Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220535

Discovery of Biomedical Databases Through Semantic Questioning

Arnaldo PEREIRA[a,1], João Rafael ALMEIDA[b], Rui Pedro LOPES[c], Alejandro PAZOS[b], and José Luís OLIVEIRA[a]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*
[c] *CeDRI, Polytechnic Institute of Bragança, Portugal*

Abstract. Many clinical studies are greatly dependent on an efficient identification of relevant datasets. This selection can be performed in existing health data catalogues, by searching for available metadata. The search process can be optimised through questioning-answering interfaces, to help researchers explore the available data present. However, when searching the distinct catalogues the lack of metadata harmonisation imposes a few bottlenecks. This paper presents a methodology to allow semantic search over several biomedical database catalogues, by extracting the information using a shared domain knowledge. The resulting pipeline allows the converted data to be published as FAIR endpoints, and it provides an end-user interface that accepts natural language questions.

Keywords. Clinical Studies, Health Data, Ontology, Semantic Questioning

1. Introduction

Over the last years, the secondary use of data from medical and biomedical practice gained attention from the health community, leading to new strategies aiming to enable access to clinical data from distributed databases without losing patient data privacy [1]. These strategies help researchers identify datasets of interest to design and conduct a multi-institutional study. One of the goals of the European Medical Information Framework (EMIF) (http://www.emif.eu) project was to improve the access of researchers to patient-level data from distinct health databases across Europe. In this project, EMIF Catalogue was developed to provide metadata information about each database to overview the database content without exposing the data.

Identifying databases of interest is one of the challenges researchers face in this platform. The main issue is the lack of a strategy to correlate concepts that are similar in distinct databases, for instance, procedures that can have different designations due to separate institutional policies. This work proposes a strategy to extract biomedical data through an ontology that guides the data conversion into a semantic format. To facilitate access to this semantic repository, we propose an interface that receives queries written in natural language.

[1] Corresponding author: E-mail: arnaldop@ua.pt

2. Methods and Discussion

The proposed methodology aims to convert questions written in the English language into formal queries based on question-answering templates. There is a clear benefit in simplifying the access to semantic data and saving time to end-users. For instance, it enables the evaluation of a research study's feasibility before spending time in the definition of a study protocol for a topic that may not be viable.

In this use case, we want to improve medical researchers' user experience when navigating the database catalogue by retrieving databases with free-text questions. This feature has some limitations due to the data scope. However, it can expand the usual match-based features available on this platform. This methodology was implemented as a plugin in the EMIF Catalogue, which requires the manual mapping of the catalogue concepts to the ontology concepts. This initial effort creates a new description layer for the catalogue entities, as well as, it also provides additional knowledge to each concept. For instance, with ontology is possible to have a relation between the entities, while in the original version, there is no relation.

3. Results and Conclusions

The proposed system was integrated into the EMIF Catalogue platform as a plugin which was validated in the study of Alzheimer's disease containing the metadata of 130 cohorts, of which 62 are public available. Each dataset is characterised by 472 entities. The searching features in this platform are provided through exact matching or through the use of boolean operation to build a nested query. With the inclusion of the proposed methodology, the search capabilities were expanded, allowing less restricted searches.

In a previous work, we used a recommender system to enable researchers to discover cohorts of interest [2]. Although these systems provide good results, they are not suitable for verifying studies' feasibility. The use of semantic queries over biomedical databases enables the exploration of data through a higher level of representation. However, to take benefit of this technology, it is required some technical background and deep knowledge of the ontology used in each scenario.

Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. AP and JRA are funded by the FCT - Foundation for Science and Technology (national funds) under the grants PD/BD/142877/2018 and SFRH/BD/147837/2019 respectively.

References

[1] Almeida JR, Fajarda O, et al. Strategies to access patient clinical data from distributed databases. In HEALTHINF, pages 466–473. SciTePress, 2019.

[2] Almeida JR, Monteiro E, et al. A recommender system to help discovering cohorts in rare diseases. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pages 25–28. IEEE, 2020.