

LETTER TO THE EDITOR

Open Access



Overcoming challenges in rare disease registry integration using the semantic web - a clinical research perspective

Karl Gisslander^{1*} , Aladdin J Mohammad^{1,2}, Augusto Vaglio^{3,4} and Mark A Little^{5,6}

Abstract

The growing number of disease-specific patient registries for rare diseases has highlighted the need for registry interoperability and data linkage, leading to large-scale rare disease data integration projects using Semantic Web based solutions. These technologies may be difficult to grasp for rare disease experts, leading to limited involvement by domain expertise in the data integration process. Here, we propose a data integration framework starting from the perspective of the clinical researcher, allowing for purposeful rare disease registry integration driven by clinical research questions.

Keywords Rare disease, Semantic web, FAIRification, Data integration, Vasculitis, Registry

Main text

Disease-specific patient registries are essential for our increased understanding of rare diseases, but these registries remain fragmented, siloed, and lack data sharing mechanisms. Under the auspices of the European Joint Programme for Rare Diseases (EJP RD) and the European Reference Networks (ERNs), and guided by the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for scientific data management, several initiatives have been undertaken to tackle these challenges using Semantic Web based solutions [1, 2]. The Semantic Web

is an initiative to represent and publish data resources on the internet to facilitate data linkage and processing. By the extensive use of ontologies, a uniform vocabulary is created across resources, adding “semantic” meaning to the data. This allows for machine-readability of heterogeneous data and is the foundation of the FAIR initiative. However, to achieve a meaningful exchange across existing resources, such as rare-disease registries, there is a need for semantic harmonisation of the data-items to uniform ontology concepts. This transformation of registry data to a uniform vocabulary generates what we call the “semantic layer” of the registries. The FAIR-VASC project is one of the initiatives utilising Semantic Web technologies, enabling the federation of knowledge between multiple independent registries concerning the rare disease anti-neutrophil cytoplasmic antibody associated vasculitis (AAV) [3].

Through semantic data representation, integrated access and querying, Semantic Web technologies hold the potential for benchmarking of patient care and personalised medicine in diseases where traditional research suffers from small sample sizes. However, the concept of

*Correspondence:

Karl Gisslander

karl.gisslander@med.lu.se

¹Department of Clinical Sciences - Rheumatology, Lund University, Lund SE-221 85, Sweden

²Department of Medicine, University of Cambridge, Cambridge, UK

³Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy

⁴Nephrology and Dialysis Unit, Meyer Children's Hospital IRCCS, Florence, Italy

⁵TTMI, Trinity Health Kidney Centre, Dublin, Ireland

⁶ADAPT, Trinity College Dublin, Dublin, Ireland



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

the Semantic Web and its benefits are often difficult and time-consuming to grasp for rare disease researchers and clinicians, leading to limited involvement from domain expertise in the data integration process. This may lead to the development of a semantic layer not fit for the purpose of clinical research. We propose a data integration framework highlighting the need for clinical researcher involvement based on an iterative approach.

Central to the concept of the Semantic Web are ontologies, such as the Orphanet Rare Disease Ontology (ORDO), the Human Phenotype Ontology (HPO), SNOMED CT, Online Mendelian Inheritance in Man (OMIM) and, the Drug Ontology (DRON) enabling interoperability between resources through a common language [4]. In the siloed registries of today, the same concept may be represented in several ways. A translation step to allow for interoperability between the siloed registries is needed. This is the semantic layer, allowing for mapping of registry data to a shared vocabulary. However, exploring the rich environment of biomedical ontologies where obvious standards are lacking and finding the appropriate ontology terms, thus creating the semantic layer requires domain expertise. In FAIRVASC we tackled this through iterative cycles of data harmonisation, technical implementation, and query development driven by competency questions generated by domain expertise, building on existing methodology for ontology development [5–7].

First, a team of domain experts identify a competency question, a clinical research question to be answered through the semantic layer of the federated registries. This competency question is passed on to a team of registry experts with extensive knowledge of local registry data availability, semantics, and structure. In collaboration this team searches standard or widely adopted ontologies for appropriate concepts matching the registry data needed to answer the competency question. The identified ontology concepts, their matching registry variable names and the competency question are then transferred to a team of data managers or computer scientists for the technical implementation of the semantic layer. This technical implementation team are now tasked with two key objectives. First, they build the semantic layer by representing the objects and relationships of the registry data using the identified ontology concepts. Secondly, they build the query to retrieve the information needed to answer the competency question from the semantic layer of the federated registries. Following this cycle of activity, feedback from the implementation from all teams then informs the next iteration.

This easy-to-follow framework for the development of the semantic layer in rare disease registry integration actively involves the clinical researcher and is assuring data interoperability which is fit for the purpose of

clinical research. By employing research question driven semantic layer design, we minimise the exposed data, further strengthening the privacy-preserving mechanisms already in place in a federated approach to registry integration. The result is a semantic layer built for the needs of the clinical researcher as opposed to data integration for the sake of data integration. By proposing this easy-to-follow framework to ensure clinical researcher involvement in data integration projects, we hope to incentivise registry owners to share their data and unlock the full potential of the growing number of rare disease registries.

We, as rare disease researchers, see a growing number of data integration projects within the community and an extended implementation of FAIR. From our experience with the FAIRVASC project, successful rare disease registry integration and FAIRification using the Semantic Web requires extensive involvement from domain expertise, starting from research-driven objectives rather than implementation of data integration for the sake of data integration. We, therefore, here propose a framework for the development of the semantic layer needed for data integration in an iterative approach purposefully starting from the clinical research questions.

Acknowledgements

Not applicable.

Authors' contributions

KG wrote the first version of the manuscript which was then reviewed and revised by AJM, AV and MAL for intellectual content. All authors approved the final version to be published.

Funding

Vetenskapsrådet: 2019–00263.

The Crafoord Foundation: 20220623.

European Union's Horizon 2020 research and innovation programme under the EJP RD: COFUND-EJP N° 825575.

Data Availability

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 21 March 2023 / Accepted: 21 July 2023

Published online: 29 August 2023

References

1. Sernadela P, González-Castro L, Carta C, van der Horst E, Lopes P, Kaliyaperumal R, et al. Linked Registries: connecting Rare Diseases patient registries through a semantic web layer. *Biomed Res Int.* 2017;2017:8327980.
2. dos Santos Vieira B, Bernabé CH, Zhang S, Abaza H, Benis N, Câmara A, et al. Towards FAIRification of sensitive and fragmented rare disease patient data: challenges and solutions in european reference network registries. *Orphanet J Rare Dis.* 2022;17(1):436.
3. FAIRVASC – building registry interoperability to inform clinical care. <https://fairvasc.eu>. Accessed 20 March 2023.
4. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat.* 2012;33(5):803–8.
5. Noy NFM, Deborah L. Ontology Development 101: A Guide to Creating Your First Ontology. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf. Accessed 20 March 2023.
6. Peroni S. A simplified Agile Methodology for Ontology Development. OWL: Experiences and Directions; 2016.
7. McGlenn K, Rutherford MA, Gisslander K, Hederman L, Little MA, O'Sullivan D. FAIRVASC: a semantic web approach to rare disease registry integration. *Comput Biol Med.* 2022;145:105313.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.