Wikidata to build 5-star Linked Open biological databases: A case study of PanglaoDB

This manuscript (permalink) was automatically generated from jvfe/paper_wdt_panglao@c1731ff on January 26, 2021.

Authors

- **João Vitor Ferreira Cavalcante**
  0000-0001-7513-7376 · jvfe
  Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte

- **Tiago Lubiana**
  0000-0003-2473-2313 · lubianat
  Computational Systems Biology Laboratory, University of São Paulo

Abstract

PanglaoDB is a database of cell type markers widely used for single cell RNA sequencing data analysis. The genes, tissues, organs and cell types mentioned in the database, however, are described by free text and lack identifiers. Wikidata, is a freely editable knowledge graph database useful for the integration of biomedical knowledge. Its linked data model can improve significantly the handling and distribution of scientific information.

In this study we explore the feasibility of enriching PanglaoDB with Wikidata identifiers. We accessed the state of reconciliation at the beginning of the project, comparing the modelling of genes, tissues, organs and cell types on Wikidata. Taking advantage of the openess of Wikidata, we leveraged our initial analysis to contribute towards Wikidata completeness and enable full reconciliation. As a final product, we released the first SPARQL endpoint for cell marker information, in a 5-star open linked data format. We hope that this study encourages further reconciliations of databases to Wikidata.

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

Introduction

PanglaoDB

PanglaoDB [1] [2] is a publically-available database that contains data and metadata on hundreds of single-cell RNA sequencing experiments. It provides extensive information on cell types, genes and tissues, as well as manually and community curated cell type markers (Tables 1 and 2). It also displays a rich web user interface for easy data acquisition, including database dumps for bulk downloads.

Table 1: 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 2: 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

As of 30 December 2020, the article describing PanglaoDB has been cited 88 times. Despite its use by the the community, the database is on a 3-star category for Linked Open Data [3] as it does not use the semantic open standards from W3C (RDF and SPARQL) needed for a 4-star rank, neither the links to external data via common identifiers that makes datasets 5-star. Improving the data format toward

W3C's gold standards is a valuable step in making biological knowlegde FAIR (Findable, Acessible, Interoperable and Reusable).

The OBO Foundry provides a rich collection of linked biological identifiers [4]. However, reconciliation to OBO is challenging, as there are many ontologies, each with slightly different contribution guidelines. For that reason, we decided to reconcile PanglaoDB to Wikidata, which allows simple creation of new terms, provided they follow Wikidata`s notability criteria [5].

Wikidata

Wikidata [6] is an open, freely editable, knowledge graph database within the semantic web [7] 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.

