Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230074

An Annotation Workbench for Semantic Annotation of Data Collection Instruments

Julia SASSE[a] and Juliane Fluck[a,b,1]
[a] *ZB MED - Information Centre for Life Sciences* https://ror.org/0259fwx54
[b] *University of Bonn, Germany* https://ror.org/041nas322
ORCiD ID: Julia Sasse https://orcid.org/0000-0002-0660-7597,
Juliane Fluck https://orcid.org/ 0000-0003-1379-7023

Abstract. Semantic interoperability, i.e., the ability to automatically interpret the shared information in a meaningful way, is one of the most important requirements for data analysis of different sources. In the area of clinical and epidemiological studies, the target of the National Research Data Infrastructure for Personal Health Data (NFDI4Health), interoperability of data collection instruments such as case report forms (CRFs), data dictionaries and questionnaires is critical. Retrospective integration of semantic codes into study metadata at item-level is important, as ongoing or completed studies contain valuable information, which should be preserved. We present a first version of a Metadata Annotation Workbench to support annotators in dealing with a variety of complex terminologies and ontologies. User-driven development with users from the fields of nutritional epidemiology and chronic diseases ensured that the service fulfills the basic requirements for a semantic metadata annotation software for these NFDI4Health use cases. The web application can be accessed using a web browser and the source code of the software is available with an open-source MIT license.

Keywords. Interoperability, FAIR data, semantic metadata, metadata annotation

1. Introduction

The FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data management and stewardship [1] are developed to optimize data reuse. Operationalizing this guidelines increase the use of data beyond its original purpose. For all research data collected, data descriptions and information about the corresponding variables are essential for data analysis and reuse. Plenty semi-structured and structured study documents exist in the area of clinical and epidemiological studies, that are critical for metadata collection but that are not semantically annotated. Semantic interoperability can be realized by terminology- or ontology-based semantic annotation, in which item-level metadata is enriched with terminology standards and linked to unique semantic concepts e.g. to the SNOMED Clinical Terms collection of medical terms (SNOMED CT) [2] (Table 1).

Semantic annotation is still a challenge due to various standards and semantic richness of the data. Despite of the laborious annotation process, semantic annotation is largely done manually and annotators have to manage formats of study documents and a

[1] Corresponding Author: Juliane Fluck, E-mail: fluck@zbmed.de.

variety of complex terminologies and ontologies for annotation. Therefore, a semi-automatic approach to support researchers in semantic annotation and handling the different terminologies is desirable. Several metadata enrichment tools exist [3]. Either they are domain specific (e.g. the ODMedit annotation service [4] integrated into the Medical Data Models-Portal [5]) or they are terminology specific (e.g. the SNOMED International [2] Snap2Snomed annotation service [6]). Some metadata annotation services offer only prospective metadata annotation support (e.g. the CEDAR Workbench [7]), others are not open accessible at all.

NFDI4Health [8] aims to improve the FAIR access to structured health data originating from epidemiology studies, public health and clinical studies and support the harmonization of (meta-) data. In this context, an open accessible Metadata Annotation Workbench was developed to primarily support standardized metadata annotation in the NFDI4Health use cases but with the potential to find application in other domains.

Table 1. Linking of variables to SNOMED CT [9] semantic concepts.

Subject_id	Subject_label	Object_label	Object_id	IRI
p1_3	Age	Age (qualifier value)	397669002	http://snomed.info/id/397669002
q01	How old are you?	Age (qualifier value)	397669002	http://snomed.info/id/397669002
p1_4	Gender	Gender (observable entity)	263495000	http://snomed.info/id/263495000
q02	Sex	Gender (observable entity)	263495000	http://snomed.info/id/263495000

2. Methods

The Metadata Annotation Workbench is direct accessible via a web browser [10]. The source code of the service is shared via GitHub under an open-source MIT license [11]. Next to source code, prebuild container images [12] are available.

The application is developed with the microservice architecture pattern [13]. Main rationale for the architecture of the developed service is the reuse and integration of existing terminology and ontology services that already expose semantic concepts. This separations of concerns reduces the needed development effort. Additionally, it lays grounds for future adaptions. The overall software system is depicted in Figure 1. The application consists of four services, a user interface, the aforementioned terminology service, a parsing service to read and write variable files and last the semantic annotation service. All components are interconnected using REST [14] protocol. Data is persisted in a relational database (PostgreSQL). The UI was developed in the React framework [15] with the design library Elastic UI [16] for a uniform web layout. This web application consists of a landing page with the form for uploading a file and the annotation area with the search field and a semantic concept information view.

