Wikidata as an intuitive resource towards semantic data modeling in data FAIRification

Annika Jacobsen[1][0000−0003−4818−2360], Andra Waagmeester[2][0000−0001−9773−4008], Rajaram Kaliyaperumal[1][0000−0002−1215−167X], Gregory S. Stupp[3][0000−0002−0644−7212], Lynn M. Schriml[4][0000−0001−8910−9851], Mark Thompson[1][0000−0002−7633−1442], Andrew I. Su[3][0000−0002−9859−4104], and Marco Roos[1][0000−0002−8691−772X]

[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Micelio, Antwerp - Ekeren, Belgium
[3] The Scripps Research Institute, San Diego CA 92037, USA
[4] University of Maryland School of Medicine, Baltimore, MD, USA

Abstract. Data with a comprehensible structure and context is easier to reuse and integrate with other data. The guidelines for FAIR (Findable, Accessible, Interoperable, Reusable) data for humans and computers provide handles to transform data existing in silos into well connected knowledge graphs (linked data). Semantic data models are key in this transformation and describe the logical structure of the data and the relationships between the data entities. This description is provided through IRIs (Internationalized Resource Identifiers) which link to existing ontologies and controlled vocabularies. Creating a semantic data model is a labour-intensive process, which requires a solid understanding of the selected domains and the applicable ontologies. Moreover, in order to achieve a useful degree of Interoperability between datasets, either the datasets need to use the same (set of) ontologies, or the ontologies themselves need to be aligned and mapped. The former requires implementation of extensive (social) processes to achieve consensus, while the latter requires relatively advanced semantic engineering. We argue that this poses a significant obstacle for (otherwise capable) novice data modelers and even experienced data stewards.

Here, we propose that Wikidata can be used as an intuitive resource for resolvable IRIs both for teaching and studying semantic data modeling. In this way Wikidata serves as a hub in the linked data cloud connecting different but similar ontologies. We elaborate current problems and how Wikidata can be used to tackle these. As an example we describe two genetic variant models, one generated in a workshop and one generated using Wikidata. This shows how Wikidata can be instrumental in mapping similar concepts in different ontologies in a way that can benefit FAIR data stewardship processes in education and research.

Keywords: semantic modeling · FAIR · Wikidata · ontology · concept mapping

1 Opportunistic semantic data modeling

Data integration between heterogeneous data sets can be enabled by making them machine-readable, which formally captures their structure and context. One common way of generating such linked data is by using the combination of Resource Description Framework (RDF) triples and Internationalized Resource Identifiers (IRIs), the latter providing semantics to the data. In addition it is crucial that the context is made explicit through IRIs. This orchestration between IRIs can be captured in a semantic data model. Generating linked, interoperable data by using semantic modeling is central to making data FAIR (Findable, Accessible, Interoperable, and Reusable) for humans and computers [1]. However, creating a semantic data model is a laborious process. This process, which requires expertise both in the field under scrutiny and in ontologies and controlled vocabularies, is coordinated by a data steward. Further, given the relative novelty of the field, the availability of data stewards does not scale to the demand in the life sciences field.

To disseminate expertise of FAIRification, so-called Bring Your Own Data (BYOD) workshops have been conducted for the last five years, where FAIR and domain experts work together to FAIRify heterogeneous data [2]. A large part of this process consists of creating the underlying semantic data model. An example of such a model, generated in a BYOD held 6-8 June 2017 in Utrecht, The Netherlands [3, 4], can be seen in Figure S1. This model describes data that reflect measurements in samples from whole genome sequencing experiments, available in a variety of non-linked data formats. These models tend to be rather opportunistic in their onset, because the BYOD participants typically have diverse backgrounds. The different ontologies and controlled vocabularies are often cherry-picked based on the respective preferences of the participants and experts. Different resources exist to semantically express the same data even within the same domain: for example, OBO [5], Bioschemas [6] and SIO [7] serve partially overlapping and partially distinct areas of semantics in the life sciences. Therefore, initial BYOD models often use multiple namespaces, because it is difficult to dictate a clear guideline to select one single source for semantics. This is demonstrated by the large number of results per term in several state-of-the-art ontology search tools (Table S1). Even if the model was harmonized on a small set of ontologies and controlled vocabularies, the numbers mentioned in Table S1 suggest that different data modeling groups still would end up using different harmonized sets. This raises the question on how interoperable and reusable the resulting linked, FAIR data really is. For linked data that uses distinct sets of ontologies and vocabularies to be interoperable, it is essential to have mappings between their vocabulary terms and ontological concepts, otherwise the resulting linked data effectively remains a data silo.

