A Semantic Representation of the NFDI4Health Metadata Schema Utilizing the CEDAR Workbench

Matthias Löbe [1], Aliaksandra Shutsko [2], Carsten O. Schmidt [3], Johannes Darms [2], Sophie A. I. Klopfenstein [4], Carina N. Vorisek [4], Xioaming Hu [5] , Martin Golebiewski [5] and Juliane Fluck [2]

[1] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
[2] *ZB MED - Information center for Live Sciences, Cologne, Germany*
[3] *Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany*
[4] *Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité, Berlin, Germany*
[5] *Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany*

Abstract

Rich metadata is required for a comprehensive description of research assets. We developed a metadata schema for clinical, epidemiological and public health research studies based on existing generic and domain-specific metadata vocabularies. It forms the basis for various search and data management services provided by the German National Research Data Infrastructure for Personal Health Data (NFDI4Health). Interoperability remains a challenge, as various health research standards are to be supported in the medium term. At the same time, embedding our infrastructure in national and international resources requires the use of overarching syntactic and semantic standards and vocabularies. In this paper we present a prototypical implementation in CEDAR Workbench. This not only provides a graphical web interface for collaboration and a possibility for form-based data entry for testing purposes. CEDAR also enables the use of standard vocabularies, annotation of concepts with medical terminologies, and a serialization in an RDF-JSON format.

Keywords

Metadata Schema, Clinical Trial Registries, Dublin Core, DataCite, DCAT, RDF, CEDAR, BioPortal

1. Introduction and Methods

NFDI4Health is an initiative to foster data sharing in the clinical and epidemiological research community in Germany. To improve findability and reusability of structured health data from clinical trials, epidemiological studies, disease registries, administrative health databases and public health surveillance, a metadata schema (N4H MDS) was developed unifying these different types of research studies particularly in the advent of COVID-19 [1]. The realization took place in Microsoft Excel, in order to be able to bring together the community quickly and without special software knowledge and not to anticipate any software-technical realization. Based on the experience gained, this schema will be further developed and also opened up for other types such as nutritional studies. However, due to the increasing complexity of the MDS, the large number of experts involved and the different domains of health research, the work is becoming increasingly difficult.

The CEDAR Workbench [3] is a web-based tool for the collaborative authoring of metadata schemas. It allows the creation of individual metadata elements including versioning and the reuse of high-quality templates that map to metadata standards such as Dublin Core or W3C DCAT. Metadata elements can be grouped and published in a form view. Completed forms are stored persistently and are therefore also suitable for prototypical tests before implementation in self-developed software.

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences
EMAIL: matthias.loebe@imise.uni-leipzig.de
ORCID: 0000-0002-2344-0426

2. Results

All metadata elements of the N4H MDS version 0.8 (equivalent to version 1.0 after final consensus) were implemented in CEDAR. We divided the implementation into 109 simple data elements that can be reused and 14 more complex structures composed of simple data elements[2]. This allowed the schema to be mapped completely. As in many technical implementations, certain idiosyncrasies of the software had to be anticipated, such as unusual types of form fields. Although the MDS follows established vocabularies such as Dublin Core, DataCite, or the specifications of international study registries, machine interpretation of such embeddings requires dedicated references, which the MDS currently does not provide. The annotation of data elements with ontologies from BioPortal was therefore only investigated prototypically, but appears very promising. It was not possible to fully map the complex conditional conditions under which certain data elements become mandatory fields, must satisfy certain formats, or should not be filled in at all. However, this would also not be possible with other tools like REDCap without programming effort.

3. Discussion and Outlook

Collaborative development of metadata vocabularies with domain experts has suffered for many years from limited support by intuitive tools. Mostly, the focus is on content work and the experts are not willing to learn new software tools in parallel. As a result, Microsoft Excel is still a quasi-standard, although its limitations in tracking changes, enforcing naming conventions and technical constraints, and implementation in software or APIs are well known. The use of CEDAR can also only partially resolve these conflicts. Further work should investigate the usability of the RDF serialization. Desirable would be a plugin mechanism that would allow syntactic compatibility to the RDF variant of the HL7 FHIR standard, as this is expected to play a major role in health research in the future [4].

4. Acknowledgements

This work supported by the DFG grant no. 442326535 and WI 1605/10-2.

5. References

[1] M. Golebiewski, M. Löbe, C.O. Schmidt, M. Lehne, A. Shutsko, and J. Darms, NFDI4Health Task Force COVID-19 Metadata Schema, FAIRDOMHub, 2021. doi: 10.15490/FAIRDOMHUB.1.DATAFILE.3972.1

[2] C.O. Schmidt, J. Darms, A. Shutsko, M. Löbe, R. Nagrani, B. Seifert, B. Lindstädt, M. Golebiewski, S. Koleva, T. Bender, C.R. Bauer, U. Sax, X. Hu, M. Lieser, V. Junker, S. Klopfenstein, A. Zeleke, D. Waltemath, I. Pigeot, and J. Fluck, Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies. Studies in health technology and informatics **281** (2021), 794–798. doi: 10.3233/SHTI210284

[3] R.S. Gonçalves, M.J. O'Connor, M. Martínez-Romero, A.L. Egyedi, D. Willrett, J. Graybeal, and M.A. Musen, The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. Semant Web ISWC 10588 (2017), 103–110. doi: 10.1007/978-3-319-68204-4_10

[4] S.A.I. Klopfenstein, C.N. Vorisek, A. Shutsko, M. Lehne, J. Sass, M. Löbe, C.O. Schmidt, and S. Thun, Fast Healthcare Interoperability Resources (FHIR) in a FAIR Metadata Registry for COVID-19 Research. Studies in health technology and informatics 287 (2021), 73–77. doi: 10.3233/SHTI210817

[2] Open view link: https://cedar.metadatacenter.org/instances/create/https://repo.metadatacenter.org/templates/6293220a-9f68-419e-9577-d055cea8ae93?folderId=https:%2F%2Frepo.metadatacenter.org%2Ffolders%2F2d0ab99f-01cb-4e74-bbe2-c5329fb77950