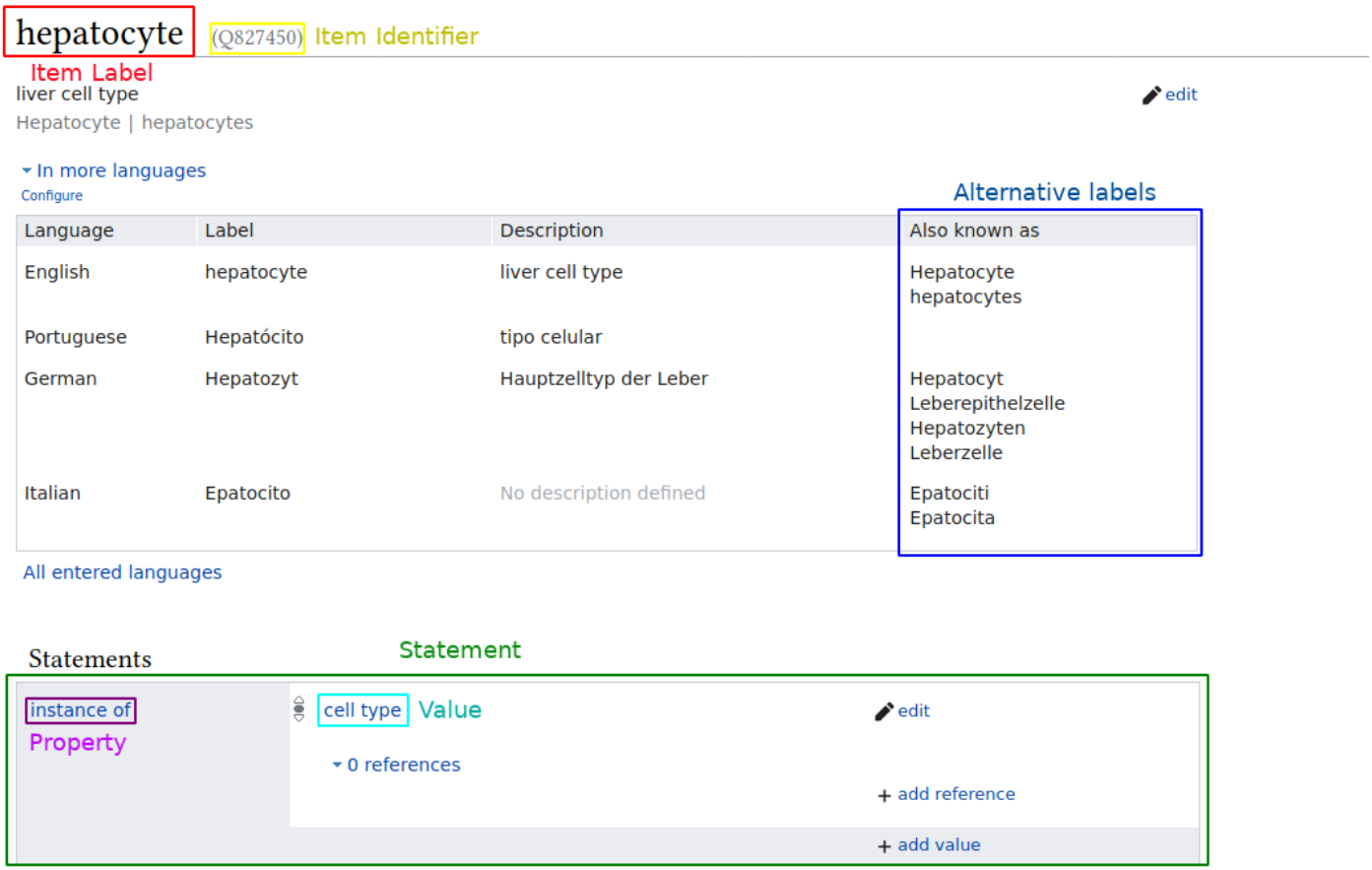

Figure 1: Wikidata item example, showing item hepatocyte (Q827450), the labels change according to the user's language, but each item has a universal identifier, called QID.

Several advances towards biological data integration and biological data analysis in Wikidata have been made before, yielding positive results [8] [9] and showcasing it's potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [10]. 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 [11].

Wikidata is nevertheless a collaborative database, and content is available on different levels of quality. For example, as of August 2020, cell type information was still very lacking, with only 264 items

being categorized as “instances of cell types (Q189118)” (https://w.wiki/b2w), while other projects describe over 2.000 cell types [12,13]. Of those 264 items, only 9 have a “Cell Ontology ID”[14] (P7963) associated, and most have a varying amount of statements (Table 3). As an additional problem, there are also 23 items being categorized as “instances of cell (Q7868)” (https://w.wiki/b2x), an imprecision, as an instance of cell would be an individual named cell from a single named individual.

Table 3: 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

This study was motivated by the increasing importance of cell-type concepts in light of the Human Cell Atlas [15], and the utter need for improved inteoperability of biological data. We aimed, thus, at providing a case study of the re-release PandlaoDB in a 5-star Linked Open Data Format while improving the modelling of the necessary concepts on Wikidata.

Methodology

Data acquisition

Gene data from Wikidata was acquired using the Wikidata Query Service [16] - https://w.wiki/bWc for *Homo sapiens* genes and https://w.wiki/bWe for *Mus musculus* genes.

Data for quality acessment from PanglaoDB was acquired through their metadata database dump repository [17].

The markers dataset was dowloaded manually from PanglaoDB’s website (https://panglaodb.se/markers/PanglaoDB_markers_27_Mar_2020.tsv.gz). It contains 15 columns and 8256 rows.

For the reconciliation, only the columns `species`, `official gene symbol` and `cell type` were used.

All data used was handled using the Pandas [18] library, with the Seaborn [19] and Matplotlib [20] libraries being used for plotting.

Automated matching

The metadata from PanglaoDB on cell types, tissues (including germ layers) and organs was matched to Wikidata items using the reconciler [21] library with the Wikidata Reconciliation Service [22], tools that operate under the W3C Reconciliation Service specification [23].

Further matching was done using a custom stemming function on the item labels, via PorterStemmer from the NLTK library [24].

Matches were considered perfect if the reconciliation service or the stemming function returned a value of “match” equals to “True”. Matches were manually analysed for false matches, such as items with same labels but used for different concepts.

Gene data was matched manually using a Pandas [18] inner merge, since both data sources contained identifiers, which should be the same.

Item quality assessment

Wikidata items were assessed for their quality by their number of statements, which were acquired using a custom wrapper on the MediaWiki API [25] and, in the case of gene data, via Wikidata's own query service, as stated in the Data acquisition section.

Furthermore, items were also assessed by the presence of external identifiers - all of which are Wikidata properties:

- Ensembl Gene [26] (P594) and Entrez Gene [27] (P351) IDs for genes,
- Cell Ontology [14] (P7963) IDs for cell types
- Uberon [28] (P1554) IDs for organs and tissues.

Class creation on Wikidata

Classes corresponding to species-neutral classes were retrieved from Wikidata manually using Wikidata's Graphic User Interface. The dictionay matching terms in PanglaoDB to Wikidata identifiers were stored in a reference csv table.

Cell types which were not represented on Wikidata were added to the database via the graphical user interface (https://www.wikidata.org/wiki/Special:NewItem) and logged in the reference table.

Species-specific cell types for human and mouse cell types were created for every entry in the reference table, connected to the species-neutral concept via a "subclass of" property (e.g. every single "human neutrophil" is a also "neutrophil"). Our approach was analogous to the one taken by the CELDA ontology to create species-specific cell-types, with the difference that they used the `rdfs:subClassOf` class to denothe the subclass relationship [29].

The reference sheet for species-neutral concepts was used to obtain the "subclass of" for every newly created item. Each item was labeled either "human" + the label for the neutral cell type, described as "cell type found in Homo sapiens" and tagged with the statement "found in taxon" *Homo sapiens*. An analogous framework was used for mouse cell types, assuming that mouse in PanglaoDB meant *Mus musculus*. Batch creations were added to Wikidata via the tool Quickstatements (https://quickstatements.toolforge.org/#/).

All genes in PanglaoDB either were already present on Wikidata or resolved to multiple entities and thus were excluded.

Property creation on Wikidata

Properties on Wikidata need to be supported by the users in a public forum before creation. To represent the cell-type marker relation, we proposed a property called `has marker` to the Wikidata community.
We posted a message in 17th of November presenting the property, domain and range constraints, as well as additional comments.