In this paper, we propose that Wikidata [8] may be used as a source for IRIs and serve as a potential hub linking different opportunistic semantic data models both for education and research. Wikidata is a linked database contributed by both humans and machines. We first, describe an opportunistic semantic data

model of genetic variants generated in a BYOD and show how Wikidata can be used for data model construction and ontology mappings.

2 Semantic Data modeling with Wikidata IRIs

Wikidata is a linked database and a sister Wikimedia project of Wikipedia [8]. What Wikipedia is to text, Wikidata is to data: anyone, both humans and machines, can contribute to Wikidata as long as its primary source of the contribution is available under a public license. Wikidata has an RDF representation using Wikidata namespaces, which enables Wikidata concepts (items) to be embedded into RDF knowledge graphs. Issuing Wikidata items and statement values are open to all. On the other hand, properties link items in the Wikidata namespace (e.g. molecular function in Figure S2) are predefined and new properties need to go through a proposal process before they can be instantiated. Although Wikidata is not limited to a constrained set of domains, there are various active initiatives in the Biomedical domain to synchronize Wikidata with knowledge from authoritative biomedical resources [9, 10], such as the Disease Ontology [11] or Gene Ontology [12], where respective items have mappings to the original ontologies. Instead of having to sort through a wide variety of suggestions provided by the different ontology search tools, Wikidata can act as a single entry point for IRIs to create a semantic data model for data FAIRification. Wikidata provides three different ways that make it a viable source for IRIs to be used in initial steps of transforming unstructured research data into FAIR data. Firstly, the various (language) labels and descriptions associated with an item may be modified or extended by Wikidata users to enrich the definition of an item. The direct link to related wikipedia articles helps to disambiguate items so as not to (unintentionally) change the intended semantics of an item. Secondly, one could use the wikidata items with mappings to existing ontologies. Finally, one could choose to mint a new Wikidata item to reflect a specific concept that does not exist yet as a wikidata item.

Using Wikidata properties and items we expressed a semantic data model for genetic variants (see Figure 1). This is the same use case as illustrated in Figure S1, but here only Wikidata was used as a controlled vocabulary, and we added the identified IRIs from the used ontologies as mappings to this wikidata model, thus increasing the potential for semantic interoperability. Creating mappings was easily achieved using the Wikidata community edit interface by attaching a wikidata exact match property to the relevant items to connect them to external ontology IRIs. If one group's opportunistic model used for example an OBO ontology instead of the SIO ontology used by a second group, then data integration may be challenging. However, since both of those terms can be reconciled through Wikidata using the available mapping properties, the introduction of Wikidata facilitates automated or semi-automated harmonization of independently-authored opportunistic models.

