Standardized metadata collection to reinforce collaboration in Collaborative Research Centers

Manuel Watter[1], Laura Kahle[1], Birger Brunswiek[1], Urs Fichtner[1], Michelle Pfaffenlehner[1], Frank Werner[1], Denis Gebele[1], Harald Binder[1], Jochen Knaus[1]

[1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Germany

An essential requirement for cross-project research data management is standardized data documentation that is accepted and actively used by the users. We propose a technical and organisational solution to integrate different standards and to lower the threshold of use without compromising the depth and accuracy of the documentation.

Introduction

Our institute currently supports four Collaborative Research Centers (CRCs) in medicine as infrastructure sub-projects. Although each CRC works on thematically related issues, there are major methodological differences, which are reflected in the data documentation requirements. The framework conditions are also heterogeneous: While there is a broad consensus for genetic data from acquisition to further processed formats, this is much less the case for imaging data and not at all for other experimental data. While excellent terminologies such as SNOMED CT exist for human data, this is not generally the case.
We try to improve data documentation ”bottom-up” by means of pilot projects and gradually generalise it in the CRCs. For this purpose, we create tools and accompany the process organisationally.

Data documentation

General description standards such as DublinCore or DataCite [1] help structuring datasets at administrative level [2]. Due to the lack of structured information from specific domains, this information is of limited use in further assessing the usability of a dataset. Minimal reporting guidelines started early [3] and data standards exist in some laboratory areas, but only partially cover the scientific process.
We hypothese 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.
In addition to the simple search for vocabulary terms, as offered in many systems, our approach also supports the structural embedding of substructures:

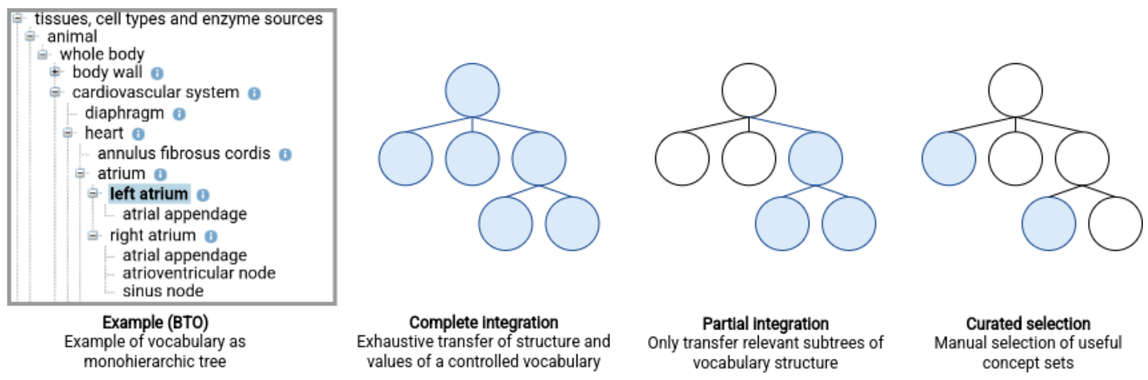

Figure: Integration choices of existing vocabularies like taxonomies and ontologies

Tools for scientists

To enable scientists to maintain data description structures themselves, an Excel input is provided. This is especially useful because local lists (e.g. antibodies, mouse lines) are usually managed in the laboratories, which can then be more easily compared with (potentially) existing standards during revision.

