LUMC Clinical Ontology for Biomedical Research

Núria Queralt-Rosinach[1], César H. Bernabé[1], Qinqin Long[1], Rajaram Kaliyaperumal[1] and Marco Roos[1]

[1]Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands

Abstract

The application of ontologies to solve biomedical problems in hospitals is increasing. In the Leiden University Medical Centre (LUMC) physicians, clinical researchers, data managers and FAIR specialists are addressing the question of how to manage research data in the hospital to enable efficient research for patient care and treatment. We hypothesized to improve data integration based on machine readable ontology, i.e. improve the Interoperability aspect of FAIR. In this paper, we describe the development and evaluation of the LUMC Clinical Ontology for biomedical research in academic hospitals.

Keywords

Ontologies, FAIR, Patient Data, Biomedical Research

1. Motivation

The application of ontologies to solve biomedical problems in hospitals is increasing. They are applied for different uses: to build knowledge graphs for computational analysis, to perform knowledge-based data analytics such as semantic similarity or machine learning for patient diagnosis or rationalized treatments, and to implement ontology-based integrative medical informatics frameworks to handle health data for clinical research and personalized patient management. The COVID-19 outbreak has brought into focus the need for faster response in hospitals. For example, the accurate prediction of mortality is important during the pandemic for the management of intensive care units allowing 'smarter' healthcare. To facilitate efficient clinical research, health data needs to be Findable, Accessible, Interoperable and Reusable (FAIR). However, this pandemic highlighted the data management crisis in hospitals as a hindrance to efficient research, and the need for the application of the FAIR principles [1].

In the LUMC, a research hospital in Leiden, the Netherlands, physicians, clinical researchers, data managers and FAIR specialists are addressing the question of how to manage research data in the hospital to enable efficient analytics for patient management and treatment. Generated health data in the hospital is heterogeneous in nature, representation and storage. Monitoring data for hospitalized patients encompasses clinical observations, laboratory measurements, and various omics such as transcriptomics or metabolomics that need to be integrated for computational analysis. Ontologies are suitable to structure data for machine readability and data harmonization over computation. We hypothesized to improve data integration based on machine readable ontology, i.e. improve the Interoperability aspect of FAIR. We aimed at

ICBO'21: Symposium on the irreproducible science, June 15–18, 2021, Bozen-Bolzano, Italy
n.queralt_rosinach@lumc.nl (N. Queralt-Rosinach)

developing an application ontology to integrate different data in the hospital. Our approach is based on following knowledge-engineering best practices such as modularization, the reuse of common biomedical ontology terms, design patterns and data models to be as FAIR as possible. In this paper, we describe the development and evaluation of the LUMC Clinical Ontology or LCO for biomedical research in academic hospitals.

2. Method

2.1. Data Sources

A workflow to create synthetic datasets of cytokine laboratory measurements was developed to facilitate simulation of patient health data and avoid complications with the General Data Protection Regulation (GDPR) too early in development. Regulators and scientists are still adjusting to the European GDPR, a new data privacy and security regulation that is possibly the most rigorous in the world. Though it was drafted and passed by the European Union, it imposes obligations onto organizations anywhere, so long as they target or collect data related to people in the EU. This synthesized dataset contains basic information related to clinical observations, lab measurements and biosamples used per patient and time point. This data comes from information systems to manage health data within the LUMC.

2.2. Ontology Development

We developed an application ontology for data integration and management of FAIR research data in the hospital. It is represented in OWL 2, a Semantic Web W3C recommended standard to improve interoperability of patient data with the biomedical Semantic Web. We build the ontology using Protégé 5.5.0 [2], which is a free, open-source editor and framework for developing and maintaining ontologies.

2.3. Evaluation with Competency Questions

Following medical doctors' research questions, we defined a set of Competency Questions (CQs) to evaluate the ontology. These questions ranged from simple retrieval of metadata to more sophisticated queries to analyse correlations in data.

3. Results

3.1. Ontology Design

We developed an OWL ontology following the EJP RD core semantic model [3], which is used to represent common data elements in rare disease patient registries [4]. The model is based on a design pattern for measurements resulting from some process in the Semanticscience Integrated Ontology (SIO) [5]. It allows describing different types of measurements using a common structure. SIO provides a simple, upper-middle level, integrated structure of types and relations for rich descriptions of objects, processes and their attributes [6]. It also provides design patterns to help in defining a structure for common data types, and it is used in biomedical

Semantic Web projects such as DisGeNET [7]. We designed the ontology by modules, one for each semantic type: clinical observations, score calculations, lab measurements, and biosamples, and we created a semantic model for each module. Moreover, we also modelled the LUMC disease severity score, which is developed by clinical researchers in the LUMC to facilitate the study of correlations for prediction. Each model shares the same fundamental structure reusing the EJP RD core semantic model. We represent the ontology using the OWL Semantic Web standard and we used Protégé to build it. The ontology is online[1] and publicly available for reuse in GitHub[2].

3.2. Evaluation

We used the LUMC disease severity score phenotypes to answer medical doctors' questions such as 'what are the clinical parameters that can predict the disease course of a patient?'. We created a set of CQs to evaluate the ontology. Our evaluation results demonstrate that the ontology enables answering sophisticated questions. CQs and results are accessible at [8].