The first core microservice, the parsing service, is responsible for the conversion of input files to an internal format for exporting the resulting annotations. The import and export file format is the column wise oriented Microsoft Excel Spreadsheet. The datasets and annotations are stored in a database within the data layer. The second microservice, the annotation service, enables the annotation of a variable with a concept of a terminology resource by associating a term or question (subject_label) with the semantic label of the concept (object_label), the concept identifier (object_id) and the International Resource Identifier (IRI) based on the Simple Standard for Sharing Ontological Mappings (SSSOM) [17] (see Table 1). For the initial search suggestion of concepts, the variable is preprocessed for querying by simple natural language processing like tokenization and stemming. The search for concepts is performed via API requests to the terminology service API of the Semantic Lookup Service (SemLookP) [18]. This service is based on software developed by the EBI [19]: the Ontology Lookup Service (OLS) [20] and the mapping service Ontology Xref Service (OxO) [21]. SemLookP serves as single point of access to the latest ontology and terminology versions of NFDI4Health and allows to annotate research data in a FAIR manner by browsing the resources through the website as well as programmatically via an API.

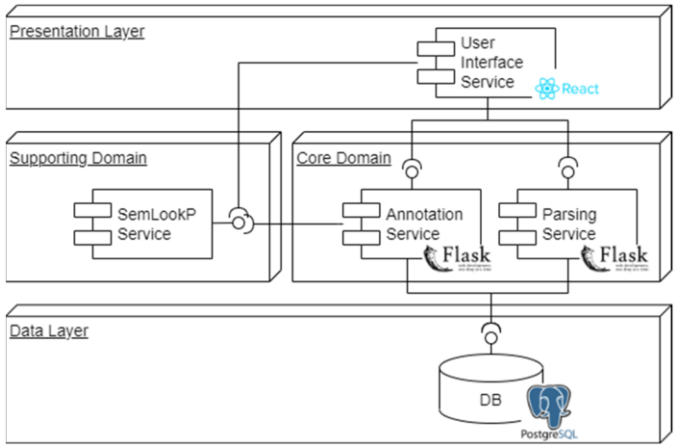

Figure 1. Metadata Annotation Workbench service interconnection.

3. Results

The web application has the following functionalities: data collection instruments can be uploaded in the convenient and common Microsoft Excel-Spreadsheet format, the user has to select a column for annotation and can select an ontology/terminology for annotation from all resources included in the terminology service SemLookP.

The metadata annotation is performed item-wise. In a first step, the user is provided initially with automatic annotation results. Detailed information about a concept is displayed in a semantic information widget that is provided by the terminology service. The user can accept the proposed annotation or modify the search term and perform a manual text search to find a matching concept. One or more concepts for the annotation of each item can be selected. Finally, the annotated instrument can then be downloaded comprising the data items and corresponding annotations.

Some terminologies such as SNOMED CT [2] or FoodEx2 [22] have terminology specific properties. SNOMED International classifies the concepts into domains indicated by a parenthetical notation at the end of a concept name. Furthermore, it supports post-coordinated expression: the required meaning is expressed by combining several concepts by logical rules. FoodEx2 uses concatenations of concepts and

additional terms (facets) describing properties and aspects of foods from various perspectives to add further detail to the information [22]. Based on user requirements of these use cases, the Metadata Annotation Workbench usability was improved. For example, annotation of a data item with multiple concepts is allowed, thereby supporting concatenation for FoodEx2 and a simple logical operation for SNOMED CT.

4. Discussion

We developed the Metadata Annotation Workbench in a user-driven development process to support standardized metadata annotation in order to foster the interoperability of data collection instruments for clinical, epidemiological and public health studies.

The terminologies for the Metadata Annotation Workbench were made available via the external terminology service SemLookP. This achieves outsourcing of the provision of the latest terminology resources and at the same time allows a replacement of the terminology service.

Several challenges for semantic metadata annotation were identified in 'nutritional epidemiology' and 'chronic diseases' use cases: the quality of the data dictionaries varies greatly and has a major influence of the automatic and human annotation. Also, the synonym content of the terminologies is important. A further challenge for ontology-based annotation is the heterogeneity of semantic annotation requirement: FoodEx2 semantic terms are often concatenations of several FoodEx2 concepts and for SNOMED CT, post-coordination might be required. The annotation service must consider terminology specifics while staying usable for various terminologies. Ontology specific functions also influences the usability for users with less expertise.

A number of necessary enhancements have been already identified:

Currently, the Metadata Annotation Workbench only supports Excel file-format as input. To enable integration of metadata annotation in workflows or data harmonization processes, other in- and output formats such as the Health Level Seven (HL7) Fast Health Interoperability Resources (FHIR) [23] format or the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) [24] should be provided. Further future steps are the integration of advanced AI methods to optimize the pre-annotation. Also, structured usability testing will follow to enable further improvement of the service.

