# Standardized Metadata Collection to Reinforce Collaboration in Collaborative Research Centers

Manuel Watter, Laura Kahle, Birger Brunswiek, Urs A. Fichtner, Michelle Pfaffenlehner, Frank Werner, Denis Gebele, Harald Binder, Jochen Knaus

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg

The availability of good metadata in referenced terminologies is a prerequisite for data interoperability and the associated reliable retrieval. This interoperability of data through their documentation is considered one of the more complex problems in the creation of FAIR datasets (Jacobsen et al. 2020; Guizzardi 2020).

Standardization of data collection depends not only on the field of research, but also on the object of research: while excellent standards such as SNOMED CT[1] have been established in clinical trials and medical routine care, this is usually not the case in basic biomedical science using cell cultures or animal models. This is also reflected in the organization of large-scale research projects such as Collaborative Research Centers: in addition to highly standardized data types, such as for genetic analyses, there is also long-tail data with sometimes individual signatures. In both cases, however, there is a need for a standardized description of the experimental set-up.

As any documentation of datasets is labor-intensive, it is often only of medium-term benefit to the researcher. Therefore, the additional workload is more likely to be accepted if there are clear guidelines, e.g., from data repositories. If data documentation is to be incentivized instead of forced, a reduction in the effort required to collect the data is certainly a prerequisite.

In our bottom-up approach, scientists are empowered to define minimal datasets that are iteratively aligned with existing terminologies and standards by RDM managers.

## 1 Data documentation

General description standards such as DataCite (DataCite Metadata Working Group 2021) help to document datasets at an administrative level. Due to the lack of structured

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18089 

1 https://www.snomed.org

information from specific domains, this information is of limited use for further assessing the usability of a dataset.

We hypothesize that a "collage" of the useful parts of different standards and controlled vocabularies can keep the effort of collection low and thus increase adoption, without reducing the interoperability of the data sets for machine analysis too much.

## 2 Schematic integration of terminologies and ontologies

Transferring existing terminology can be difficult, even with a search function. Presenting exhaustive lists in a narrow use case can feel overwhelming and might waste a user's precious time. A complete hierarchy of possible tissue sources is not relevant to a cardiologist, for example.

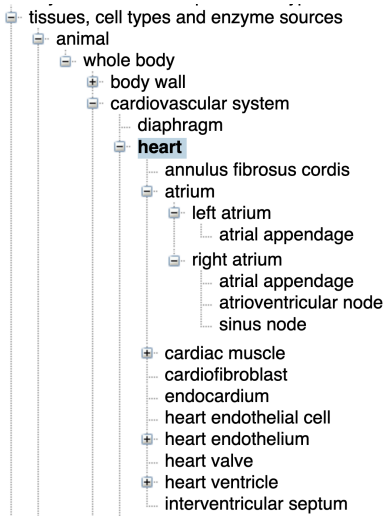


Figure 1: Example of a hierarchical vocabulary (Brenda Tissue Ontology BTO).

Accordingly, even when using terminologies, we propose a selection that is geared to the particular input case, covering it completely from a technical point of view, but reducing it to the minimally required areas (see Figure 1 for an example). Three strategies are potentially possible (Figure 2). Reducing the range of possible values is optimal for speeding up input without compromising the precision of the description and thus interoperability.

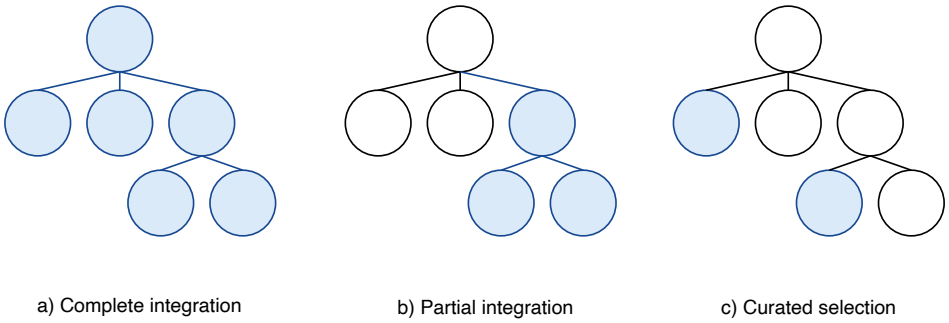


Figure 2: Options to integrate existing vocabularies such as taxonomies and ontologies.