The proposal was accompanied by the following motivation statement:

```
"Even though the concept of a marker gene/protein is not clear cut, it is
very important, and widely used in databases and scientific articles.

This property will help us to represent that a gene/protein has been
reported as a marker by a credible source, and should always contain a
reference.

Some markers are reported as proteins and some as genes. Some genes don`t
encode proteins, and some protein markers are actually protein complexes.

The property would be inclusive to these slightly different markers.
Some cell types are marked by absence of expression of
genes/proteins/protein expression. As these seem to be less common than
positive markers (no organized databases, for example) they are left outside
the value range for this property"
```

The proposal contained specifications of the property such as:

- Description:
  - “a gene or a protein published as a marker of a species-specific cell type”
- Data type:
  - Item (internal entities in Wikidata)
- Domain:
  - ?subject instance of (P31) cell type (Q189118)
- Allowed values:
  - {?object instance of (P31) protein (Q8054) .}
  - UNION {?object instance of (P31) gene (Q7187) .}
  - UNION {?object instance of (P31) macromolecular complex (Q22325163) .}
- Planned use:
  - Reconcile knowledge from the PanglaoDB marker database to Wikidata. In the future, expand to other trusted sources of cell type marker information.

More details can be in the archived Wikidata:Property proposal page (https://www.wikidata.org/wiki/Wikidata:Property_proposal/has_positive_marker).

Integration to Wikidata

The reconciled dataset was uploaded to Wikidata via the WikidataIntegrator python package [30], a wrapper for the Wikidata Application Programming Interface. The details of the integration can be seen in the accompanying Jupyter notebook.

Access to reconciled data

Wikidata dumps

Wikidata provides regular dumps in a variety of formats, including RDF dumps: https://www.wikidata.org/wiki/Wikidata:Database_download. It is possible to also download partial dumps of the database with reduced size (ex: https://wdumps.toolforge.org/dump/987 for all cell types with the `has_marker` property).

SPARQL queries

Besides the Wikidata Dumps, Wikidata provides an SPARQL endpoint with a Graphical User Interface (https://query.wikidata.org/). Updated data was immediately accessible via this endpoint, enabling integrative queries integrated with other database statements.

Source code and data availability

All source code used for the study and data created during the study are available in a GitHub repository, https://github.com/jvfe/wikidata_panglaodb, as well as archived in a zenodo repository, https://doi.org/10.5281/zenodo.4438614.

Results

Cell Marker information on Wikidata

Adding marker information on Wikidata was not possible before this study and became possible after community approval of the property "has marker" (P8872) (see Methods). Figure 2 shows 2 of the current markers of "human colinergic neuron"(Q101405051), CHAT and ACHE, as they seen on Wikidata. The PanglaoDB is referenced both via URL to the website (https://panglaodb.se/markers.html) and a pointer to the PanglaoDB item on Wikidata, Q99936939.

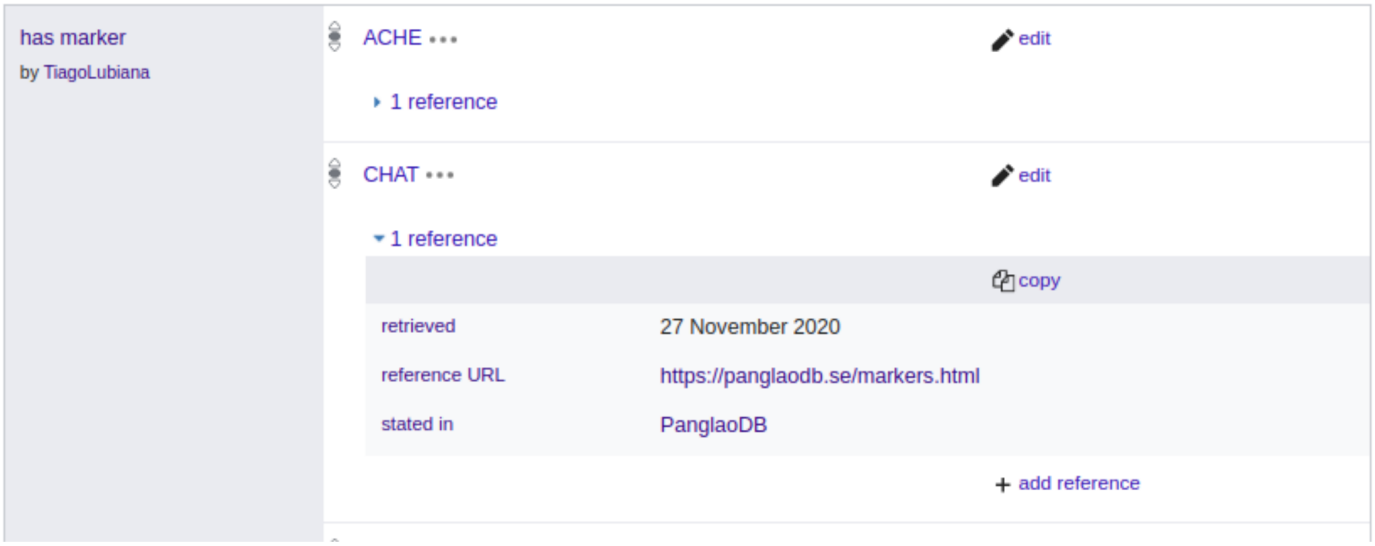

Figure 2: Subset of the marker genes for item Q101405051 (human cholinergic neuron)

Since Wikidata is an open system information about markers will be complemented by user contributions. To date, no other project has systematically integrated cell type markers to Wikidata, and most information is provenient from PanglaoDB. The queries below show an update view of the marker count for cell types of humans and mice on Wikidata.