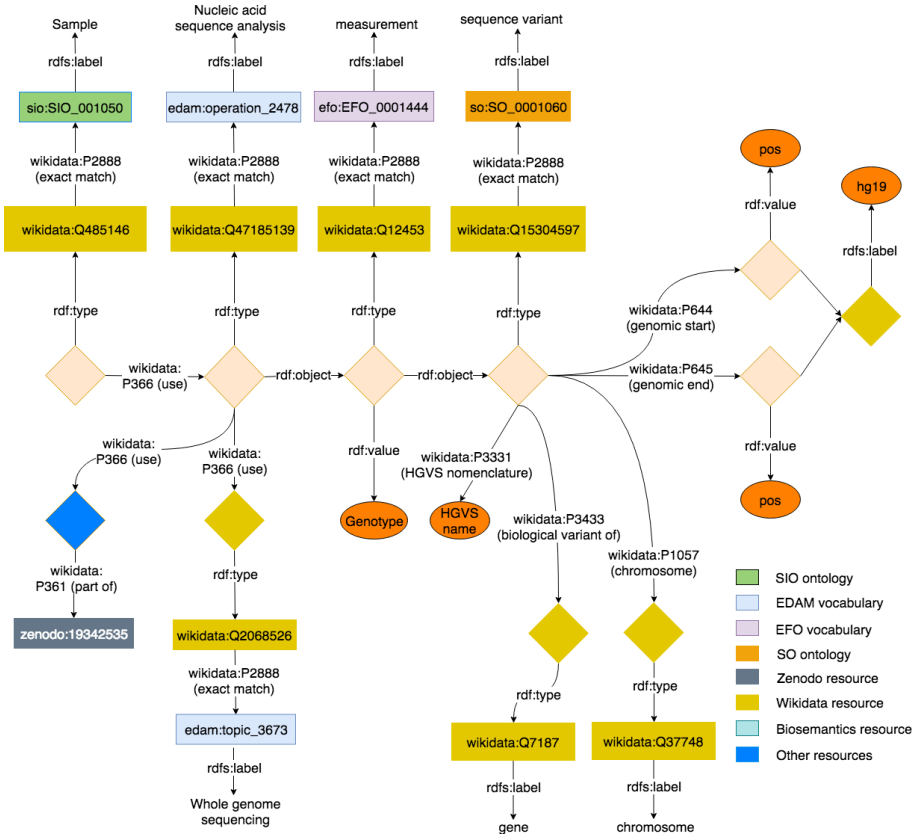

Fig. 1. Semantic data model of a genetic variant using Wikidata.

3 Conclusion and future work

In this paper, we put forward the position that Wikidata is a well-suited controlled vocabulary for data FAIRification in the life sciences, especially in initial stages, where it is either too early to adhere to specific domain descriptions, or specific data or project constraints and requirements are yet to emerge. As said, this leads to opportunistic data models where parts of different ontologies are cherry picked. By using Wikidata namespaces, this decision can be postponed or (partially) obviated. Once applicable ontologies become apparent, one could update Wikidata with mappings, thus linking the Wikidata namespace to external ones. For many data managers and data stewards, linked data approaches and technologies have a relatively steep learning curve and dealing with the wide variety of available ontologies is a major contributing factor. We argue that Wikidata is a good starting point to create initial models while maintaining the flexibility to evolve those models in the future. Additionally, it is easy to add new concepts and to do mappings using Wikidata so that resolvable IRIs

are instantly available for representing data as linked data. In comparison, extending other commonly used ontologies requires engaging with their curators or maintainers, which is not always possible or easy.

In conclusion, we argue that Wikidata is a viable source for IRIs in the process making data FAIR. Wikidata is open to everyone to add terms, properties and mappings to external ontologies. This together with the fact that every Wikidata items has a resolvable IRI, makes data using Wikidata items as IRI and its properties - also with IRIs - interoperable. Wikidata is useful resource in any FAIRification process. As part of future work, we would like to investigate how different semantic data models representing the same data compare to each other, what their respective limitations are and how Wikidata can be used to map from one to the other in a more extensive interoperability use case. We plan to do such a modelling exercise in the foreseeable future and welcome collaborations in doing so.

References

1. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016, 3:160018. doi: 10.1038/sdata.2016.18.
2. Bring Your Own Data Workshops, `http://www.dtls.nl/fair-data/byod/`, Last accessed 1 Oct 2018
3. UBEC FAIR datapoint, `http://www.ubec.nl/data/fair-data-point/`, Last accessed 1 Oct 2018
4. Bring Your Own Data Workshop - OncoXL, `https://www.dtls.nl/wp-content/uploads/2017/09/BYOD-OncoXL-June-2017-report.pdf`, Last accessed 1 Oct 2018
5. Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnology. 2007, 25, 1251-5. doi: 10.1038/nbt1346
6. Franck Michel and The Bioschemas Community. Bioschemas and Schema.org: a Lightweight Semantic Layer for Life Sciences Websites. Biodiversity Information Science and Standards. 2018, 2, e25836. Doi: 10.3897/biss.2.25836
7. Dumontier M, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics. 2014, 5, 14. doi: 10.1186/2041-1480-5-14.
8. Vrandečić D. and Krötzsch M. Wikidata: A Free Collaborative Knowledgebase. Commun ACM. 2014, 57, 78-85.
9. Burgstaller-Muehlbacher S, et al. Wikidata as a semantic framework for the Gene Wiki initiative. Database (Oxford). 2016, pii, baw015. doi: 10.1093/database/baw015.
10. Mitraka E, et al. Wikidata: A platform for data integration and dissemination for the life sciences and beyond. bioRxiv. 2015. doi: 10.1101/031971.
11. Kibbe WA, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015, 43(Database issue), D1071-8. doi: 10.1093/nar/gku1011.
12. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015, 43(Database issue), D1049-56. doi: 10.1093/nar/gku1179.