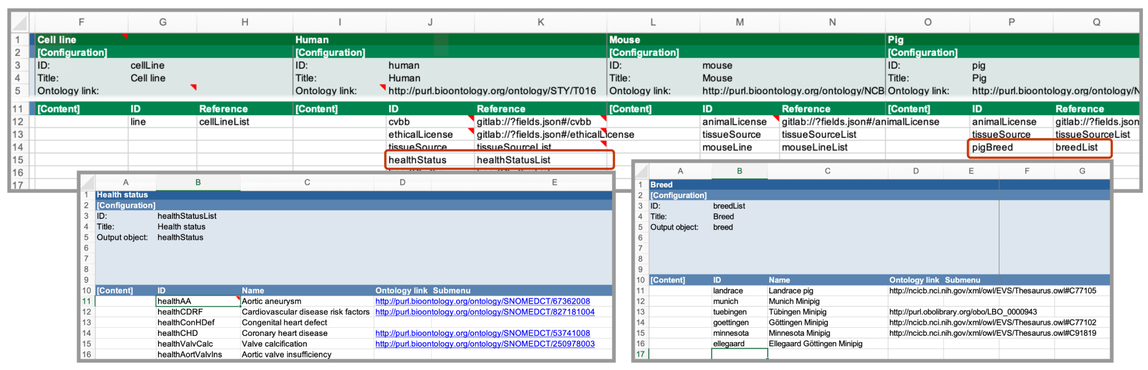

Figure: Example defining structures and relationships of documentation entities using Excel

Technical implementation

Our data documentation forms are embedded into our RDM system *fredato* and stored database-less directly in Gitlab repositories, so they are always kept with the research data and do not require explicit export processes. The user maintains full control.
The forms exist as distributed JSON schema definitions after being converted from various sources (external vocabularies, local Excel lists, manual input) and displayed in the web frontend using the VJSF library [4]. Once saved in the respective repositories, the metadata is indexed automatically in OpenSearch though Gitlab Continuous Integration.

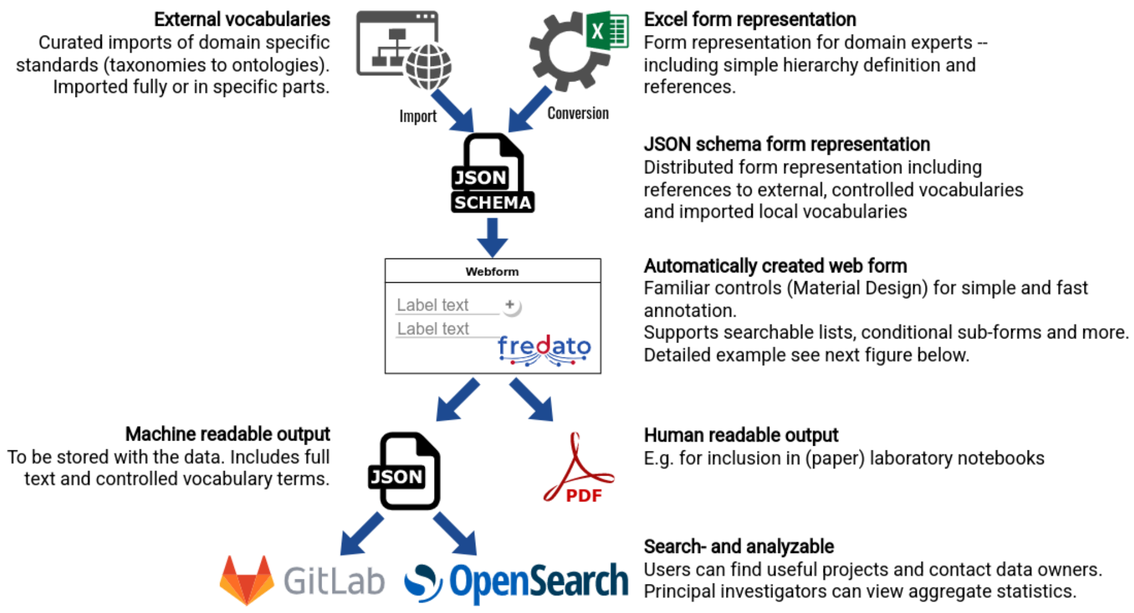

Figure: Workflow of form creation and processing using ”fredato”