Marker information on Wikidata for cell types found in *Homo sapiens*

cell_type	cell_typeLabel	mar
(http://www.wikidata.org/entity/Q101405035) wd:Q101405035 (http://www.wikidata.org/entity/Q101405035)	human interneuron	216
(http://www.wikidata.org/entity/Q101405104) wd:Q101405104 (http://www.wikidata.org/entity/Q101405104)	human neuron	203
(http://www.wikidata.org/entity/Q68621315) wd:Q68621315 (http://www.wikidata.org/entity/Q68621315)	human endothelial cell	187
(http://www.wikidata.org/entity/Q101404861) wd:Q101404861 (http://www.wikidata.org/entity/Q101404861)	human fibroblast	170
(http://www.wikidata.org/entity/Q101405101) wd:Q101405101 (http://www.wikidata.org/entity/Q101405101)	human hepatocyte	149

Wikidata Query Service
(https://query.wikidata.org/#SELECT%20%3Fcell_type%20%3Fcell_typeLabel%20%28count%28%3Fmarker%2

Marker information on Wikidata for cell types found in *Mus musculus*

cell_type	cell_typeLabel	mar
(http://www.wikidata.org/entity/Q102426621) wd:Q102426621 (http://www.wikidata.org/entity/Q102426621)	mouse neocortical interneuron	219
(http://www.wikidata.org/entity/Q104416243) wd:Q104416243 (http://www.wikidata.org/entity/Q104416243)	mouse interneuron	219
(http://www.wikidata.org/entity/Q104416303) wd:Q104416303 (http://www.wikidata.org/entity/Q104416303)	mouse neuron	210
(http://www.wikidata.org/entity/Q104416178) wd:Q104416178 (http://www.wikidata.org/entity/Q104416178)	mouse endothelial cell	188
(http://www.wikidata.org/entity/Q104416140) wd:Q104416140 (http://www.wikidata.org/entity/Q104416140)	mouse fibroblast	176

Wikidata Query Service
(https://query.wikidata.org/#SELECT%20%3Fcell_type%20%3Fcell_typeLabel%20%28count%28%3Fmarker%2

Wikidata SPARQL queries enabled by the integration

Now that the PanglaoDB is released as Linked Open Data, we can make queries that were not possible before, including federated queries with other biological databases, such as Uniprot [31] and Wikipathways [32]. Due to previous similar reconciliation projects, Wikidata already contains information about genes, including their relations to Gene Ontology (GO) terms, something that led to the development of an R package, go2cell [33], that facilitates interconnection between cell types and GO terms via their markers.

PanglaoDB's integration to the Wikidata ecosystem allows us to ask a variety of questions. The next section headers exemplify such questions.

"Which human cell types are related to neurogenesis via their markers?"

As expected, the query below retrieved a series of neuron types, such as "human purkinje neuron" and "human cajal-retzius cell." It did, however, also retrieved non-neural cell types such as the "human loop of henle cell, a kidney cell type, and"human osteoblast. These seemingly unrelated cell types markedly express genes that are involved in neurogenesis, but that does not mean that they are involved with this process. This reinforces the idea that one needs to be careful when using curated pathways to enrich one's analysis, as false positives abound.

The molecular process that gene products take part depends on the cell type. The SPARQL query below enables us to seamlessly compare Gene Ontology processes with cell marker data, providing a fruitful sandbox for generation of hypothesis and exploration of the biomedical knowledge landscape.

Query for cell types related to neurogenesis

geneLabel	cellTypeLabel
EPHB1	human oligodendrocyte
EPHB1	human osteoclast
OMP	human purkinje neuron
OMP	human olfactory epithelial cell
OMP	human neuron
PCSK9	human delta cell
PCSK9	human loop of Henle cell
CXCR4	human b cell
CXCR4	human nk cell
CXCR4	human dendritic cell
CXCR4	human megakaryocyte
CXCR4	human t helper cell

Wikidata Query Service (https://query.wikidata.org/#SELECT%20%3FgeneLabel%20%3FcellTypeLabel%0AWHERE%20%0A%7B%0A%20%20%3Fprote

“Which cell types express markers associated to Parkinson`s disease?”

Besides integration with Gene Ontology, Wikidata reconciliation makes it possible to complement the marker gene info on PanglaoDB with information about diseases. This integration is of biomedical interest, as there is a quest for detailing of mechanisms that link genetic associations and the diseases themselves.

“Disease genes” are often compiled from Genomic Wide Association Studies, which look for sequence variation in the DNA. These studies are commonly blind to the cell types related to the pathophysiology of the disease. In the query below, we can see cell types that are marked by genes genetically associated with Parkinson’s disease. Even considering the false positives (as per the previously mentioned multifunctional nature of genes) this kind of overlook can aid domain experts to come up with novel hypothesis.

Query for cell types related to Parkinson’s disease

cellTypeLabel	geneLabel	diseaseLabel
human fibroblast	COL13A1	Parkinson's disease
human erythroid-like and erythroid precursor cell	SNCA	Parkinson's disease
human podocyte	MAPT	Parkinson's disease
human b cell	HLA-DRA	Parkinson's disease
human dendritic cell	HLA-DRA	Parkinson's disease
human monocyte	HLA-DRA	Parkinson's disease
human langerhans cell	HLA-DRA	Parkinson's disease
human smooth muscle cell	ITGA8	Parkinson's disease

Wikidata Query Service (https://query.wikidata.org/#SELECT%20%3FcellTypeLabel%20%3FgeneLabel%20%3FdiseaseLab%3E%20gene%0A%20%20%0A%20%20%3FcellType%20rdfs%3Alabel%20%3FcellTypeLabel.%0A%20%20wd%3AQ11085%2

Which diseases are associated with the markers of pancreatic beta cells?

