

Analysing the extent of biological information present in Wikidata

This manuscript ([permalink](#)) was automatically generated from [jvfe/manuscript_panglaodb@d0f09d9](#) on August 31, 2020.

Authors

- **João Vitor Ferreira Cavalcante**

-  [jvfe](#)

Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte

- **Tiago Lubiana Alves**

-  [lubianat](#)

Computational Systems Biology Laboratory, University of São Paulo

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Intellegi quidem, ut propter aliam quamquam rem, verbi gratia propter voluptatem, nos amemus; Nec tamen ille erat sapiens quis enim hoc aut quando aut ubi aut unde? Duo Reges: constructio interrete. Non pugnem cum homine, cur tantum habeat in natura boni; Gloriosa ostentatio in constituendo summo bono. Me igitur ipsum ames oportet, non mea, si veri amici futuri sumus. Cur ipse Pythagoras et Aegyptum lustravit et Persarum magos adiit? Quae animi affectio suum cuique tribuens atque hanc, quam dico. Quis est autem dignus nomine hominis, qui unum diem totum velit esse in genere isto voluptatis? Qui si ea, quae dicit, ita sentiret, ut verba significant, quid inter eum et vel Pyrrhonem vel Aristonem interesset?

***Keywords:** Wikidata, Knowledge graph, Single-cell, 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. 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. Several advances towards biological data integration and biological data analysis in Wikidata have been made before, showing positive results [1] [2]. However, as of August 2020, cell type information is still very scarce, with only 264 items being categorized as instances of cell types (Q189118), of those, almost none have a “Cell Ontology ID”[3], and most have a varying amount of statements (Table 1).

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](#) [4] 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 (Table 2). It also provides a rich web user interface for easy data acquisition, including database dumps for bulk downloads.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Intellegi quidem, ut propter aliam quamquam rem, verbi gratia propter voluptatem, nos amemus; Nec tamen ille erat sapiens quis enim hoc aut quando aut ubi aut unde? Duo Reges: constructio interrete. Non pugnem cum homine, cur tantum habeat in natura boni; Gloriosa ostentatio in constituendo summo bono. Me igitur ipsum ames oportet, non mea, si veri amici futuri sumus. Cur ipse Pythagoras et Aegyptum lustravit et Persarum

magos adiit? Quae animi affectio suum cuique tribuens atque hanc, quam dico. Quis est autem dignus nomine hominis, qui unum diem totum velit esse in genere isto voluptatis? Qui si ea, quae dicit, ita sentiret, ut verba significant, quid inter eum et vel Pyrrhonem vel Aristonem interesset?

Table 2: Database statistics for PanglaoDB, as of 31st 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

Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

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/gggqgc6>
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. **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/)
4. **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/)