

Opportunities and challenges presented by Wikidata in the context of biocuration

Benjamin M. Good¹, Sebastian Burgstaller-Muehlbacher¹, Tim Putman¹, Andrew Su¹

¹Department of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, 92037

Andra Waagmeester², Elvira Mitraka³

²Micelio, Veltwijcklaan 305, 2180 Antwerp, Belgium
³University of Maryland, School of Medicine, 655 West Baltimore Street, Baltimore, MD, 21201

Abstract—Wikidata is a world readable and writable knowledge base maintained by the Wikimedia Foundation. It offers the opportunity to collaboratively construct a fully open access knowledge graph spanning biology, medicine, and all other domains of knowledge. To meet this potential, social and technical challenges must be overcome - many of which are familiar to the biocuration community. These include community ontology building, high precision information extraction, provenance, and license management. By working together with Wikidata now, we can help shape it into a trustworthy, unencumbered central node in the Semantic Web of biomedical data.

Keywords—*wikidata; semantic web; ontology; crowdsourcing; wiki; biocuration; knowledge graph;*

I. INTRODUCTION

Wikidata is a world readable and writable knowledge base currently maintained by the Wikimedia Foundation [1]. It is used by the many different language Wikipedias to manage inter-language links and to host data rendered in infoboxes. Its contents are accessible for all users via the Creative Commons CC0 1.0 Universal license¹. The data can be queried via a SPARQL endpoint², retrieved as a full database download³, and manipulated both manually and programmatically via a REST API⁴.

In addition to its function as a structured datastore for the Wikimedia projects, Wikidata is being used to integrate and distribute biomedical knowledge [2]. For example, it has been used to disseminate knowledge about drug-drug interactions [3], human genes [4], and microbial genomics [5]. Here, we suggest a few of the opportunities and associated challenges that Wikidata presents to the broad biocuration community.

II. OPPORTUNITIES

A. As a fully open public knowledge graph

Wikidata's CC0 license, Semantic Web compatible implementation and active community provide a unique opportunity to assemble and disseminate knowledge. Wikidata

is currently the only major Semantic Web resource that supports open, collaborative editing. Further, through its association with the Wikipedias, it has thousands of editors working to improve its content. If orchestrated effectively, this combination of technology and community could produce a knowledge resource of unprecedented scale and value. In terms of distributing knowledge, its direct integration with the Wikipedias can allow its community vetted content to be shared with literally millions of consumers in hundreds of languages. Outside of the Wikipedias, Wikidata's CC0 license removes all barriers on re-use and redistribution of its contents in other applications. Such legal barriers to data sharing are critical blockers to scientific progress [6]. Because of its truly open access status and its standards compliant implementation, it could become the central component of the long promised Semantic Web in the life sciences.

B. As a shared concept resource for information extraction

Apart from its use as a knowledge graph, Wikidata could provide great value to the text-mining community as a multi-lingual collection of concept labels, descriptions, and links to encyclopedic text. So-called 'items' in Wikidata are roughly analogous to the concepts in the Unified Medical Language System (UMLS) [7]. Each item may have labels and descriptions in any of hundreds of different human languages as well as links to corresponding Wikipedia articles in each of these languages. In addition, Wikidata provides links to unique concept identifiers in a growing number of controlled vocabularies and ontologies, thus easing integration with and between existing knowledge bases. For example, the Wikidata item for peritonitis⁵ provides terms, aliases and article links in approximately 50 languages. Further it provides links to equivalent concepts in 11 different external resources including e.g. MeSH, the Human Disease Ontology, and ICD10. This lexical information, coupled with the growing amount of semantic information represented in the Wikidata knowledge graph, provides a powerful resource for natural language processing. Already, applications such as ContentMine are using Wikidata for this purpose [8]. Unlike the UMLS, which is centrally curated, Wikidata's distributed curation model offers the potential for far greater scale and adaptability— at the cost of greater challenges in establishing and maintaining order.

¹ <https://creativecommons.org/publicdomain/zero/1.0/>

² <https://query.wikidata.org/>

³ https://www.wikidata.org/wiki/Wikidata:Database_download

⁴ <https://www.wikidata.org/w/api.php>

⁵ <https://www.wikidata.org/wiki/Q223102>

III. CHALLENGES

A. Community ontology building

When creating a knowledge base that spans all domains of knowledge, what are the most effective patterns for representation? How can the community work most effectively together to move iteratively closer to the most useful forms? These questions are currently being tackled by a distributed, mostly-volunteer community of ontologists, technologists, domain experts, and interested citizens in discussions held in forums such as the Wikidata property proposal page⁶. Before a property (e.g. ‘part of’, ‘MeSH id’, or ‘used to treat’) can be used in Wikidata it must be proposed and approved by community consensus. Once consensus is achieved, an elected community member with administrative powers creates the property and it can then be used to add claims to any item. This property collection, and the guidelines associated with their use, forms a major part of the active ‘ontology’ of Wikidata.