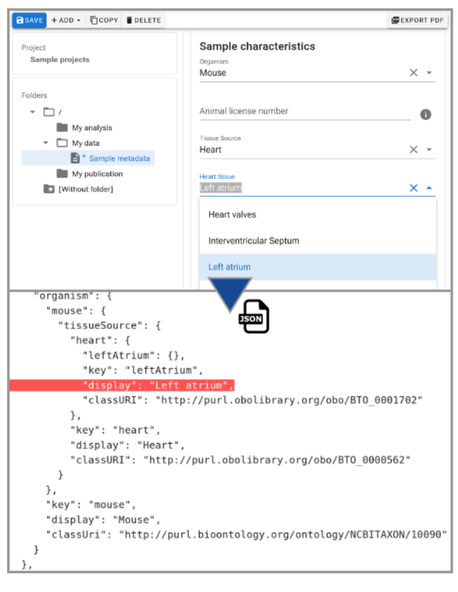

Figure: Example of web form and resulting JSON metadata file

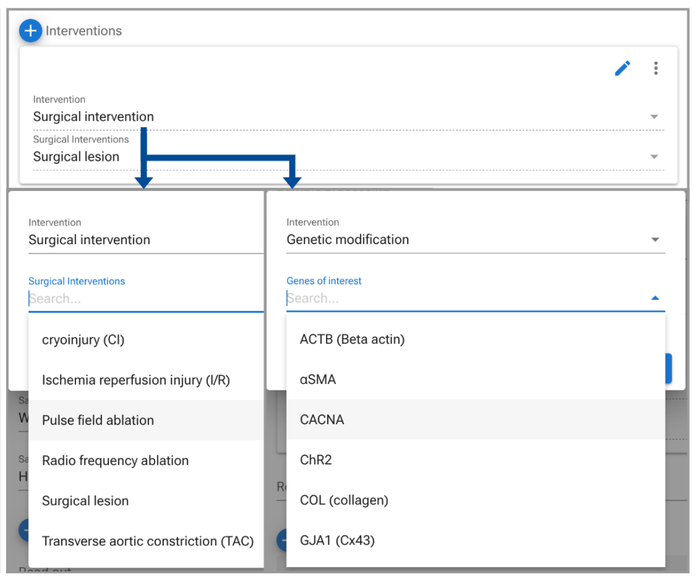

Figure: Example of a repetition field with conditional subfields

RDM organisation towards data stewards

When creating combined data documentation, it is advisable to include support from the research data management side in addition to the actual users and thus the subject experts. Both sides can benefit from each other, since the knowledge of the necessity of reporting guidelines and data standards on the part of the subject experts often has to be built up first. Ideally, candidates for local data stewards will emerge in this iterative process, tremendously speeding up future collaborative efforts.

Advantages to scientists

- For users:
  - Precise and curated documentation with relevant vocabulary instead of generic, overwhelming lists
  - Search and find CRC collaborations based on common terms
  - Intuitive modern web frontend and integrated with more tools to come
- For data stewards: Assisted process of combining RDM with lab reality
- For CRC managers: Central datasource for statistics and analytics

Acknowledgments

This work is funded by the CRCs 1425 ScarCare, CRC 1453 Nephgen, CRC 1479 Onkoescape and TR-CRC 359 PILOT, all funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Bibliography

[1] DataCite Metadata Working Group. “DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4”. In: (2021).

[2] Dan Phillips and Michael Smit. “Toward Best Practices for Unstructured Descriptions of Research Data”. In: *Proceedings of the Association for Information Science and Technology* 58.1 (2021), pp. 303–314. ISSN: 2373-9231, 2373-9231.

[3] Chris F Taylor et al. “Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project”. In: *Nature Biotechnology* 26.8 (2008), pp. 889–896.

[4] *Vuetify JSON schema form*. https://koumoul-dev.github.io/vuetify-jsonschema-form. Accessed: 2023-02-27.