A Framework to Generate, Store, and Publish FAIR Data in Experimental Sciences

Nick Garabedian[1], Ilia Bagov[1], Malte Flachmann[1], Nuoyao Ye[1], Miłosz Meller[2], Floriane Bresser[1], Christian Greiner[1]
[1] *Karlsruhe Institute of Technology, Institute for Applied Materials, Germany*
[2] *Helmholtz-Zentrum Hereon, Institute of Membrane Research, Germany*

Abstract

Purpose: FAIR data is a relatively new paradigm in research data management which aims to facilitate reproducibility of research, knowledge generation, and knowledge retention in all scientific domains. This paper presents a framework which enables the semantic generation, storage, and publication of FAIR datasets in the field of experimental materials science with the help of controlled vocabularies.

Methodology: The framework presented in this work consists of multiple software tools developed by the authors, as well as an external electronic lab notebook (ELN), which is used as a database. The centerpiece of this solution is VocPopuli, a tool for the collaborative development of FAIR SKOS-based controlled vocabularies. These vocabularies are used as the basis of further software components developed by the authors, which enable the entry, processing, and publishing of FAIR datasets.

Findings: This paper shows that SKOS-based controlled vocabularies can be used as the cornerstone of FAIR data management systems in experimental materials science, and, in research and development as a whole. Furthermore, it demonstrates how these vocabularies can be part of common laboratory workflows in a seamless fashion which simplifies the generation, storage, and publication of FAIR data.

Value: The solution presented in this work enables the simplified creation of FAIR data without any additional effort from lab scientists, as most of the infrastructure is set up by the data stewards and the rest of the community. The controlled vocabularies, which are used to define the schemas of the generated datasets, facilitate the linking of external semantic resources, and increase the reproducibility of the research results. Furthermore, using our framework, these datasets can easily be published to open science platforms, so that other researchers can also benefit.

Conclusions: Integrating FAIR metadata in the production of FAIR data is not just a technical, but also a cultural issue. That is why, separating the creation of community and lab vocabularies, as well as, the specific templates for data input by lab scientists turned out to be a strategy to be further developed.

Keywords
FAIR Data, R&D Data Management, Materials Science, Tribology

1. Introduction

Grand scientific discoveries often require the incorporation of knowledge from a variety of research domains. In the age of data-driven science, the FAIR data principles emerge as a viable strategy to integrate experimental results on a community level, and tackle global challenges. The FAIR data principles [1] stand for Findability, Accessibility, Interoperability, and Reusability. The broad intention behind following these principles when managing research data is that it will enable scientists from across geographical and domain boundaries to come together and analyze their own results in a larger context. A key part of the FAIR data principles is that, by following them, data becomes machine operable, and thus, autonomous agents can find new insights in it.

Arguably, the biggest challenge standing in the way of adopting modern research data management practices in experimental sciences, such as the application of the FAIR data

SEMANTICS 2023 EU: 19th International Conference on Semantic Systems, September 20-22, 2023, Leipzig, Germany
Nikolay.Garabedian@kit.edu (N. Garabedian); Ilia.Bagov@kit.edu (I. Bagov); Malte.Flachmann@kit.edu (M. Flachmann); Nuoyao.Ye@student.kit.edu (N. Ye); Milosz.Meller@hereon.de (M. Meller); Floriane.Bresser@student.kit.edu (F. Bresser); Christian.Greiner@kit.edu (C. Greiner)
0000-0003-4049-4212 (N. Garabedian); 0000-0002-9094-8959 (I. Bagov); 0000-0002-0802-3480 (M. Flachmann); 0000-0001-8079-336X (C. Greiner)

principles, is the lack of software solutions which come "ready-to-use" and preloaded with the metadata templates that cover the various labs' use cases. Our group has already published two "proof-of-concept" papers [2, 3] which addressed some of the specific bridges needed to be made for our domain of tribology (the field that investigates friction, wear, and lubrication). As a step further, these "proof-of-concept" ideas need to be translated into daily practice. A cultural change in scientists' workflows require that any software for FAIR data generation, storage, and publication, must be as easy to interact with as possible, and should account for the specific peculiarities of their individual research practices and domain; for example, in the field of tribology 35% of the friction testers are custom and self-developed [4], which need to be described in a FAIR manner, in order to be cross-comparable. As such, the solutions described below enable a flexible framework for making applications that assist scientists with their research data management. In turn, because of this flexibility, a successful FAIR data framework in tribology and materials science is transferrable to many other experimental domains.