While in Figure 2 a) in the entire terminology is browsable, b) only transfers a substructure, provided that suitable hierarchy levels or separation criteria are available. If only a few of the values ever occur in lab reality, a manually curated list (c)) is advantageous and can re-combine nodes scattered throughout the terminology.

# 3 Tools for scientists and data stewards

To enable scientists or data stewards to maintain data description structures themselves, a Microsoft Excel schema is provided (Figure 3). The familiarity with this tool lowers the barrier of entry. It is also particularly useful for locally specified lists that are typically already maintained in laboratories (e.g., antibodies, mouse lines), which can then be more easily compared to (potentially) existing standards during a subsequent revision step.

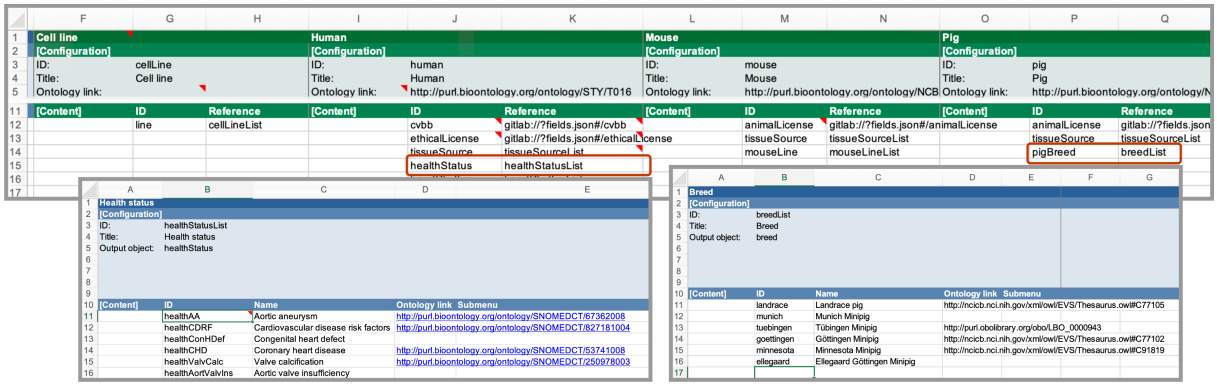


Figure 3: Example of defining structures and relationships of documentation entities using Excel.

# 4 Technical implementation

Our data documentation forms are embedded in our research data management system *fredato*, which is a thin wrapper over on-premises GitLab[2] and Nextcloud[3] instances. It stores the metadata directly in Git[4] repositories without a database, so they are always kept in sync with the research data and do not require explicit export processes, meaning no lock-in to our software and full user control.

The form definitions are also treated as data and exist as distributed JSON schema[5] definitions after being converted and merged from various sources (external vocabulary imports, local Excel lists, manual input) and displayed in the web frontend using the VJSF library[6] (see Figure 4). Once stored in their respective repositories, the metadata is automatically indexed in OpenSearch using GitLab Continuous Integration.

An example of processing a single aspect and the resulting internal representation is shown in Figure 5, an example of form logic defined in Excel is shown in Figure 6.

2 `https://about.gitlab.com`
3 `https://nextcloud.com`
4 `https://git-scm.com`
5 `https://json-schema.org`
6 `https://github.com/koumoul-dev/vuetify-jsonschema-form`; *Last accessed on May 5th, 2023.*