In comparison to other efforts to build large knowledge graphs, the Wikidata approach is on the chaotic side. There is no rigid application of an upper ontology, no automated reasoning to support class inference or quality control, and no over-arching plan to govern the system’s evolution. Instead, there is a large, motivated, highly heterogeneous community doing their best to assemble useful structures one step at a time. So far, good progress has been made as evidenced by the early applications of Wikidata content such as the new infobox for human genes in Wikipedia [4]. That being said, there is a clear need for experienced ontologists to join the conversations and help to collaboratively guide this community forward if it is to reach its full potential.

B. Establishing computable trust

A key enabling feature of the Wikidata infrastructure is the capacity to provide provenance for its claims (the triples that compose the knowledge graph) through references. Each claim can be supported by any number of references to supporting sources of information. Unfortunately, many of the claims that are currently in Wikidata were not assigned references. These unsourced claims are of uncertain quality and may weaken the chances of community uptake. Many long-time Wikipedians are hesitant to embrace Wikidata and use the lack of references as an argument against broadly deploying its contents to support infoboxes. This situation poses a challenge to the information extraction community. Given an unsourced claim (e.g. that a drug treats a particular disease) can we develop automated or semi-automated processes for finding sources to validate or invalidate these claims? Could we apply similar processes to automatically verify references that do exist to ensure high quality? If successful, such automation could greatly help Wikidata and other similarly open initiatives forward by allaying concerns about the trustworthiness of content.

C. Building up Wikidata with text mining

The majority of the world’s biomedical knowledge remains locked up in unstructured text. As text mining matures, it is increasingly possible to extract this knowledge automatically; however (1) most people, even within the bioinformatics community, do not have the skills and resources to perform this work themselves and (2) despite many advances, workflows for generating highly reliable content still require human review. If extracted knowledge could be shared through Wikidata, it would reach the broadest possible audience, eliminating the need for consumers to build and run their own extraction pipelines. However, to achieve this, the quality of such workflows would need to be at the same level as institutional biocuration processes – likely with human verification as the final step. A challenge for the text-mining research community is to identify ways to engage the thousands of Wikidata community members to define truly scalable, high quality biocuration workflows by effectively integrating machine intelligence with community intelligence.

IV. CONCLUSION

With diligence, persistence and patience, Wikidata could become the central hub of the Web of data, uniting all domains of knowledge. The biocuration community has an opportunity to help lead this process and, in doing so, benefit all aspects of biomedical research. The time is now.

ACKNOWLEDGMENT

This work was supported by the US National Institute of Health (grants GM089820 and U54GM114833 to AIS) and by the Scripps Translational Science Institute with an NIH-NCATS Clinical and Translational Science Award (CTSA; 5 UL1 TR001114).

REFERENCES

- [1] Vrandečić D, Krotzsch M: **Wikidata: a free collaborative knowledgebase**. *Commun ACM* 2014, **57**(10):78-85.
- [2] Mitraka E, Waagmeester A, Burgstaller-Muehlbacher S, Schriml LM, Su AI, Good BM: **Wikidata: A platform for data integration and dissemination for the life sciences and beyond**. *bioRxiv* 2015:031971.
- [3] Pfundner A, Schonberg T, Horn J, Boyce RD, Samwald M: **Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages: a pilot study**. *J Med Internet Res* 2015, **17**(5):e110.
- [4] Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, Naik C, Pavlidis P, Schriml L, Good BM *et al*: **Wikidata as a semantic framework for the Gene Wiki initiative**. *Database (Oxford)* 2016, **2016**.
- [5] Putman TE, Burgstaller-Muehlbacher S, Waagmeester A, Wu C, Su AI, Good BM: **Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes**. *Database (Oxford)* 2016, **2016**.
- [6] Himmelstein D, Jensen LJ, Smith M, Fortney K, Chung C: **Integrating resources with disparate licensing into an open network**. *Thinklab* 2015, doi:10.15363/thinklab.d107.
- [7] Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic acids research* 2004, **32**:267-270.
- [8] Martone M, Murray-Rust P, Molloy J, Arrow T, MacGillivray M, Kittel C, Kasberger S, Steel G, Oppenheim C, Ranganathan A *et al*: **ContentMine/Hypothes.is Proposal**. *Research Ideas and Outcomes* 2016, **2**(e8424).

⁶ https://www.wikidata.org/wiki/Wikidata:Property_proposal