Title: Semantic Metadata Annotation Service

Abstract

Operationalising the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data management and stewardship [1] enhance the use of data beyond its original purpose. Data analysis of different sources for example requires interoperability that allows machines to automatically combine and process the data. Semantic interoperability, i.e., the ability to automatically interpret the shared information in a meaningful way, is given special attention in the biomedical domain. Its realization by ontology-based semantic annotation is still a challenge due to various standards and semantic richness of the data. Plenty semi-structured and structured study documents exist that are not yet semantically annotated. Despite of the laborious annotation process, semantic annotation is largely done manually and annotators have to manage data standards and formats as well as a variety of complex terminologies and ontologies for annotation. Therefore, a semi-automatic approach to support researchers in semantic annotation and handling the different data formats is desirable. In this context, the National Research Data Infrastructure for Personal Health Data (NFDI4Health) 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 [2]. For the latter, we present an approach for a cross-domain and extensible semantic metadata annotation service, that addresses the problems stated above. Basic requirements for the metadata annotation service are an open accessible web service, open code and an interface to integrate different terminology lookup services with access to terminology concepts depending on the use case. For the medical domain, support for the SNOMED terminology and common standards such as HL7 FHIR [3] are additional requirements. While some metadata annotation services already exist (e.g. [4]), they are in many cases tailored to specific data formats and terminologies or they do not meet all basic prerequisites. A first prototype was developed that meets the basic requirements and can be adapted to specific use cases. Terminology search via the Ontology Lookup Service [5] was implemented in an open-source web application to provide access to a wide range of terminologies. Usability tests with users indicated a good user experience. Currently, the annotation service is being further adapted for a use case in the medical domain. Therefore, the HL7 FHIR standard will be integrated and the SNOMED CT terminology will be used for semantic annotation of medical terms.

References

[1] Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 2016, 3, 160018, doi:10.1038/sdata.2016.18.

[2] Fluck, J.; Lindstädt, B.; Ahrens, W.; Beyan, O.; Buchner, B.; Darms, J.; Depping, R.; Dierkes, J.; Neuhausen, H.; Müller, W.; et al. NFDI4Health – Nationale

Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. *Bausteine Forschungsdatenmanagement* **2021**, [cited 2021 Oct 13] 72–85, doi:10.17192/bfdm.2021.2.8331.

[3] Bender, D.; Sartipi, K. HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange.; June 1 2013; pp. 326–331.

[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 Medical Research Methodology 2016, 16, 65, doi:10.1186/s12874-016-0164-9.

[5] Côté, R.; Jones, P.; Apweiler, R.; Hermjakob, H. The ONTOLOGY Lookup Service, a Lightweight Cross-Platform Tool for Controlled Vocabulary Queries. *BMC bioinformatics* **2006**, *7*, 97, doi:10.1186/1471-2105-7-97.