

Analysing the extent of cell type information present in Wikidata: A case study on PanglaoDB

This manuscript ([permalink](#)) was automatically generated from [jvfe/manuscript_panglaodb@852b94a](#) on September 2, 2020.

Authors

- **João Vitor Ferreira Cavalcante**

 [0000-0001-7513-7376](#) ·  [jvfe](#)

Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte

- **Tiago Lubiana Alves**

 [0000-0003-2473-2313](#) ·  [lubianat](#)

Computational Systems Biology Laboratory, University of São Paulo

Abstract

[Wikidata](#), a freely editable knowledge graph database, presents a great opportunity for the integration of biomedical knowledge, it's well thought linked data model can improve significantly the handling and distribution of scientific information. On the other hand, Wikidata is still lacking in various aspects, in particular to what pertains to cell type information. This study aims to analyse how cell type knowledge is currently modelled in Wikidata and how it differs from other type of biological information, using, as a reference point, metadata from the well known single cell RNA sequencing database, [PanglaoDB](#).

Keywords: wikidata, knowledge graph, cell type, ontology.

Introduction

Wikidata

[Wikidata](#) is an open, freely editable, knowledge graph database within the [semantic web](#) that stores knowledge across a multitude of domains, such as arts, history, chemistry and biology, using an item-property-value linked data model (Figure 1). It is easy to use and edit, by both humans and machines, with a rich web user interface and wrapper packages available in common programming languages such as R and Python. All the data within Wikidata is linked and inherently public domain, thus, it presents a great opportunity to make scientific data more FAIR (Findable, accessible, interoperable and reusable), as well as provides the necessary tools to curate and develop ontologies.

hepatocyte

(Q827450) Item Identifier

Item Label

liver cell type

Hepatocyte | hepatocytes

edit

In more languages

Configure

Language	Label	Description	Alternative labels
English	hepatocyte	liver cell type	Hepatocyte hepatocytes
Portuguese	Hepatócito	tipo celular	
German	Hepatozyt	Hauptzelltyp der Leber	Hepatocyt Leberepithelzelle Hepatozyten Leberzelle
Italian	Epatocito	No description defined	Epatociti Epatocita

All entered languages

Statements

Statement

instance of

Property

cell type

Value

edit

0 references

+ add reference

+ add value

Figure 1: Wikidata item example, showing item hepatocyte (Q827450), the labels change according to the user's language, but each item has a universal identifier, called QID.

Several advances towards biological data integration and biological data analysis in Wikidata have been made before, yielding positive results [1] [2] and showcasing it's potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [2]. Wikidata has been proposed as a unified base to gather and distribute biomedical knowledge, with more than 50 000 human gene items indexed and hundreds of biomedical-related properties [3]. However, as of August 2020, cell type information is still very scarce, with only 264 items being categorized as "instances of cell types (Q189118)" (<https://w.wiki/b2w>), of those, only nine have a "Cell Ontology ID"[4] (P7963) associated, and most have a varying amount of statements (Table 1). As an additional problem, there are also 23 items being categorized as "instances of cell (Q7868)" (<https://w.wiki/b2x>), illustrating the absence of any formal data model.

Table 1: As of August 2020, Wikidata items regarding cell types have a varying amount of information, with most having very few statements.

Cell type Item	Number of statements
red blood cell (Q37187)	48
myocyte (Q428914)	18
mesenchymal cell (Q66568500)	2

PanglaoDB

[PanglaoDB](#) [5] is a public database that contains data and metadata on hundreds of single-cell RNA sequencing experiments, providing extensive information on cell types, genes and tissues, as well as manually and community curated cell type markers (Tables 2 and 3). It also provides a rich web user interface for easy data acquisition, including database dumps for bulk downloads.

Table 2: Database statistics for each species in PanglaoDB, as of 31st of August, 2020.

	Mus musculus	Homo sapiens
Samples	1063	305
Tissues	184	74
Cells	4,459,768	1,126,580
Cell Clusters	8,651	1,748

Table 3: Metadata statistics for PanglaoDB, gathered from their [last update on August, 2019](#).

	Number
Cell types	215 (uniquely named)
Tissues	240 (+6 germ layers)
Organs	29
Species	2 (Homo sapiens and Mus musculus)
Genes	110292

Objectives

In this study, we aim to answer questions regarding the integration of biological data from PanglaoDB in Wikidata, analysing items such as cell types, genes and tissues. Some of the questions we gathered so far are:

- How many cell types in PanglaoDB are also present in Wikidata? How many of those items are exact matches?
- Of those that are exact matches, how many statements do they have associated? Are these items well annotated?
- How does the coverage of biological items differ within Wikidata? Are the items for tissues and genes better annotated? How so?
 - Do items with alternative identifiers, such as genes, have their alternative identifiers indexed within Wikidata?

In the end, we'll have gathered and analysed enough data to formulate a report on the integration of this knowledge. We also intend to migrate the missing data, enriching Wikidata with more biological information.

Methodology

Data acquisition

Data from Wikidata will be acquired using the [Wikidata Query Service](#) and associated wrapper packages in Python, such as [WikidataIntegrator](#) and [wikidata2df](#).

Data from PanglaoDB will be acquired through their [web interface](#) and [metadata database dump repository](#).

All data used will be handled with commonly used Python data science packages, such as Pandas[6], Seaborn[7] and Jupyter[8].

