Wikidata WikiProject COVID-19: modelling the pandemic in real time

Tiago LUBIANA [a,1]

[a] *Computational Systems Biology Laboratory, University of São Paulo, Brazil*

The COVID-19 crisis has led to a surge in biomedical data related to COVID-19. This wealth of information, continuously updated, prompts efforts for processing, understanding, and interpreting these data. Wikidata is an open, public domain, knowledge graph representing concepts in a variety of domains and interlinks them by relations. The interest of the biomedical community for Wikidata is on the rise, and Wikidata is poised to become a global knowledge graph for the life sciences [1].

This work is a case report on the WikiProject COVID-19 on Wikidata and possible implications for the biomedical community. The author contributes to managing the project since its creation on the 16th of March of 2020. The project is the fruit of the effort of its more than 50 participants (spread across nations) in collaboration with the broader Wikidata community.

The Wikidata WikiProject COVID-19 is a collaborative, multinational effort to improve the representation of pandemic-related content on Wikidata. The project has developed collaborative models using EnititySchemas and natural language descriptions. The data is accessible in a web-based format via the Wikidata API and wrapper packages in Python and R. The contributions of the WikiProject COVID-19 participants have led to a powerful resource for the life sciences community to parse our collective knowledge.

The WikiProject is subdivided into branches for ontological representations of different areas of knowledge. The areas range from the modeling of epidemiological information (case, death, hospitalization, and recovery counts) to curating emergency measures, numbers on hospital beds, as well as concepts directly related to the biology of SARS-CoV-2.

Project members have developed a variety of data models to reconcile external data to Wikidata. These include, but are not limited to, data models, and Shape Expression Entity Schemas on hospitals, preprints, outbreaks themselves, emergency measures, macromolecular complexes, virus strains. These models have been used to integrate datasets in semantic format. In Wikidata, these datasets become integrated with OBO ontologies such as the Gene Ontology and the Disease Ontology and bibliometric information on scholars and institutions.

Also, Wikidata provides linked information on a range of encyclopedic topics (from physical constants to demographic information), totaling more than 90 million items. This integrated information is available openly and can be queried via a SPARQL endpoint, which enables complex queries to get insights into our collective knowledge of COVID-19.[2]

[1] Corresponding Author: Tiago Lubiana, University of São Paulo, Brazil. E-mail:tiago.lubiana.alves@usp.br

Queries such as "Which drugs inhibit proteins that bind SARS-CoV-2 proteins?" can be made in the user-friendly Wikidata SPARQL system (https://query.wikidata.org/). Third-party applications, such as the bibliometrics-oriented Scholia[3], make the data even more accessible. Noticeably, as Wikidata is fed with more knowledge by the biomedical community, these queries are automatically updated, providing an invaluable source of updated, integrated biomedical information on SARS-CoV-2 and the COVID-19 pandemic.

The WikiProject is open for new participants, and participation does not require any previous expertise. Alongside editing, project participants discuss data models, automate integration, and collaborate with other ongoing efforts to curate and improve the usability of COVID-19 data. Crucially, project participants are improving the inner workings of the Wikidata system for integrating biomedical knowledge, preparing the community for handling the next crises.

To sum up, WikiProject Wikidata COVID-19 is a collaborative international effort improving the availability of open, semantically-linked data about the virus and the pandemic. Any individual can contribute to its efforts by adding reference information about their topic of expertise or preference. This data is publicly available and provides a comprehensive resource for SARS-CoV-2 related knowledge, which is both machine-readable and human-readable. It is a significant asset for ontologists, computational biologists, and life scientists in general.

References

[1] Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM. Science Forum: Wikidata as a knowledge graph for the life sciences. Elife. 2020 Mar 17;9:e52614.

[2] Addshore, Mietchen, D, Willighagen, E. Wikidata Queries around the SARS-CoV-2 virus and pandemic. 2020: Zenodo. https://doi.org/10.5281/zenodo.3977414

[3] Nielsen FÅ, Mietchen D, Willighagen E. Scholia, scientometrics and wikidata. InEuropean Semantic Web Conference 2017 May 28 (pp. 237-259). Springer, Cham.