We can check the cell-type to disease relation in both ways. Scientists that study specific cell types (and not necessarily specific diseases) might be interested in knowing which diseases are related to their cell type of interest. In the sample query below, I looked for the diseases linked to the human pancreatic beta cells, which play an important role in controlling blood sugar levels. Reassuringly, top hits associated with markers included obesity and type-2 diabetes. Other diseases retrieved, such as Huntington disease-like 2 don't bear a clear link with sugar function, and might merit a further look by a domain expert to see if there are any hypothesis worth pursuing.

Query for cell types related to Parkinson's disease

cellTypeLabel	diseaseLabel	count	genes
human beta cell	obesity	3	PCSK2, ADCYAP1, SLC30A8
human beta cell	type 2 diabetes	2	SLC30A8, TGFBR3
human beta cell	Parkinson's disease	1	SH3GL2
human beta cell	asthma	1	SLC30A8
human beta cell	aniridia	1	PAX6
human beta cell	rheumatoid arthritis	1	CD40
human beta cell	type-1 diabetes	1	PAX4
human beta cell	Optic nerve hypoplasia	1	PAX6
human beta cell	CD40 deficiency	1	CD40
human beta cell	MODY 6	1	NEUROD1

Wikidata Query Service (https://query.wikidata.org/#SELECT%20%3FcellTypeLabel%20%3FdiseaseLabel%20%0A%28COU
%3E%20%20gene%0A%20%20%3Fdisease%20wdt%3AP2293%20%3FdiseaseGene%20.%20%20%20%20%20%20%20%20%20%
%3E%20gene%0A%20%20%0A%20%20wd%3AQ101405087%20rdfs%3Alabel%20%3FcellTypeLabel%20.%0A%20%20%3Fdi

Improvement of Wikidata data on cell types

To reconcile a database to Wikidata, we need to match names on the databases, often in natural language, to the unique identifiers on Wikidata. We first employed an automatic approach based on Entities from PanglaoDB, that is, cell types,tissue types and organ types, were matched with Wikidata items, matching summary can be seen on Table 4.

Of note, Wikidata editors often mix first-order classes such as “cells” and “organs” with second-order classes like “cell types” and “organ types” (Supplementary Information). First-order classes point to real-world individuals, like the “Dolly sheep zygote” (a real-world “cell”) and the “brain of Albert Einstein” (a real-world “organ”). Second-order classes point to classes, like “zygote” (a conceptual “cell type”) and “brain” (a conceptual “organ type”).

Table 4: Summary of the matched entities from PanglaoDB (August 2020).

	PanglaoDB (count)	Automatic matches (count)
Cell types	215	81 (37.67 %)
Tissue types	246	85 (34.55 %)
Organ types	29	22 (75.86%)

The difference between first-order classes and second-order classes is notoriously tricky, and biological databases and texts often practice unintentional punning (the use of the same concept for different levels [34]). For adding markers on Wikidata, we assumed that all information in PanglaoDB was about instances of cell types, and not specific cells.

After marker data from PanglaoDB was added to Wikidata, we tested the automatic classification method was able to detect most cell types matches for most cell types on PanglaoDB matches (Table 5). The improvement of 38% to 80% of automatically matched types is an evidence that our work improved cell type content on Wikidata, and will arguably facilitate the reconciliation of other cell-type related resources.

Table 5: Summary of matched PanglaoDB entities after improvements were made (December 2020).

	PanglaoDB (count)	Automatic matches (count)
Cell types	215	173 (80.46 %)
Tissue types	246	63 (25.60 %)
Organ types	29	18 (62.06 %)

Noticeably, the proportion of automatic matches for other entity types (tissues and organs) seems reduced in relation to the first assessment (35% to 25% and 76 to 62%). These entities were not targeted by our work, but as Wikidata is a living resource, modifications in the database, such as reclassification of entities or adding of other similar concepts, may have reduced the performance of our simple reconciler.

Analysis of item quality - final look

As can be gathered from Figure 3, nearly all cell type items have the appropriate “instance of cell type” statement, with only 4 items still missing said statement and one item being classified as an “instance of gland”.

This is a considerable advance in improving the quality of cell type data in Wikidata, as having this simple statement will make these items easier to find and be expanded upon.

