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


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

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

Supplementaries

Annika Jacobsen[0000—0003—4818—2360], Andra Waagmeester[0000—0001—9773—4008], Rajaram Kaliyaperumal[0000—0002—1215—167X], Gregory S. Stupp[0000—0002—6644—7212], Lynn M. Schriml[0000—0001—8910—9851], Mark Thompson[0000—0002—7633—1442], Andrew I. Su[0000—0002—9859—4104], and Marco Roos[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.

<table>
<thead>
<tr>
<th>Term</th>
<th>EBI OLS</th>
<th>Bioportal</th>
<th>Linked open vocabularies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>8</td>
<td>25</td>
<td>148</td>
</tr>
<tr>
<td>Whole genome sequencing</td>
<td>17</td>
<td>29</td>
<td>276</td>
</tr>
<tr>
<td>Protein sequence analysis</td>
<td>2</td>
<td>31</td>
<td>6928</td>
</tr>
<tr>
<td>Measurement</td>
<td>5</td>
<td>47</td>
<td>1066</td>
</tr>
<tr>
<td>Sequence variant</td>
<td>3</td>
<td>5</td>
<td>3759</td>
</tr>
</tbody>
</table>

Figure S2 A Wikidata item, with its descriptions, statements and site links.
Reviewer 1

The paper explores how Wikidata can be used as a controlled vocabulary for life sciences in the process of data FAIRification.

We thank the reviewer for their comments. We would like to point out that the manuscript was initially submitted as a long research paper, but has now been re-submitted as a short position paper.

The authors claim that using multiple sources (ontologies) for URIs can result in data models that are difficult to integrate which is a very valid argument especially when the underlying ontologies are not mapped to each other. Based on that they present Wikidata as a better alternative but do not prove the efficiency and coverage of Wikidata on the considered data.

This is a very valid point, but we need to point out that in our manuscript we argue that Wikidata is a good alternative for teaching and research purposes (not necessarily in all general usage) as it is an intuitive resource for resolvable IRIs.

In our manuscript we focus on two problems where we could use Wikidata, 1) the expertise required to properly select ontologies does not scale with the need and the learning curve is very steep, and 2) different curators often select different ontologies for the same knowledge, which not always are mapped. For the first problem, we propose that Wikidata should be used in bring your own data sessions as a more simple step when learning how to use ontologies in semantic data modeling. For the second problem, we propose that Wikidata can be used to study mappings between ontologies into details. We have now made the goal, problems and solutions more clear throughout the manuscript.
The proposed model based on Wikidata is applied to a very small example about genetic variants. The authors do not mention how this can be extended to additional datasets and which coverage can be attained.

The authors mention in the abstract that the paper explores how Wikidata can be used in mapping similar concepts in different ontologies but the paper do not describe this process except for adding the links to the ontologies to Wikidata. This process is not addressed in the paper.

Overall the paper is well written but contains many parts describing Wikidata and the FAIR principles instead of concentrating on the actual contribution. No evaluation of the proposed approach is presented and the presented model is not sufficient in my opinion to validate the solution.

The authors do not compare their approach to existing ones and no state of the art is presented in the paper.

These are all valid points and do need to be addressed in a future research paper (which we are actually currently drafting) on this topic. On the last point, we have now made a clear invitation in our paper for others to join us in this experiment. Please note that we have re-submitted our manuscript as a position paper (see details above).

Reviewer 2

The authors describe an approach using Wikidata as possible central data source for semantic modeling. The paper has some major drawbacks. First, I can not read any of the figures. The descriptions used in these figures are so small making it impossible to read.

We thank the reviewer for their comments. We would like to point out that the manuscript was initially submitted as a long research paper, but has now been re-submitted as a short position paper. To this end only one figure remains in the paper, where the text in the figure has been substantially increased.

Second, ontologies and other vocabularies which have been extensively discussed (and used) in the life science community are an essential part for semantic modeling. They should be reused as much as possible instead of using new concepts created by own (which sounds more like reinventing the wheel). I don’t see the point that there are so many ontologies available that extensively overlap. In cases there is an overlap, the ontology concepts can also be linked (using a 'same-as' relationship).

Wikidata is a knowledge stack that comes with both a large (global) community and an built-in infrastructure that facilitates communications. Contrast to more traditional semantic model drafting processes, modelling data in wikidata is primarily a community effort. To mint a new property, it is required to convince the larger community, versus simply minting your own. In this communication process, reuse of existing properties is core. The process is not yet waterthight, but already in state to leverage the community it brings.
Third, I’m not really sure that a central hub will be a valuable solution. In other domains, decentralization is much more successful than generating and managing a large monolith. In that way, I would argue for reusing the available ontologies and in cases there is something missing adding new concepts and relationships or completely new ontologies if there is no one available.

We also argue for reuse of available ontologies. However, often the choice of ontologies to reuse is large. More ontologies exist to describe the same concepts. Currently, a change in one ontology remains with the ontology designers and these changes are not propagated with other similar ontology. We argue that Wikidata could act a communication platform where propagation of change between similar ontologies is propagated. This has the effect that data stewards really have the liberty to choose the ontology they prefer, or are required to use due to project constraints. Wikidata will capture the mappings.

Fourth. I don’t see how the links between the new concepts created in Wikidata will be generated. Are they created manually by data stewards? Perhaps, but on what basis?