5. Conclusion

The Metadata Annotation Workbench fulfills the main requirements to support semantic annotation: it provides an user interface and the integration of external terminology services. Independent of the Metadata Annotation Workbench, the quality of the variable catalogs, the range of synonyms and descriptions of the terminology/ontology and their complexity are decisive factors that influence the success of semantic annotation.

Acknowledgement: We would like to thank Franziska Jannasch (German Institute of Human Nutrition (DIfE)), Carolina Schwedhelm (Max Delbrück Center for Molecular Medicine (MDC)), Carina Vorisek and Sophie Klopfenstein (Berlin Institute of Health (BIH)) for their contributions during the requirement specification.

Funding: This work was done as part of the NFDI4Health Consortium. We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 442326535.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[2] SNOMED Home page. SNOMED https://www.snomed.org/ (accessed February 21, 2023).

[3] Sasse J, Darms J, Fluck J. Semantic Metadata Annotation Services in the Biomedical Domain—A Literature Review. Appl Sci 2022;12:796. https://doi.org/10.3390/app12020796.

[4] Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016;16:65. https://doi.org/10.1186/s12874-016-0164-9.

[5] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database J Biol Databases Curation 2016;2016:bav121. https://doi.org/10.1093/database/bav121.

[6] Snap2SNOMED https://snap.snomedtools.org/ (accessed February 21, 2023).

[7] Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, et al. The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. In: d'Amato C, Fernandez M, Tamma V, Lecue F, Cudré-Mauroux P, Sequeda J, et al., editors. Semantic Web – ISWC 2017, vol. 10588, Cham: Springer International Publishing; 2017, p. 103–10. https://doi.org/10.1007/978-3-319-68204-4_10.

[8] Home. NFDI4Health https://www.nfdi4health.de/ (accessed February 21, 2023).

[9] SNOMED CT - Home https://browser.ihtsdotools.org/? (accessed March 1, 2023).

[10] Metadata Annotation Workbench https://mda.nfdi4health.de/ (accessed February 21, 2023).

[11] GitHub - nfdi4health/metadata-annotation-workbench: A metadata annotation service to support standardised metadata annotation developed by T3.2. https://github.com/nfdi4health/metadata-annotation-workbench (accessed March 1, 2023).

[12] Open Container Initiative - Open Container Initiative https://opencontainers.org/ (accessed February 21, 2023).

[13] Microservices. MartinfowlerCom https://martinfowler.com/articles/microservices.html (accessed February 21, 2023).

[14] [PDF] Architectural Styles and the Design of Network-based Software Architectures"; Doctoral dissertation | Semantic Scholar https://www.semanticscholar.org/paper/Architectural-Styles-and-the-Design-of-Software-Fielding/49fc9782483bc311c2bd1b902dfb32bcd99ff2d3 (accessed February 21, 2023).

[15] React – A JavaScript library for building user interfaces https://reactjs.org/ (accessed February 21, 2023).

[16] Elastic UI https://elastic.github.io/eui (accessed February 21, 2023).

[17] Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database 2022;2022:baac035. https://doi.org/10.1093/database/baac035.

[18] Terminology Service < Semantic Lookup Platform < ZB MED https://semanticlookup.zbmed.de/ols/index (accessed February 21, 2023).

[19] Institute EB. EMBL-EBI homepage https://www.ebi.ac.uk/ (accessed February 21, 2023).

[20] Jupp S, Burdett T, Leroy C, Parkinson H. A new Ontology Lookup Service at EMBL-EBI. SWAT4LS, 2015.

[21] Jupp S, Liener T, Sarntivijai S, Vrousgou O, Burdett T, Parkinson HE. OxO - A Gravy of Ontology Mapping Extracts. In: Horridge M, Lord P, Warrender JD, editors. Proc. 8th Int. Conf. Biomed. Ontol. ICBO 2017 Newctle.--Tyne U. K. Sept. 13th - 15th 2017, vol. 2137, CEUR-WS.org; 2017.

[22] Authority (EFSA) EFS, Nikolic M, Ioannidou S. FoodEx2 maintenance 2021. EFSA Support Publ 2022;19:7220E. https://doi.org/10.2903/sp.efsa.2022.EN-7220.

[23] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst., 2013, p. 326–31. https://doi.org/10.1109/CBMS.2013.6627810.

[24] ODM-XML | CDISC https://www.cdisc.org/standards/data-exchange/odm (accessed February 21, 2023).