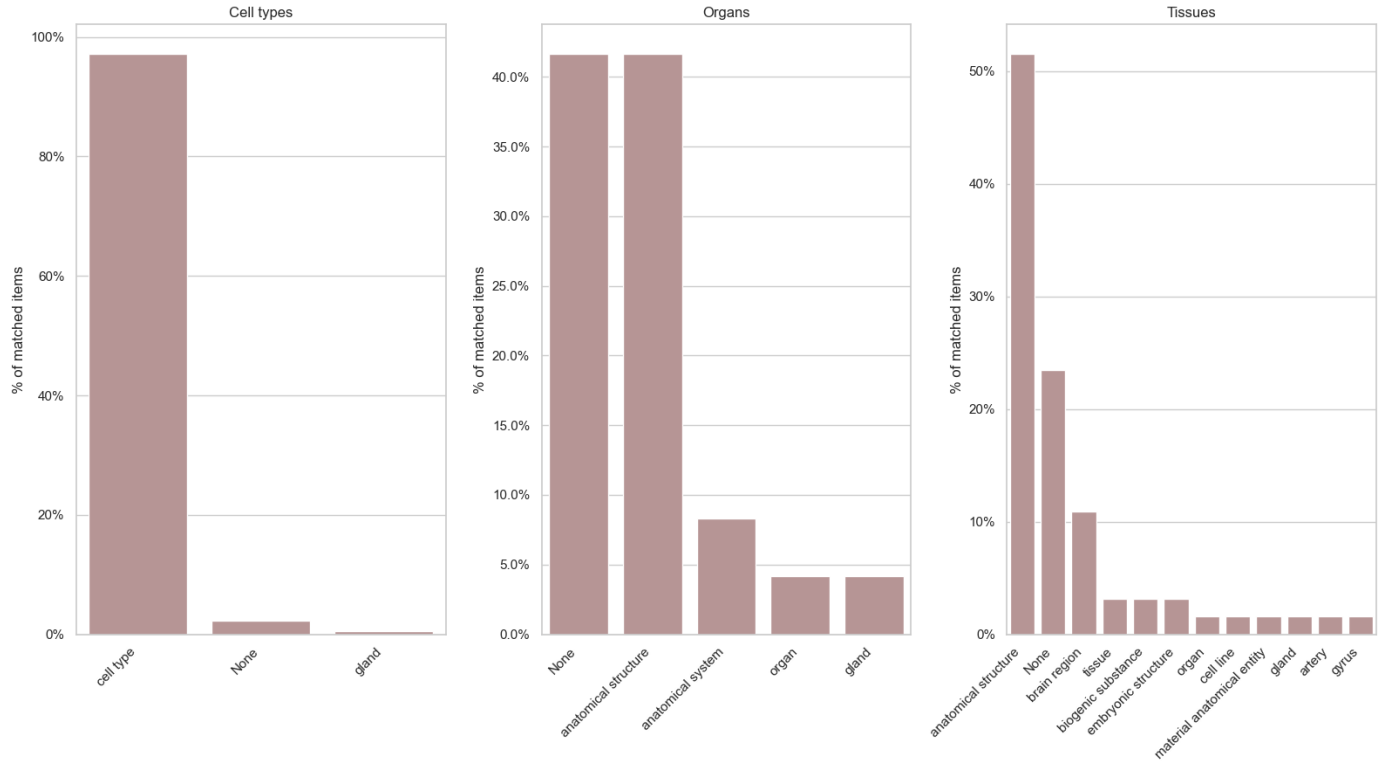

Figure 3: Percentage of reconciled entities gathered during the second and final reconciliation, divided by which item type they belong to.

Discussion

In this work, we re-released the knowledge curated in PanglaoDB on Wikidata, connecting it to the semantic web. Each cell-type/marker statement was added to Wikidata with a pointer to PanglaoDB and a citation of the article, providing proper provenance. At the same time, we documented the process of database integration to Wikidata, providing a blueprint for future efforts.

It is important to note that not all data on PanglaoDB was added to Wikidata. Fine-grained, database-specific details were too granular for a general-purpose database like Wikidata (e.g. the sensitivity and specificity attached to each marker-cell type pair). Eventhough, these data could be released in RDF format and be connected to independent SPARQL endpoints (as done in the Bio2RDF effort [35]), we focused on integration to Wikidata to take advantage of the built-in integration with various types of knowledge, as well as the tooling developed by the Wikidata community.

As described in the methods session, we added species-specific terms to Wikidata for cell types of *Homo sapiens* and *Mus musculus* described in the PanglaoDB database. The use of species-specific cell-types is necessary because genes in Wikidata are also species-specific, connected to their taxon by the "found in taxon" properties. In the biomedical literature, however, genes and cell types are sometimes referred to broadly, in a multi-species or species-neutral way. The fuzzy, humane meanings are not always compatible with formalized data models. Thus, the reconciliation endeavor is not merely finding the right match on Wikidata, but largely of crafting coherent interpretations of data.

The complexity of biomedical communication adds to the argument pro-Wikidata. Sometimes, as happened for us, it is just impossible to find a suitable term in an existing ontology. OBO Foundry ontologies are open to contribution, but require a large investment. For starters, one must learn a lot about description logic, a field that is often exotic for biologists and software developers alike. Moreover, to contribute, one needs to acquire the tooling. That includes learning to use GitHub (https://github.com/) and Protegé (https://protege.stanford.edu/), but also learning community conventions and social norms that are slightly different for every single ontology. Wikidata bypasses this steep learning curve by providing a web interface which requires little to no previous experience with ontologies and programming. The reconciliation process becomes smoother, as if a concept is not previously catalogued, we can add a new one on the fly.

Additionally, knowledge added to Wikidata is not locked in the ivory tower of academia. Data on Wikidata can be easily reused on Wikipedia, a major source of information for scientists and lay people alike. Wikipedia's thriving mutualism with academia is well documented. [36,37,38] Wikidata information can enhance the quality of articles about life-science subjects in semi-automated ways (as has been done before [39]). Thus, Wikidata is directly connected to the well-established science education platform of Wikipedia, a feature unrivaled by any other structured knowledge system.

