The impact of Wikimedia on the life sciences knowledge landscape

Tiago Lubiana
Univ. of São Paulo/Wiki Movimento Brasil

João Vitor Cavalcante
Fed. Univ. of Rio Grande do Norte

Pedro Medeiros
Fed. Univ. of ABC

Abstract

In this research, we propose to evaluate the use of Wikipedia, Wikidata, and Wikimedia Commons on the life sciences biocuration landscape, focusing on ontologies and knowledge bases. We will combine development, data analysis, and community engagement to gain insights on this interface. By understanding the reuse of Wikimedia content, we can better position it as an *essential infrastructure of the ecosystem of free knowledge* while benefiting the life sciences community.

Introduction

Ontologies and knowledge bases are central to modern life sciences research. Computational ontologies power data annotation and knowledge bases are the go-to places for domain-specific information. These biocurated resources , such as the Rfam database and Cellosaurus,[1,2] re-use, link, and point to Wikimedia content. These re-uses signal trust and pave new ways for Wikimedia knowledge to flow. However, the details of that reuse are undocumented, hindering informed actions.

We will provide a panorama of the impact of Wikimedia projects in core life sciences ontologies, answering questions like: how many cross-references do they have to English Wikipedia? How many entities use Commons as a source? Which media do they link? Which kind of relations are used to map to Wikidata, if any?

The OBO Foundry gathers over 200 life sciences ontologies sharing technological standards. The homogeneity enables software processing of their content, and preliminary analysis shows they often cross-reference Wikipedia pages (Figure 1).

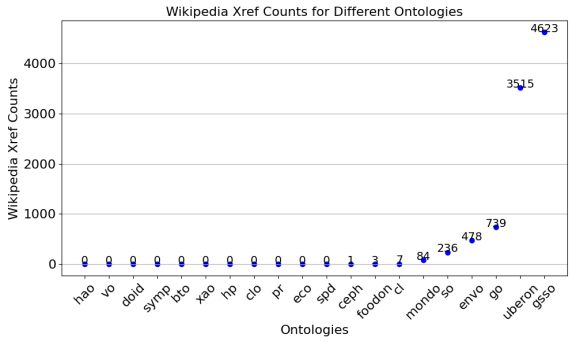

Figure 1: Cross-references to Wikipedia in a selected set of OBO Foundry ontologies as of October 2023.

Beyond ontologies, we will investigate how core knowledge bases link to Wikimedia content. We will create a taxonomy of how Wikipedia, Wikidata, and Commons are reused from the data, gaining insights on how to foster Wikimedia-biocuration collaborations.

With our results, we expect biocurators to leverage Wikimedia for their resources better while bringing Wikimedia content to a wider audience. At the same time, we expect to entice

academics to join the movement, as biocurators share values of open, organized knowledge with Wikimedians. [3–5] Finally, we expect the additional knowledge and engagement to increase the value of Wikimedia for society, and consequently towards increased sustainability.

Date:
Start date: July 01, 2024
End date: June 30, 2025

Related work

We are unaware of research works focused on the details of the reuse of Wikimedia by biocuration resources. Works on the subject include:

- T.L.'s (author) Ph.D. project on Wikidata for biocuration of cell types. [7]
- A previous Wikimedia Research project by Houcemedine Turki and colleagues, including coverage of OBO Foundry on Wikidata.
- Papers about individual databases and their use of MediaWiki or Wikimedia content. [3–5]
- The Gene Wiki project, which promoted Wikipedia and Wikidata as direct biocuration interfaces. [6,8]

Methods

The project includes two connected parts. For ontologies:

- Develop software to parse the ontologies source code using ROBOT, extract Wikimedia references, and generate plots.
- Build a pipeline to re-run analysis periodically as ontologies are updated
- Create a web portal showcasing the results

For knowledge bases:

- Develop software to detect Wikimedia links in webpages
- Analyze the Global Core Biodata Resources and the ELIXIR Core Data Resources for Wikimedia links [9,10]

Expected output

- A web portal, a preprint on bioRxiv, and a peer-reviewed publication in Database or similar, making results available for the biocuration and the Wikimedia communities.