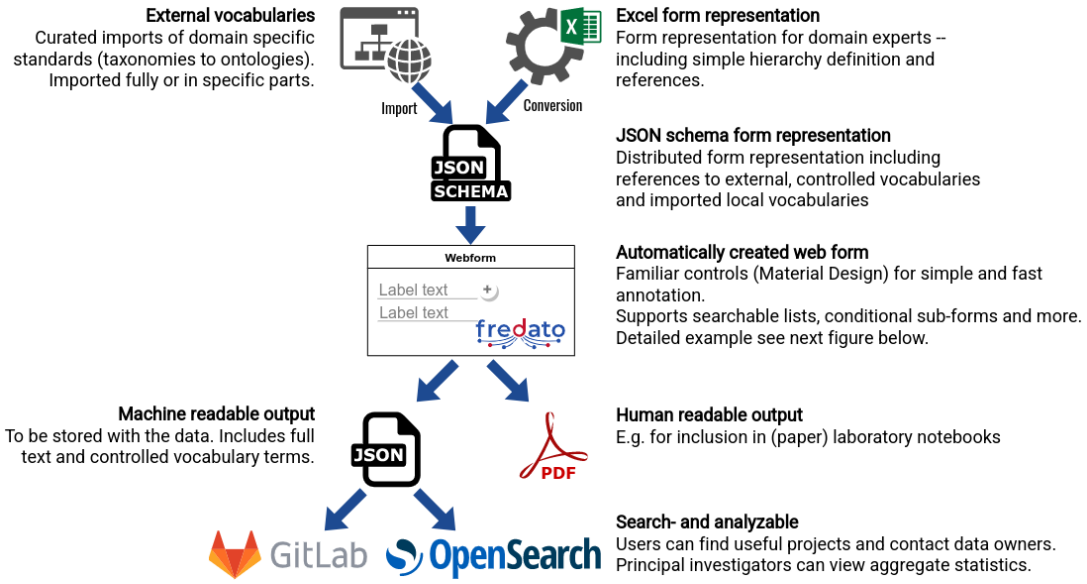


Figure 4: Workflow of form creation and processing using *fredato*.

# 5 Everyday use

Currently, for example, a template is available for the documentation of data sets in basic cardiological research, which was introduced as a recommendation in the Collaborative Research Centre 1425. In everyday life, there are two different procedures: Researchers use this to document the end of an experimental series and thus the generation of the raw data set in the laboratory. Alternatively, researchers, especially those who still work without an electronic lab book, use the metadata editor on a daily basis to document experimental progress. A copy function is available for this purpose, which only requires the new parts to be changed. A data set documentation is thus created from the compilation of the metadata of the individual laboratory days.

# 6 Discussion

When developing data documentation schemes in a bottom-up manner, it is advisable to include support from the research data management side in addition to the actual users, i.e. the subject experts. Both sides can benefit from each other, as knowledge of the need for reporting guidelines and data standards often needs to be built up by the subject experts. Ideally, candidates for local data stewards will emerge from this iterative process, greatly accelerating future collaborative efforts.

Our solution improves metadata interoperability, but does not produce fully machine-understandable grammars (Jacobsen et al. 2020). However, simply referencing published terminologies is usually not enough context for software agents to understand naming. The context can be re-created later in the export process by translating terminology look-ups into grammars.

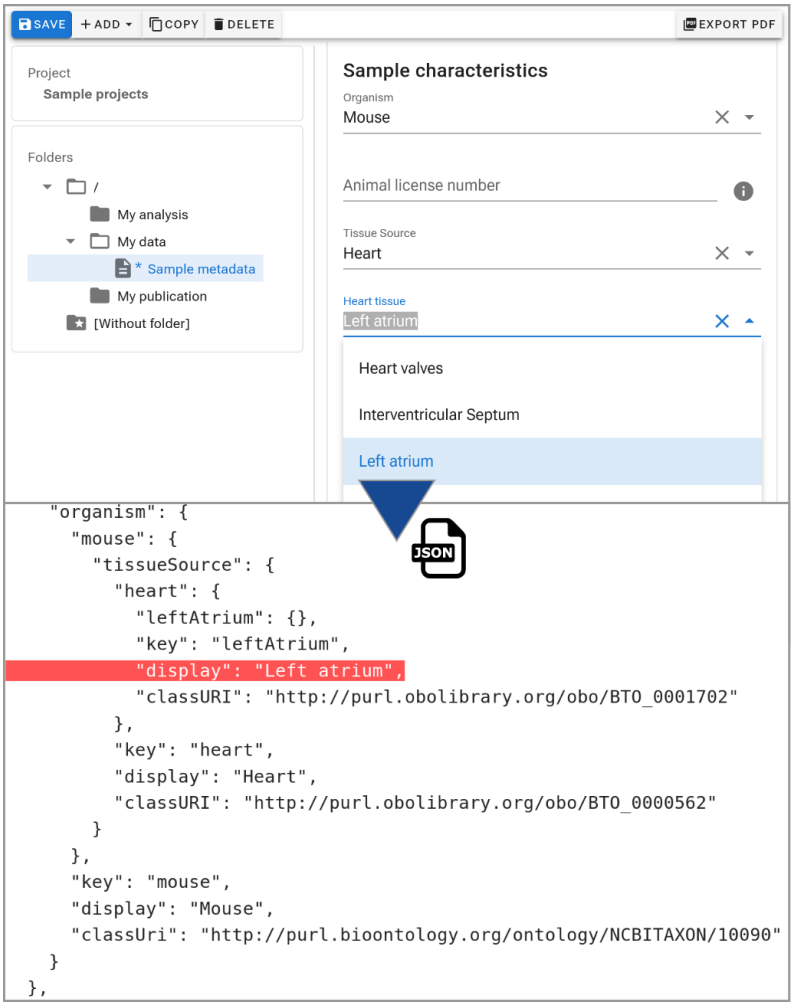


Figure 5: Web form and resulting JSON metadata file.

