, Federation, Integration, Research Data Management

1 Motivation

Persisting research data in a findable, accessible, interoperable, and reusable way, therefore implementing the FAIR principles stated by Wilkinson et al. [1], is an essential requirement for efficient (re)use of data and reproducible work and thus an important part of the rules of good scientific practice.

Research data can be collected by and stored in various systems, from raw data outputs, log files, or measurement control software to entries in electronic lab notebooks (ELNs), data catalogs, and repositories designated for data publication. In particular, inter-institutional and interdisciplinary collaborations require an efficient exchange between these systems, which might be operated by different organizations and use various data descriptions, concepts, and software solutions. Differing data descriptions and the use of various metadata schemas, ontologies, and vocabularies can result in isolated data silos, complicating data discovery and access. As these systems may be tailored to specific requirements, they can not easily be replaced by a shared infrastructure. Therefore solutions for an efficient and FAIR data exchange, keeping access restrictions, and preventing inconsistent duplicates and missing information on different systems are required.

The open-source, web-based research metadata database and electronic lab notebook SampleDB [2] allows the definition of individual metadata schemas fitted to specific environments and use cases. These schemas allow the description of process-specific metadatasets using various datatypes and conditions, that have to be met by a new record and are validated on creation. These are completed by attached files, a location history, comments, and publications to track the entire lifecycle of a datum,

e. g. a sample. To be able to keep track of measurements performed at other facilities and to make the locally-defined metadata useful for users from other institutions, SampleDB requires such methods for exchanging data with instances at other institutions or with other ELNs or metadata catalogs.

2 Data Exports

Sharing data with partners or preparing it for publication requires different ways of data export. Depending on the requirements this might include simple file exports in open formats that can be easily interpreted by the receiving end, as well as making information directly accessible to other computer systems via APIs.

In SampleDB there are several file-based export methods, such as a PDF document containing the metadata or archives as `.zip`, `.tar.gz`, containing the full metadata, including location assignments, comments, linked publications, and files.

To exchange data with other ELNs, the export and import of files using the standardized `.eln` file format, as defined by The ELN Consortium in [3], can be used. These exports can include multiple related objects as well, for example, to describe a whole process flow.

Besides these file exports, there are also methods of direct data export to other software systems, e. g. for data publication or data exploration and analysis.

Objects can be directly exported to Dataverse [4] repositories, with the process-specific metadata exported from SampleDB represented using the EngMeta [5] *Process Metadata* metadata block. A researcher can decide which metadata and related files should be shared with the Dataverse and when a SampleDB record is exported, a draft dataset is created to be reviewed, extended, and published by the researcher.

Records can also be made available to the SciCat [6] data catalog, by mapping the data description used in SampleDB to the categories and metadata fields used in SciCat.

To facilitate data analysis using the metadata stored in a SampleDB instance, SampleDB also supports a JupyterHub [7] integration. JupyterHub is a web service that can run Jupyter notebooks for multiple users, which can then be used for data analysis and exploration. SampleDB can provide relevant metadata to a notebook template server, which can then combine it with predefined notebook templates to create ready-to-run notebooks on JupyterHub. This way experienced researchers, e.g. instrument scientists, can prepare templates that allow guest scientists to more easily analyze and explore their data.

SampleDB also provides an HTTP API that allows users to implement custom export programs for systems that neither support the export file formats nor are supported by SampleDB directly.

3 Federation

The concept of SampleDB federations has been introduced in [8] to allow sharing of information in loose associations of SampleDB instances of collaborating institutions and facilities while keeping unique identifiers and tracking records across institutional boundaries.

This is accomplished by assigning universally unique identifiers (UUID) to SampleDB instances, which – combined with the unique identifiers of the objects – create a unique identifier for every dataset in a federation. These unique identifiers allow for keeping valid references within the federation, even when not all referenced information is shared. Authentication between databases is accomplished by exchanging pair-wise tokens when setting up a collaboration in a federation so that the federation can grow organically as its institutions and their researchers collaborate.

When a record is released to a federated instance it is first processed, for example, to add license information or to prevent personal or confidential data from being shared. Therefore the shared data and schema might not be exact copies of the original data and are transferred into an export schema. On the receiving side, data is then checked for consistency and validity to be accepted or declined.

In [8] a protocol to update imported data and send back the changes to the origin is proposed as well. To prevent conflicting object versions within the federation, the instance that created a record can review changes before accepting and probably redistributing them.

The federation API and data exchange format could as well be used by other services to be integrated into a more heterogeneous federation of metadata catalogs or electronic lab notebooks. However, a less complex data ex- or import, like the methods described above, might be more suitable and practical in many cases.

Underlying and related material

- SampleDB source code: https://github.com/sciapp/sampledb, DOI: 10.5281/zenodo.4012175
- SampleDB documentation: https://go.fzj.de/sampledb

Competing Interests

The authors declare that they have no competing interests.

Funding

–

Acknowledgements

We thank Daniel Kaiser for his contributions, especially in code review.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 160018, Mar. 2016. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://doi.org/10.1038/sdata.2016.18.

[2] F. Rhiem, "SampleDB: A sample and measurement metadata database," *Journal of Open Source Software*, vol. 6, no. 58, p. 2107, 2021. DOI: 10.21105/joss.02107. [Online]. Available: https://doi.org/10.21105/joss.02107.

[3] The ELN Consortium. "TheELNFileFormat." (2022), [Online]. Available: https://github.com/TheELNConsortium/TheELNFileFormat (visited on 04/21/2023).

[4] M. Crosas, "The dataverse network: An open-source application for sharing, discovering and preserving data," *D-Lib Magazine*, vol. Volume 17, 2011. DOI: 10.1045/january2011-crosas. [Online]. Available: https://doi.org/10.1045/january2011-crosas.

[5] B. Schembera and D. Iglezakis, "EngMeta - Metadata for Computational Engineering," *International Journal of Metadata, Semantics and Ontologies*, vol. 14, pp. 26–38, DOI: 10.1504/IJMSO.2020.107792. [Online]. Available: https://doi.org/10.1504/IJMSO.2020.107792.

[6] SciCat Project, *SciCat Project*. [Online]. Available: https://scicatproject.github.io/ (visited on 04/26/2023).

[7] Project Jupyter, *JupyterHub*. [Online]. Available: https://jupyter.org/hub (visited on 04/26/2023).

[8] M. Deckers, "Entwicklung eines föderierten Datenbanksystems zur verteilten Verwaltung von Forschungsdaten," Master's Thesis, FH Aachen, 2021. [Online]. Available: https://juser.fz-juelich.de/record/904981.