Of course, Wikidata has its limitations. Concerns with the reliability of Wikipedia are as old as the encyclopedia itself (for a discussion, see https://en.wikipedia.org/wiki/Reliability_of_Wikipedia) and Wikidata likely shares many of such concerns. The ontological modelling on Wikidata is often far from perfect, and inconsistencies and logical mistakes abound. [40]. It has been argued, though, that bio-ontologies generally lack "strict, explicit and well defined semantics" (at least in 2008 [41]). While a comprehensive analysis of pros and cons of scientific Wikidata is not available, we extend Don Fallis' view on Wikipedia and argue that Wikidata has a number of "epistemic virtues (e.g., power, speed, and fecundity) that arguably outweigh any deficiency in terms of reliability." [42]

This work exemplifies the power of releasing Linked Open Data via Wikidata, and provides the biomedical community with the first semantically accessible, 5-star LOD dataset of cell markers, easily

reachable from Wikidata's SPARQL Query Service(https://query.wikidata.org/). The work also paves the way for Wikidata reconciling of other databases for cell-type markers, such as CellMarker [43], labome [44], CellFinder [12] and SHOGoiN/CELLPEDIA [45]). The approach we took here can in essence be applied to any knowledge set of public interest, providing a low-cost and low-barrier platform for sharing biocurated knowledge in gold standard format.

We hope that community will keep improving marker and overall biological content on Wikidata, and that the interlinked marker information will be helpful. We invite the reader to improve information on Wikidata for their favorite cell types, adding markers and a link to the reference works, and make ourselves available for aiding anyone interested in using or editing marker information on Wikidata.

General Ideas

Temporary file containing ideas for the project. Interesting references and concepts.

med2rdf[46] is a project to migrate biomedical knowledge bases to RDF format, facilitating integration with the semantic web.

15 years ago, in the original Cell Ontology paper, they mention the idea to integrate their knowledge with gene expression databases, something not done as far as we know [47]

References

1. **PanglaoDB - A Single Cell Sequencing Resource For Gene Expression Data**
   https://panglaodb.se/index.html

2. **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 · PMID: 30951143 · PMCID: PMC6450036

3. **Linked Data - Design Issues** https://www.w3.org/DesignIssues/LinkedData.html

4. **The OBO Foundry** http://www.obofoundry.org/

5. **Wikidata:Notability - Wikidata** https://www.wikidata.org/wiki/Wikidata:Notability

6. **Wikidata** https://www.wikidata.org/wiki/Wikidata:Main_Page

7. **Semantic Web - W3C** https://www.w3.org/standards/semanticweb/

8. **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
   Cold Spring Harbor Laboratory (2015-11-16) https://doi.org/gg9dk4
   DOI: 10.1101/031971

9. **Wikidata as a knowledge graph for the life sciences**
   Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M. Good, Malachi Griffith, Obi Griffith, Kristina Hanspers, Henning Hermjakob, Toby Hudson, Kevin Hybiske, ... Andrew I. Su
   eLife (2020-03-17) https://www.wikidata.org/wiki/Q87830400
   DOI: 10.7554/elife.52614

10. **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/ggqqc6
    DOI: 10.7554/elife.52614 · PMID: 32180547 · PMCID: PMC7077981

11. **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 · PMID: 31557529

12. **CellFinder: a cell data repository**
    Harald Stachelscheid, Stefanie Seltmann, Fritz Lekschas, Jean-Fred Fontaine, Nancy Mah, Mariana Lara Neves, Miguel A. Andrade-Navarro, Ulf Leser, Andreas Kurtz

Nucleic Acids Research (2013-12-03) https://www.wikidata.org/wiki/Q28660708
DOI: 10.1093/nar/gkt1264

13. **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, … Chris Mungall
Journal of Biomedical Semantics (2016-07-04) https://www.wikidata.org/wiki/Q36067763
DOI: 10.1186/s13326-016-0088-7

14. **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 · PMID: 27377652 · PMCID: PMC4932724

15. **The Human Cell Atlas.**
Aviv Regev, Sarah Teichmann, Eric Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, … Human Cell Atlas Meeting Participants
eLife (2017-12-05) https://www.wikidata.org/wiki/Q46368626
DOI: 10.7554/elife.27041

16. https://query.wikidata.org/

17. **oscar-franzen/PanglaoDB**
Oscar Franzén
(2020-09-02) https://github.com/oscar-franzen/PanglaoDB

18. **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

19. **mwaskom/seaborn: v0.11.0 (Sepetmber 2020)**
Michael Waskom, Olga Botvinnik, Maoz Gelbart, Joel Ostblom, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, Jordi Warmenhoven, … Thomas Brunner
Zenodo (2020-09-08) https://doi.org/ghcq2j
DOI: 10.5281/zenodo.4019146

20. **matplotlib/matplotlib: REL: v3.3.2**
Thomas A Caswell, Michael Droettboom, Antony Lee, John Hunter, Elliott Sales De Andrade, Eric Firing, Tim Hoffmann, Jody Klymak, David Stansby, Nelle Varoquaux, … Paul Ivanov
Zenodo (2020-09-15) https://doi.org/ghcq2k
DOI: 10.5281/zenodo.4030140

21. **reconciler: Python utility to reconcile Pandas DataFrames**
João Vitor F. Cavalcante
https://github.com/jvfe/reconciler

22. **OpenRefine-Wikidata interface** https://wikidata.reconci.link/

23. **Reconciliation Service API v0.1** https://reconciliation-api.github.io/specs/0.1/

24. **Natural language processing with Python**
Steven Bird, Ewan Klein, Edward Loper
O'Reilly (2009)
ISBN: 9780596516499

25. **API:REST API - MediaWiki** https://www.mediawiki.org/wiki/API:REST_API

26. **Ensembl 2020**
Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, 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 · PMID: 31691826 · PMCID: PMC7145704

27. **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 · PMID: 23193264 · PMCID: PMC3531099

28. **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 · PMID: 22293552 · PMCID: PMC3334586