# 7 Conclusion

The burden of data documentation can be selectively reduced, without loss of technical interoperability, by presenting only the information from the terminologies that is necessary for a particular group based on standardized controlled vocabularies.

# 8 Author contributions

Manuel Watter developed the metadata editing and importing software and contributed to the original draft. Birger Brunswiek developed the metadata search software and added metadata indexing. Urs Fichtner and Michelle Pfaffenlehner contributed with writing - reviewing & editing. Denis Gebele, Laura Kahle and Frank Werner contributed to software testing. Harald Binder contributed with funding acquisition and monitoring. Jochen Knaus contributed to the conception, writing of the original draft and supervision.

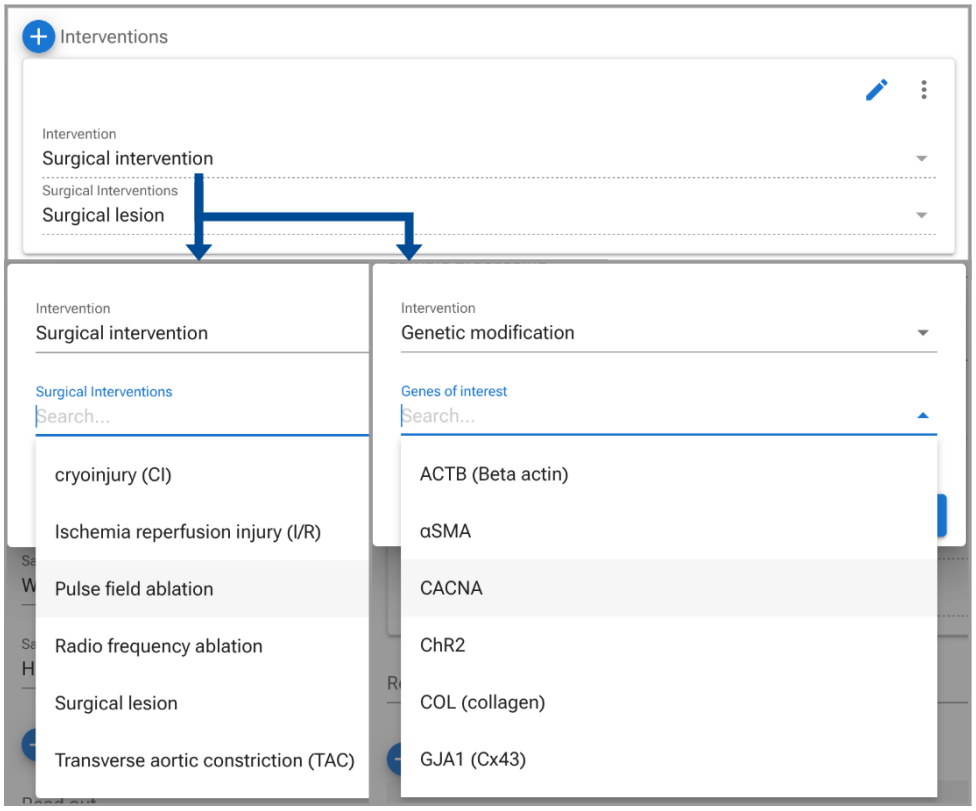


Figure 6: Example of a repeating field with conditional subfields.

## Acknowledgements

This work is funded by the Collaborative Research Centers 1425 (project number #42268-1845), 1453 NephGen (#431984000), 1479 OncoEscape (#441891347) and TR-CRC 359 PILOT (#491676693), all funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation).

## References

DataCite Metadata Working Group. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". DOI: `https://doi.org/10.14454/3W3Z-SA82`. `https://schema.datacite.org/meta/kernel-4.4/`.

Guizzardi, Giancarlo. 2020. "Ontology, Ontologies and the 'I' of FAIR". *Data Intelligence* 2 (1-2): 181–191. DOI: `https://doi.org/10.1162/dint_a_00040`.

Jacobsen, Annika, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, et al. 2020. "FAIR Principles: Interpretations

and Implementation Considerations". *Data Intelligence* 2 (1-2): 10–29. DOI: https://doi.org/10.1162/dint_r_00024.