2. Methodology

The following components of our framework present one route in which FAIR data can be produced, stored, and published (Figure 1). The motivation for the creation of this framework comes from the needs of lab scientists whose workflows need to be digitalized. The flexible and open protocols easily allow for the incorporation of other applications that subscribe to the FAIR data initiative. The different components serve different purposes, namely: (1) VocPopuli is the metadata schema manager which provides templates to (2) ELNs; (3) data in the ELN is then paired with the VocPopuli vocabularies, and exported on Zenodo; (4) the communication between scientists and the ELN is enabled through a set of applications that accompany scientists in their various activities, either manual (e.g., specimen cleaning) or digital (e.g., data analysis). **A demonstration of part of the framework is available at:** https://youtu.be/lS8w-LwwGU4.

Figure 1. A visual overview of our FAIR data production pipeline. VocPopuli is the application for creation of FAIR vocabularies – these vocabularies can then be used by either our suite of FAIR-Save applications, or in conjunction with ELNs, for FAIR data publication in open repositories.

2.1. VocPopuli

The centrepiece of the framework is a software tool called VocPopuli, available at https://gitlab.com/metacook/vocpopuli. It enables the development of FAIR controlled vocabularies (FAIR assessment as per [5]) in a collaborative fashion. These vocabularies describe the experiments of interest for a given group, as well as all other equipment, processes, and data which pertain to them. A detailed description of VocPopuli's concept is offered in [6]. In short, as it is relevant to the overall framework of FAIR data collection, VocPopuli offers the following features:

- Login through a GitLab account, and user management via GitLab's user right scopes.
- The definition of contextual types such as *procedure, experiment, equipment, data*.
- The creation of hierarchy of terms (Figure 2).
- Each term can be defined through free text, synonyms, links to already existing terms.

- Each term is assigned a unique ID upon its creation. A finer ID scheme is used to distinguish different versions of the same term.
- For terms that describe quantities, constraints on their ranges and units can be placed.
- Multiple users can edit or comment on the terms collaboratively.
- The final approval of the terms and their inclusion in the lab vocabulary is managed by a user with administrative rights.
- The serialization of the vocabulary using SKOS and its publication.
- The assignment of Persistent Uniform Resource Locators (PURLs).

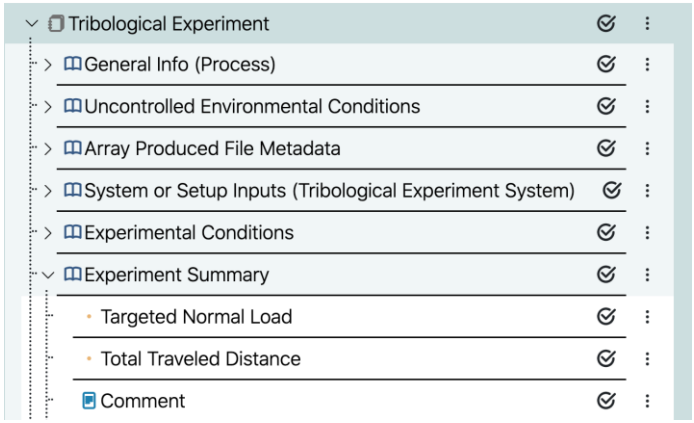

Figure 2. A tree of terms describing a tribological experiment in VocPopuli. Each of terms contain additional information. Once this schema is used within a lab database and populated with entries about a specific experiment, it can become a FAIR data artefact.

Afterwards, the vocabularies can be reused by the various other tools, developed as part of the presented work, and also published online, to be accessed by other interested parties. The motivation for this is that vocabularies that are composed by the scientists themselves contain the specific knowledge about a particular lab in the native way which they would express it in. A method for vocabulary sharing via GitLab and Zenodo is available, thus creating an ecosystem of vocabularies that can be examined by new users, and customized to a particular lab. The task of handling vocabulary and term versions across time and place is handled internally through the use of a graph database and externally through storage on GitLab. This makes these vocabularies a reliable starting point for the creation of individual user interfaces in other lab software.