Wikidata has a variety of mappings properties, such as equivalent class, equivalent property, exact match etc. mapping relational is done by adding statements using these properties

Fifth, using the acronym BYOD for 'Bring You Own Data' is somehow misleading since it is widely known for 'Bring Your Own Device'.

BYOD has been used as an acronym for Bring Your Own Data for the last 5 years (www.dtls.nl/fair-data/byod/). This is not something were suggesting in this paper. We are merely referring to one of these events.

Sixth, I do not agree that semantic modeling using RDF is a long term process, since you describe the entities by their properties in a graph. You can add new properties for a single entity, a set of entities or all entities at any time.

From our own experience it is a very laborious process. First the (non-linked) data needs to be understood. Subsequently, the relations needs to be made explicit. Often the resulting links are not simply a 1-1 translation from for example a column name to a property, but more an aggregation of multiple fields that are converted to RDF by drafting a set of mapping rules. We agree with the reviewer that indeed adding new properties is relatively easy, but this also shows that modelling a semantic model is more a perpetual process (i.e. laborius process) than a one time effort.

Seventh, the abstract is in relationship to the length of the paper very long - I’ve never seen such along abstract!

The abstract had 353 words. We have now cut it down to 277 words.
Reviewer 3

This paper introduced a method to connect concepts from different ontologies through the intermediary of an IRI provided by Wikidata. Not only introducing the method, the authors also presented an example of an improved data model for genetic variants by using Wikidata IRIs through their experiences in a BYOD workshop. This paper could be acceptable because their method is practical especially for ordinary researcher having their own data. One point that I am concerned about is whether Wikidata IRIs can cover various and heterogeneous data in life sciences such as datasets of minor organisms although activities like a BYOD workshop could solve the issue to some extent.

We thank the reviewer for their kind comments. We would like to point out that the manuscript was initially submitted as a long research paper, but has now been re-submitted as a short position paper. As for the point regarding coverage: the reviewer is right that there may be particular niche areas that are still underrepresented in Wikidata. However, we would like to point out that Wikidata is already a very rich resource (currently 50m+ items, more than all ontologies and vocabularies in BioPortal combined) and that it is ever expanding as communities (even niche ones) are invited to contribute the items relevant to their domain.

Reviewer 4

Summary: This is a study to explore if Wikidata is useful for generating a preliminary semantic model for your dataset in either cases in the first steps during the FAIRification process of your dataset when there is not a clear application scenario for data or in teaching how to provide your data more FAIR. Authors also claim that wikidata can be a potential hub for choosing the appropriate ontology/vocabulary concepts in next FAIRification steps. Authors' experiment was to convert a preliminary semantic model developed during teaching FAIRification into a semantic model only expressed using semantic web resources from wikidata, ie they used wikidata IRIs for naming and wikidata vocabulary for meaning. Finally authors claim that this approach has application in the research and also teaching communities since it makes more easy to make data FAIR and to teach how to make it. Finally authors also claim the application of wikidata to evaluate semantic overlap between ontologies.

Reasons to accept: The authors proposal of using Wikidata as a resource hub for providing both IRIs for the RDF layer and semantics for schema language to create the first preliminary semantic model in the FAIRification process of datasets, is interesting. I understand that what the authors proposed is that can be used as the UMLS metathesaurus for terms, but wikidata as a metathesaurus for ontology IRIs because by authors definition, Wikidata is not constrained to any set of domains, on the contrary it embeds all the domains of human knowledge, and it can be extendable by users request. I think this idea can foster very interesting discussions for the biomedical semantic
web community, for semantic web data providers who are interested in the FAIRification process of current data and for data experts who teach to provide FAIR data. I would strongly recommend to compare Wikidata with current solutions as identifiers.org [1] for IRI providers and OxO service from EBI [2] for ontology mapping to make the proposal more solid and the discussion more interesting.


We thank the reviewer for their comments. We would like to point out that the manuscript was initially submitted as a long research paper, but has now been re-submitted as a short position paper. We thank the reviewer for their suggestions. These will be addressed in a future research paper on this topic. Further, we have now made a clear invitation in our paper for others to join us in this experiment.

Reasons to reject: –

Minor comments: Please provide FAIR semantic models in order to interpret results (Fig1 and 3 are poor quality). Wrt Table 1, please provide query and the date query was executed. Please revise some typos, eg in the Abstract "though IRIs" should be fixed to "through IRIs".

As we have now re-submitted as a short position paper only one figure remains in the paper, where the text in the figure has been substantially increased. We have also corrected the typo in the abstract.

Further comments: Where are the links to the FAIR outcome of this study?
The link to the FAIR outcome of the BYOD that we describe has been included to the re-submitted paper.

Reviewer 5

This paper deals with an important question: how to teach practitioners ontology based on FAIR principle. The lessons of this paper seem to be useful in consolidating various ontology sources via Wikidata.

Considering that the paper’s main argument is the usefulness in teaching, it doesn’t look sufficient that the content of the paper is mostly filled with arguments not with empirical data such as survey results.

This paper would have been much better one if the researchers performed controlled experiments with and without using the proposed method. Comparing the outcomes would have helped to make a more informed judgement.

We thank the reviewer for their kind comments. We would like to point out that the manuscript was initially submitted as a long research paper, but has now been re-submitted as a short position paper. We do plan to do research as suggested in this position paper in the foreseeable future, and your comments will be very useful in this process.