Wikidata as an intuitive resource towards semantic data modeling in data FAIRification

Supplementaries

Annika Jacobsen[1][0000−0003−4818−2360], Andra Waagmeester[2][0000−0001−9773−4008], Rajaram Kaliyaperumal[1][0000−0002−1215−167X], Gregory S. Stupp[3][0000−0002−0644−7212], Lynn M. Schriml[4][0000−0001−8910−9851], Mark Thompson[1][0000−0002−7633−1442], Andrew I. Su[3][0000−0002−9859−4104], and Marco Roos[1][0000−0002−8691−772X]

[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Micelio, Antwerp - Ekeren, Belgium
[3] The Scripps Research Institute, San Diego CA 92037, USA
[4] University of Maryland School of Medicine, Baltimore, MD, USA

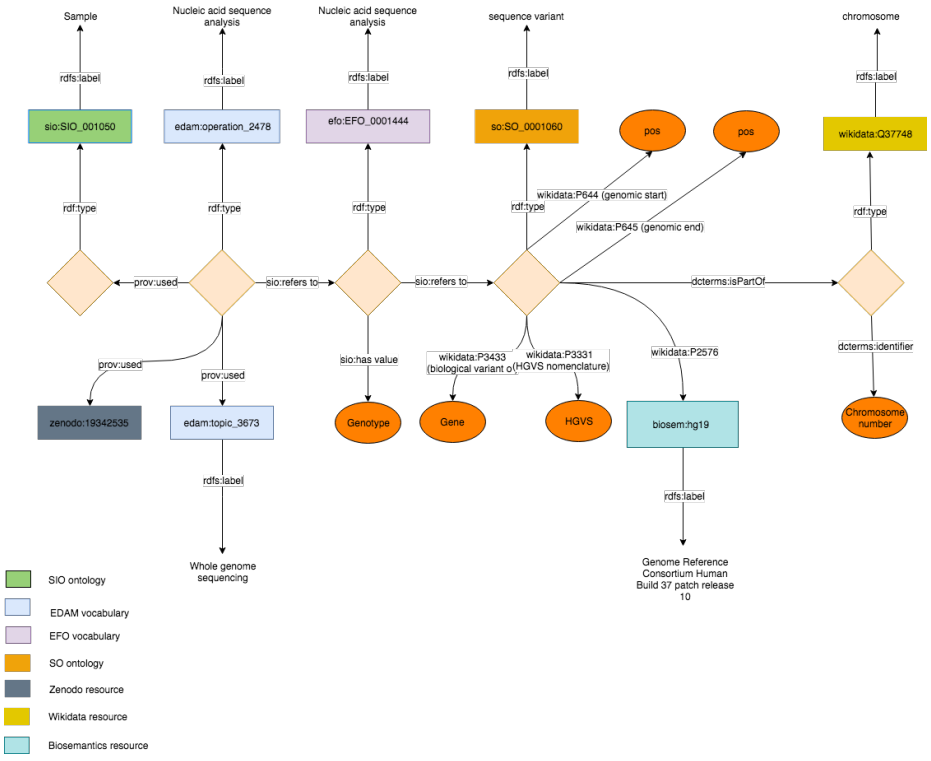

Figure S1 'Opportunistic' semantic data model of a genetic variant constructed during a 'Bring Your Own Data' (BYOD) workshop.

Table S1 Terms used to describe the data reflecting measurements in samples from whole genome sequencing experiments and number of results in EBI OLS (`www.ebi.ac.uk/ols/index`), BioPortal (`bioportal.bioontology.org/`) and Linked open vocabularies (`lov.linkeddata.es/dataset/lov`), respectively.

Term	EBI OLS	Bioportal	Linked open vocabularies
Sample	8	25	148
Whole genome sequencing	17	29	276
Protein sequence analysis	2	31	6928
Measurement	5	47	1066
Sequence variant	3	5	3759

Figure S2 A Wikidata item, with its descriptions, statements and site links.