- Two open software pipelines, tailored to detect Wikimedia links in ontologies and knowledge bases

- Present the work virtually in 2+ academic conferences and 2+ Wikimedia events.

Risks

We outline two main risks:

- Risk of low prevalence of Wikimedia in the target resources. While it would not decrease the project's usefulness, it might make the research less flashy
- Risk of failure in meeting deadlines due to tight schedule. While we are designing flexible milestones, we acknowledge the inherent risk of delays.

Community impact plan

While our project targets the academic community, it includes engaging researchers and other communities with Wiki projects. For example via

- Dialog with the OBO Foundry community via Slack

- Building bridges with the International Society for Biocuration

- Connecting the Wiki Movimento Brazil community with Wikimedia Research

Evaluation

We will evaluate progress on a continuous base, with the final milestones being a preprint and a journal submission. Continuous check interfaces, besides the on-wiki reports, include:

- An evolving open academic article via Manubot (https://manubot.org/).
- An open portal with plots and results generated with GitHub pages.
- An open repository on GitHub, enabling checks for good documentation and test coverage, for example.

Budget:

We estimate an approximate 25k USD total budget, divided initially into:

- ~19k for stipends to fund three part-time research software developers with experience with Wikimedia activities. We calculated an hourly rate of 30 USD/hour for the budget estimate.
- ~2.5k for open-access publishing fees
- ~3.5k for institutional overhead (15%)

Prior contributions

- Tiago Lubiana is a Ph.D. candidate in Bioinformatics at the University of São Paulo in Brazil to graduate in early 2024. He has authored several peer-reviewed papers in the life sciences, some related to the role of Wikidata. He is a long-term Wikimedia contributor and a Wiki Movimento Brasil User Group member. He is also a member of the International Society for Biocuration and actively contributes to the OBO Foundry initiative.
- João Vitor Cavalcante is a MSc student in Bioinformatics at the Federal University of Rio Grande do Norte, Brazil. He has worked in Bioinformatics for over 5 years and has authored peer-reviewed papers in metagenomics and transcriptomics. He has also been a Wikidata editor since 2020, collaborating on the Wikiproject COVID-19 and others.
- Pedro Medeiros is a Biologist and Pharmacist from the University of Sao Paulo with an MSc in Biosystems from the University of ABC, both in Brazil. He has worked in Bioinformatics in academia and industry and participated in several projects regarding the use of open databases in biological research.
- The authors have worked together on mapping the Brazilian bioinformatics landscape on Wikidata (dashboard on https://lubianat.github.io/bioinfo_brasil/dashboard)

References

1. Gardner, P. P. *et al.* Rfam: Wikipedia, clans and the ‘decimal’ release. *Nucleic Acids Res.* **39**, D141–D145 (2011).

2. SIB Swiss Institute of Bioinformatics RDF Group Members *et al.* The SIB Swiss Institute of Bioinformatics Semantic Web of data. *Nucleic Acids Res.* gkad902 (2023) doi:10.1093/nar/gkad902.

3. Biocuration: Distilling data into knowledge.

PLoS Biol **16**, e2002846 (2018).

4. Wiki's wild world. *Nature* **438**, 890–890 (2005).
5. Finn, R. D., Gardner, P. P. & Bateman, A. Making your database available through Wikipedia: the pros and cons. *Nucleic Acids Res.* **40**, D9–D12 (2011).
6. Waagmeester, A. *et al.* Wikidata as a knowledge graph for the life sciences. *eLife* **9**, e52614 (2020).
7. Lubiana, T. & Nakaya, H. Building a biological knowledge graph via Wikidata with a focus on the HumanCell Atlas. (2021) doi:10.5281/ZENODO.4723818.
8. Huss, J. W. *et al.* A Gene Wiki for Community Annotation of Gene Function. *PLoS Biol* **6**, e175 (2008).
9. Cook, C. & Cochrane, G. Global Biodata Resources: Challenges to long-term sustainability of a crucial data infrastructure. *Biodivers. Inf. Sci. Stand.* **6**, e90946 (2022).
10. Drysdale, R. *et al.* The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. *Bioinformatics* **36**, 2636–2642 (2020).