2.2. FS-DigitalBook

The FAIR-Save (FS)-DigitalBook is a tablet-based application, which allows lab experimentalists to describe processes and objects relevant to their field of research in a FAIR fashion, without leaving the lab bench. The application utilizes the vocabularies developed with the help of VocPopuli as metadata schemata. Each process or object described using FS-DigitalBook must have been previously defined as part of a VocPopuli vocabulary. These schemata are used to create data input forms which can afterwards be filled with specific values, linked with files generated by the specific procedure at hand, and stored in a laboratory database. Additionally, FS-DigitalBook includes the option for taking pictures which accompany the manually-collected data and metadata from the scientists. All of the collected information is then linked with other already existing data artefacts, in order to form a linked data network. The infrastructure for this process is currently provided by the lab database, which the FAIR data sets are stored in.

2.3. FS-Validator

FS-Validator is used to inspect and verify the consistency between vocabularies and collected data. It serves as independent audit which checks whether the vocabularies from VocPopuli are

used correctly within the FAIR data generation process. The application connects to the lab database through an API, which offers users the option to correct and update their saved entries. The specific database used for this first version of FS-Validator is an ELN called Kadi4Mat [7].

2.4. FS-Analysis

FS-Analysis consists of Python-based MATLAB functions used in the processing of FAIR raw data. The problem this application tackles is the documentation of research data analyses. Once raw data is collected by experimentalists, usually, it goes through multiple iterations of reformatting, merging, cropping, and scaling, until a visual representation, for example, is produced and published. FS-Analysis downloads raw data and its metadata from the lab database, then lets users work with it in any way they wish, and finally uploads the processed data to the lab database. The key attribute of FS-Analysis is that it links the raw and processed data, while including any analysis scripts, if available.

2.5. FS-Instrument

FS-Instrument is a set of LabVIEW addons, which automate the export of FAIR data directly from experimental setups. The addons take the terms from existing vocabularies, automatically collect details about a running experiment, and eventually upload them to the lab database. The benefit of using these addons is that they do not add an extra step in the daily lab practices of scientists.

2.6. Electronic Lab Notebooks (ELNs)

ELNs can serve as the lab database and offer a collaborative and traceable platform for scientific data management. Our group is involved with lab procedures which currently include the use of Kadi4Mat [7] and Herbie (https://codebase.helmholtz.cloud/hereon-mb/herbie), but is open to others. Each of these solutions is tailored to serve the technical needs of scientists from specific domains. The goal of our applications for FAIR vocabularies is to supplement ELNs with traceable semantics for the data they contain. Once all data is collected, then the ELNs have functionality to package and export the data and metadata as interoperable packages, such as the RO-Crate [8].

2.7. Research Data Repositories

Once the data is ready for publication, it can be shared via platforms for hosting scientific data. In our case this is done through Zenodo (https://zenodo.org) for smaller files and metadata, and RADAR4KIT (https://radar.kit.edu) for larger files. The platforms offer the assignment of digital object identifiers (DOIs), which make the group of scientific data referenceable.

3. Conclusions

Traversing the trail of producing FAIR data from vocabulary creation to data publication from the ground up resulted in the development of the many tools described above. One of the outcomes of this is a FAIR data package of tribological results (https://doi.org/10.5281/zenodo.7923127) which contains 151,045 semantic triples and 412 GB of data. Keeping our local data in a FAIR way for this project has been instrumental in locating information quickly. This is essential when analyses that include information from multiple pieces of test equipment have to be performed. Figure 3 shows a snapshot of some of the data collected by our lab for one project. Having the traceability from any lab procedure to the specimen it has analyzed and its entire history let us gain insights about friction and oxidation of copper that have previously not been discovered.

