2018 IEEE 14th International Conference on e-Science

Fast and Easy Mapping of Relational Data to RDF for Rapid Learning Health Care

Martine de Vos
Netherlands eScience Center
Amsterdam, The Netherlands
m.devos@esciencecenter.nl

Berend Weel
Netherlands eScience Center
Amsterdam, The Netherlands
b.weel@esciencecenter.nl

Adriënne M. Mendrik
Netherlands eScience Center
Amsterdam, The Netherlands
a.mendrik@esciencecenter.nl

Andre Dekker
Department of Radiation Oncology (MAASTRO)
GROW, Maastricht University
Maastricht, The Netherlands
andre.dekker@maastro.nl

Johan van Soest
Department of Radiation Oncology (MAASTRO)
GROW, Maastricht University
Maastricht, The Netherlands
johan.vansoest@maastro.nl

***Index Terms*—FAIR data, mapping, Rapid Learning Health Care, ontologies, validation**

I. Introduction

This abstract describes a system that supports the creation and validation of relational data to ontologies. The proposed system is developed for rapid learning health care (RLHC), i.e., an evidence based approach to train cancer prediction models on clinical care data that is stored in multiple networked hospitals. Since clinical care data cannot leave the hospital due to privacy issues, distributed (machine) learning can be used [1], [2], where the models are transferred, rather than the actual patient data. This requires clinical care data to be represented in a findable, accessible, inter-operable and reusable (FAIR) [3] manner. The proposed system uses an approach called Ontology Based Data Access (OBDA) to query the hospital data. To this end the relational data is mapped to a conceptual layer in the form of ontologies, i.e., shared vocabularies. These ontologies represent the meaning and values of the data while hiding the complicated structure of the original data sources. Currently, the creation of these mappings forms an obstacle in RLHC, as it requires considerable effort and knowledge from users, and validation of the mappings is difficult.

II. Approach

The proposed system (Figure 1) enables database administrators and clinical experts to create and validate mappings and is based on a few basic principles:

1) No expertise of semantic web technologies required
2) Evaluation using SQL based queries as a proxy for a gold standard
3) Different approaches for structure and terminology mappings

The complete tool suite, installation and an example dataset are available on GitHub [1].

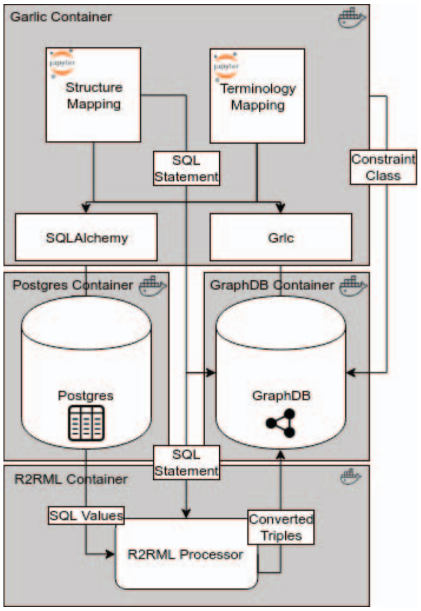

Fig. 1: Schematic overview of the technology stack.

III. Evaluation

The system is evaluated on a trial data set in two hospitals and during an international workshop. The proposed system shows high potential for efficient creation and evaluation of R2RML mappings of relational data to existing ontologies for Rapid Learning Health Care.

More extensive validation of the proposed system in other hospitals is required to ensure its applicability to a wide variety of database structures and content. We expect that there is overlap in the terminology used in different hospitals. Therefore it is expected that, by storing multiple terminology mappings in a common knowledge base, this knowledge could be reused as mapping suggestions for new hospitals.

[1] https://github.com/NLeSC/DataFAIRifier

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00110

REFERENCES

[1] A. Dekker, S. Vinod, L. Holloway, C. Oberije, A. George, G. Goozee, G. P. Delaney, P. Lambin, and D. Thwaites, “Rapid learning in practice : A lung cancer survival decision support system in routine patient care data,” *Radiotherapy and Oncology*, vol. 113, no. 1, pp. 47–53, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.radonc.2014.08.013

[2] T. Lustberg, M. Bailey, D. I. Thwaites, A. Miller, L. Holloway, E. R. Velazquez, F. Hoebers, and A. Dekker, “Implementation of a rapid learning platform : Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting,” *Oncotarget*, vol. 7, no. 24, 2016.

[3] M. D. Wilkinson et al., “Comment : The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific data*, vol. 3, pp. 1–9, 2016.