Reconciliation

The metadata from PanglaoDB on cell types, genes, tissues (including germ layers) and organs will be matched to Wikidata items using the [reconciler](#) Python package, which is itself a wrapper around the well known [OpenRefine](#) reconciliation service, as well as manual intersections of both data sources. Data from the reconciliation service will be considered a match if the service returns a value of "match" equals to "True".

Item quality assessment

Wikidata items will be assessed for their quality by their number of statements, which can be acquired via both the [MediaWiki API](#) and Wikidata's own query service. And also by the presence of external identifiers, such as Ensembl Gene[9] and Entrez Gene[10] IDs for genes, Cell Ontology[4] IDs for cell types and Uberon[11] IDs for organs and tissues.

References

1. **Wikidata: A platform for data integration and dissemination for the life sciences and beyond**

Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M Schriml, Andrew I Su, Benjamin M Good
bioRxiv (2015-11-16) <https://doi.org/gg9dk4>
DOI: [10.1101/031971](https://doi.org/10.1101/031971)

2. **Wikidata as a knowledge graph for the life sciences**

Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, ... Andrew I Su
eLife (2020-03-17) <https://doi.org/gggqc6>
DOI: [10.7554/elife.52614](https://doi.org/10.7554/elife.52614) · PMID: [32180547](https://pubmed.ncbi.nlm.nih.gov/32180547/) · PMCID: [PMC7077981](https://pubmed.ncbi.nlm.nih.gov/PMC7077981/)

3. **Wikidata: A large-scale collaborative ontological medical database**

Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, Helmi Hamdi
Journal of Biomedical Informatics (2019-11) <https://doi.org/gg9dnt>
DOI: [10.1016/j.jbi.2019.103292](https://doi.org/10.1016/j.jbi.2019.103292) · PMID: [31557529](https://pubmed.ncbi.nlm.nih.gov/31557529/)

4. **The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability.**

Alexander D Diehl, Terrence F Meehan, Yvonne M Bradford, Matthew H Brush, Wasila M Dahdul, David S Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, ... Christopher J Mungall
Journal of biomedical semantics (2016-07-04) <https://www.ncbi.nlm.nih.gov/pubmed/27377652>
DOI: [10.1186/s13326-016-0088-7](https://doi.org/10.1186/s13326-016-0088-7) · PMID: [27377652](https://pubmed.ncbi.nlm.nih.gov/27377652/) · PMCID: [PMC4932724](https://pubmed.ncbi.nlm.nih.gov/PMC4932724/)

5. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**

Oscar Franzén, Li-Ming Gan, Johan LM Björkegren
Database (2019) <https://doi.org/ggkzxr>
DOI: [10.1093/database/baz046](https://doi.org/10.1093/database/baz046) · PMID: [30951143](https://pubmed.ncbi.nlm.nih.gov/30951143/) · PMCID: [PMC6450036](https://pubmed.ncbi.nlm.nih.gov/PMC6450036/)

6. **pandas-dev/pandas: Pandas 1.0.0**

Jeff Reback, Wes McKinney, Jbrockmendel, Joris Van Den Bossche, Tom Augspurger, Phillip Cloud, Gfyoung, Sinhrks, Adam Klein, Matthew Roeschke, ... Thomas Kluyver
Zenodo (2020-01-29) <https://doi.org/gg9gtt>
DOI: [10.5281/zenodo.3630805](https://doi.org/10.5281/zenodo.3630805)

7. **mwaskom/seaborn: v0.10.1 (April 2020)**

Michael Waskom, Olga Botvinnik, Joel Ostblom, Maoz Gelbart, Saulius Lukauskas, Paul Hobson, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, ..., Brian
Zenodo (2020-04-26) <https://doi.org/gg4t5p>
DOI: [10.5281/zenodo.3767070](https://doi.org/10.5281/zenodo.3767070)

8. **IPython: A System for Interactive Scientific Computing**

Fernando Perez, Brian E. Granger
Computing in Science & Engineering (2007) <https://doi.org/dcs6r3>
DOI: [10.1109/mcse.2007.53](https://doi.org/10.1109/mcse.2007.53)

9. Ensembl 2020

Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amodé, Irina M Armean, Andrey G Azov, Ruth Bennett, ... Paul Flicek
Nucleic Acids Research (2019-11-06) <https://doi.org/ggqp72>
DOI: [10.1093/nar/gkz966](https://doi.org/10.1093/nar/gkz966) · PMID: [31691826](https://pubmed.ncbi.nlm.nih.gov/31691826/) · PMCID: [PMC7145704](https://pubmed.ncbi.nlm.nih.gov/PMC7145704/)

10. Database resources of the National Center for Biotechnology Information

NCBI Resource Coordinators
Nucleic Acids Research (2012-11-26) <https://doi.org/gg9gtr>
DOI: [10.1093/nar/gks1189](https://doi.org/10.1093/nar/gks1189) · PMID: [23193264](https://pubmed.ncbi.nlm.nih.gov/23193264/) · PMCID: [PMC3531099](https://pubmed.ncbi.nlm.nih.gov/PMC3531099/)

11. Uberon, an integrative multi-species anatomy ontology.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, Melissa A Haendel
Genome biology (2012-01-31) <https://www.ncbi.nlm.nih.gov/pubmed/22293552>
DOI: [10.1186/gb-2012-13-1-r5](https://doi.org/10.1186/gb-2012-13-1-r5) · PMID: [22293552](https://pubmed.ncbi.nlm.nih.gov/22293552/) · PMCID: [PMC3334586](https://pubmed.ncbi.nlm.nih.gov/PMC3334586/)