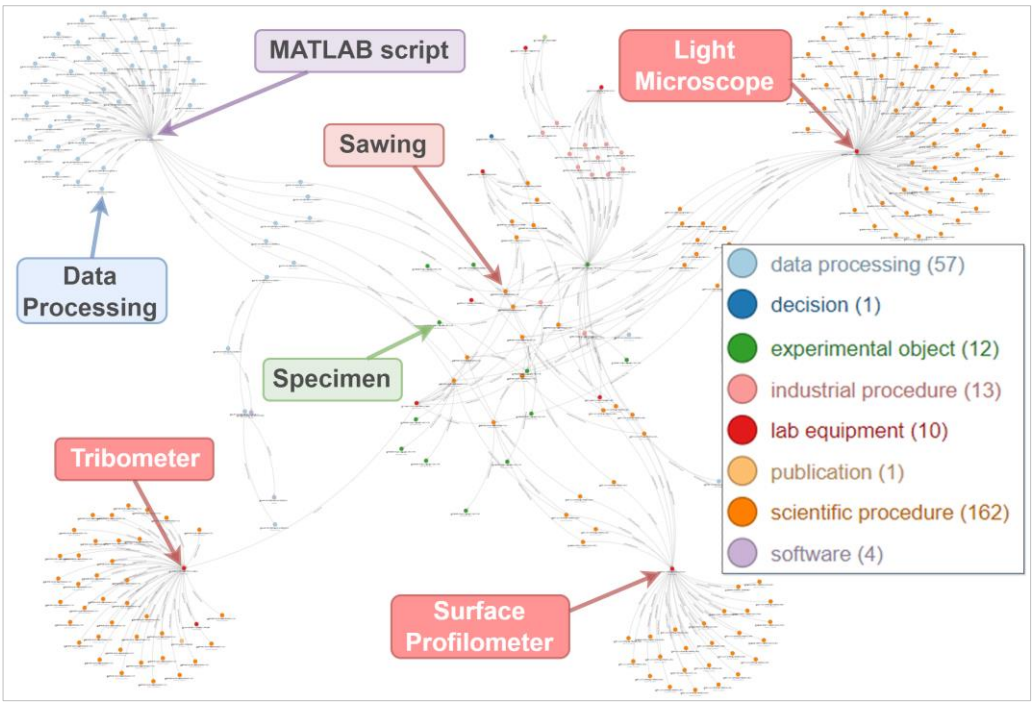

Figure 3. A visualization of a small part of the records in the ELN Kadi4Mat. This visual already shows that the data forms clusters around equipment and procedures that are often reused.

References

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. **3**, 160018 (2016)
2. Garabedian, N.T., Schreiber, P.J., Brandt, N., Zschumme, P., Blatter, I.L., Dollmann, A., Haug, C., Kümmel, D., Li, Y., Meyer, F., Morstein, C.E., Rau, J.S., Weber, M., Schneider, J., Gumbsch, P., Selzer, M., Greiner, C.: Generating FAIR research data in experimental tribology. Sci. Data. **9**, 315 (2022)
3. Brandt, N., Garabedian, N.T., Schoof, E., Schreiber, P.J., Zschumme, P., Greiner, C., Selzer, M.: Managing FAIR tribological data using Kadi4Mat. Data. **7**, 15 (2022)
4. Zhou, Y., Tian, Y., Meng, S., Zhang, S., Xing, X., Yang, Q., Li, D.: Open-source tribometer with high repeatability: Development and performance assessment. Tribol. Int. **184**, 108421 (2023)
5. Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., Marinescu, M.C.: Ten simple rules for making a vocabulary FAIR. PLoS Comput. Biol. **17**, 1–15 (2021)
6. Bagov, I., Greiner, C., Garabedian, N.: Collaborative Metadata Definition using Controlled Vocabularies, and Ontologies. Res. Ideas Outcomes. **8**, (2022)
7. Brandt, N., Griem, L., Herrmann, C., Schoof, E., Tosato, G., Zhao, Y., Zschumme, P., Selzer, M.: Kadi4mat: A research data infrastructure for materials science. Data Sci. J. **20**, 1–14 (2021)
8. Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L.J., Coppens, F., Fernández, J.M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P., Goble, C.: Packaging research artefacts with RO-Crate. Data Sci. **5**, 97–138 (2022)