4. Discussion

Healthcare suffers from a *data silo* problem. In a typical healthcare-delivery organization such as a hospital, several different information systems are used for data management of different types: EHR (electronic health record), RIS (radiology information system), LIS (laboratory information system) and HIE (health information exchange) to mention just a few. This creates heterogeneity in clinical data structures (syntactic and semantic), and in using different tools and software systems. We created an OWL ontology to facilitate meaningful integration of different patient data within a hospital as well as to improve interoperability of clinical data for queries across external Linked Open Data and other clinical datasets from other hospitals on the Web. Beyond the scope of this paper is our work on applying ontologies (DCAT2 [9] and extensions thereof) to also make the metadata of clinical data containers machine readable, and thereby addressing F, A, and R principles of FAIR.

The ontology is designed to integrate four semantic types: clinical observations, score calculations, lab measurements and patient biosamples, and represented in OWL to enable interoperability of clinical data, data sharing and reasoning. We used knowledge-engineering good practices such as reuse of ontological terms, design patterns and modularization guidelines to increase interoperability with biomedical Linked Data. The ontology is publicly accessible at [10] and FAIRsharing[3].

The main challenge was to access clinical data for modelling the ontology. To preserve the privacy of patients in the hospital, GDPR enforces compliance with a rigorous regulatory framework. However, the intention of the GDPR is not to close off data, but to regulate access and make access rules transparent. Therefore, we recommend setting up a FAIR data governance policy for clinical data as soon as possible. FAIR facilitates a policy for making data as open

[1] https://lumc-biosemantics.github.io/beat-covid/docs/LUMC-Clinical-Ontology.html
[2] https://github.com/LUMC-BioSemantics/beat-covid/tree/master/fair-data-model/lumc-clinical-ontology
[3] https://fairsharing.org/bsg-s001616/

as possible and as closed as necessary [11]. A machine readable representation of access rules can be included with the metadata that describes the data resource, such as via a DCAT2-based FAIR Data Point[4]. With this ontology we demonstrated how to increase interoperability of clinical data, i.e. the "I" in the FAIR principles, independent of the imposed access rules. Other ontologies already exist in the clinical domain that are applied in similar contexts such as LOINC [5], SNOMED CT [6] and ICD [7]. They are used to represent observations, medical terms and diagnoses in clinical information models and constitute robust artifacts for data acquisition and retrieval of data associated with these terms. The problem is that they are ontology-like terminologies, i.e. term-centered and they represent linguistic entities and no semantic types. With the creation of LCO based on logically defined representations and on reusing the same biomedical design pattern to represent heterogeneous health data that can be applied in diverse contexts from clinical measurements in hospitals to elements in patient registries, we aim to support semantic interoperability, algorithmic reasoning and computable concepts for analysis in multicentric clinical and biomedical research projects. We next envision to create FAIR data 'in terms of' the ontology and build knowledge-based applications in the hospital for analysis and hypothesis generation.

Acknowledgments

N. Queralt-Rosinach, R. Kaliyaperumal, C. Bernabé, Q. Long, and M. Roos are supported by funding from the European Union's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575. We would also like to thank to the EJP RD, the GO FAIR VODAN, and the ZonMW Health Holland under the Trusted World of Corona, for supporting the research on FAIR data that was reused here. We would like to acknowledge that work in the BEAT-COVID project was partly funded by the Wake Up To Corona crowdfunding initiated by the Leiden University Fund (LUF).

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016).

[2] M. A. Musen, et al, The protégé project: A look back and a look forward, AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence 1 (2015). doi:10.1145/2557001.2757003.

[3] Ejp rd sio core model graph, 2021. URL: https://github.com/ejp-rd-vp/CDE-semantic-model/wiki/Core-model-SIO.

[4] R. Kaliyaperumal, M. D. Wilkinson, P. Alarcón Moreno, N. Benis, R. Cornet, B. dos Santos Vieira, M. Dumontier, C. H. Bernabé, A. Jacobsen, C. M. A. Le Cornec, M. P.

[4] https://github.com/FAIRDataTeam/FAIRDataPoint-Spec

[5] https://loinc.org/

[6] https://www.snomed.org/snomed-ct/five-step-briefing

[7] https://www.who.int/classifications/classification-of-diseases

Godoy, N. Queralt-Rosinach, L. J. Schultze Kool, M. A. Swertz, P. van Damme, K. J. van der Velde, N. van Lin, S. Zhang, M. Roos, Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data, medRxiv (2021). URL: https://www.medrxiv.org/content/early/2021/07/30/2021.07.27.21261169. doi:10.1101/2021.07.27.21261169. arXiv:https://www.medrxiv.org/content/early/2021/07/30/2021.07.27.21261169.full.pdf.

[5] Sio dp measurements homepage, 2014. URL: https://github.com/MaastrichtU-IDS/semanticscience/wiki/DP-Measurements.

[6] M. Dumontier, et al, The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014). doi:10.1186/2041-1480-5-14.

[7] J. Piñero, et al, The disgenet knowledge platform for disease genomics: 2019 update, Nucleic Acids Research 48 (2020) D845–D855. doi:10.1093/nar/gkz1021.

[8] The lco evaluation homepage, 2021. URL: https://github.com/LUMC-BioSemantics/beat-covid/tree/master/fair-data-model/lumc-clinical-ontology/competency-questions.

[9] DCAT2 W3C Homepage, https://www.w3.org/TR/vocab-dcat-2/, ???? Last accessed 2020/08/24.

[10] The lumc clinical ontology (lco) owl file, 2021. URL: https://github.com/LUMC-BioSemantics/beat-covid/blob/master/fair-data-model/lumc-clinical-ontology/owl/lco.owl.

[11] A. Landi, et al, The "a" of fair – as open as possible, as closed as necessary, Data Intelligence 2 (2020) 47–55. doi:10.1162/dint_a_00027.