29. **CELDA – an ontology for the comprehensive representation of cells in complex systems**
Stefanie Seltmann, Harald Stachelscheid, Alexander Damaschun, Ludger Jansen, Fritz Lekschas, Jean-Fred Fontaine, Throng Nghia Nguyen-Dobinsky, Ulf Leser, Andreas Kurtz
BMC Bioinformatics (2013-07-17) https://www.wikidata.org/wiki/Q21284308
DOI: 10.1186/1471-2105-14-228

30. **SuLab/WikidataIntegrator**
Su Lab
(2021-01-24) https://github.com/SuLab/WikidataIntegrator

31. **UniProt** https://sparql.uniprot.org/sparql

32. **Portal:Semantic Web - WikiPathways**
https://www.wikipathways.org/index.php/Portal:Semantic_Web

33. **jvfe/go2cell**
João Vitor
(2021-01-02) https://github.com/jvfe/go2cell

34. **Punning - OWL** https://www.w3.org/2007/OWL/wiki/Punning

35. **Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data**
Alison Callahan, José Cruz-Toledo, Peter Ansell, Michel Dumontier, Michel Dumontier
Lecture Notes in Computer Science (2013-01-01) https://www.wikidata.org/wiki/Q56989268
DOI: 10.1007/978-3-642-38288-8_14

36. **Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial**
Neil Thompson, Douglas Hanley
(2017-09-20) https://www.wikidata.org/wiki/Q42013239
DOI: 10.2139/ssrn.3039505

37. **The Gene Wiki in 2011: community intelligence applied to human gene annotation**
Benjamin M. Good, Erik L. Clarke, Luca de Alfaro, Andrew I. Su
Nucleic Acids Research (2012-01-01) https://www.wikidata.org/wiki/Q21629969
DOI: 10.1093/nar/gkr925

38. **Ten simple rules for editing Wikipedia**
Darren W. Logan, Massimo Sandal, Paul P. Gardner, Magnus Manske, Alex Bateman
PLOS Computational Biology (2010-01-01) https://www.wikidata.org/wiki/Q21145331
DOI: 10.1371/journal.pcbi.1000941

39. **Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages: a pilot study**
Alexander Pfundner, Tobias Schönberg, John Horn, Richard David Boyce, Matthias Samwald
Journal of Medical Internet Research (2015-05-05) https://www.wikidata.org/wiki/Q21503276
DOI: 10.2196/jmir.4163

40. **Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata**
Freddy Brasileiro, João Paulo A. Almeida, Victorio A. Carvalho, Giancarlo Guizzardi
Proceedings of the 25th International Conference Companion on World Wide Web (2016-04-01)
https://www.wikidata.org/wiki/Q27037396
DOI: 10.1145/2872518.2891117

41. **Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology**
Mikel Egaña Aranguren, Erick Antezana, Martin Kuiper, Robert Stevens
BMC Bioinformatics (2008-01-01) https://www.wikidata.org/wiki/Q21093639
DOI: 10.1186/1471-2105-9-s5-s1

42. **Toward an epistemology of Wikipedia**
Don Fallis
Journal of the Association for Information Science and Technology (2008-08-01)
https://www.wikidata.org/wiki/Q101955295
DOI: 10.1002/asi.20870

43. **CellMarker: a manually curated resource of cell markers in human and mouse**
Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, ... Yun Xiao
Nucleic Acids Research (2019-01-01) https://www.wikidata.org/wiki/Q56984510
DOI: 10.1093/nar/gky900

44. **Cell Markers**
Konstantin Yakimchuk
Materials and Methods (2013-05-02) https://doi.org/ghq494
DOI: 10.13070/mm.en.3.183

45. **SHOGoiN: Shogoin Human Omics database for the Generation of iPS and Normal cells**
https://stemcellinformatics.org/

46. **MED2RDF**
website
http://med2rdf.org/

47.**:(unav)**
Jonathan Bard, Seung Y Rhee, Michael Ashburner
Genome Biology (2005) https://doi.org/dfxc74
DOI: 10.1186/gb-2005-6-2-r21 · PMID: 15693950 · PMCID: PMC551541

48. **Wikidata as a semantic framework for the Gene Wiki initiative**
Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mitraka, Julia Turner, Tim Putman, Justin Leong, Chinmay Naik, Paul Pavlidis, Lynn Schriml, Benjamin M Good, Andrew I Su
Database (2016-03-17) https://doi.org/f9bbk9
DOI: 10.1093/database/baw015 · PMID: 26989148 · PMCID: PMC4795929

Supplementary text and figures

Only *Homo sapiens* genes and Organs reconciled more than 50%. In the case of genes, this is probably due to the Gene Wiki initiative [48], a long-running project to improve biological information in Wikipedia and its sister-projects, including Wikidata.

This is further illustrated by Figure 5, in which we can see that all *Mus musculus* gene items - and nearly all *Homo sapiens* items - analysed had the Entrez ID alternative identifier present. Most of the data from the Gene Wiki project came from NCBI, creator and maintainer of Entrez. Nevertheless, there are still many gene items without an “Ensembl Gene ID” property, showcasing the need for further work in migrating this important source of information.
In the case of Organ data, there was a high number of matches both due to the fact that there were only a few number of items, but also since most Organ entities have Wikipedia pages, that are, therefore, cross-linked using Wikidata, requiring the creation of these items.

Regarding alternative identifiers, what was observed for genes cannot be said for histological entities. While there is significant progress in integrating UBERON IDs, there is near to no items with a Cell Ontology ID property (Figure 4).

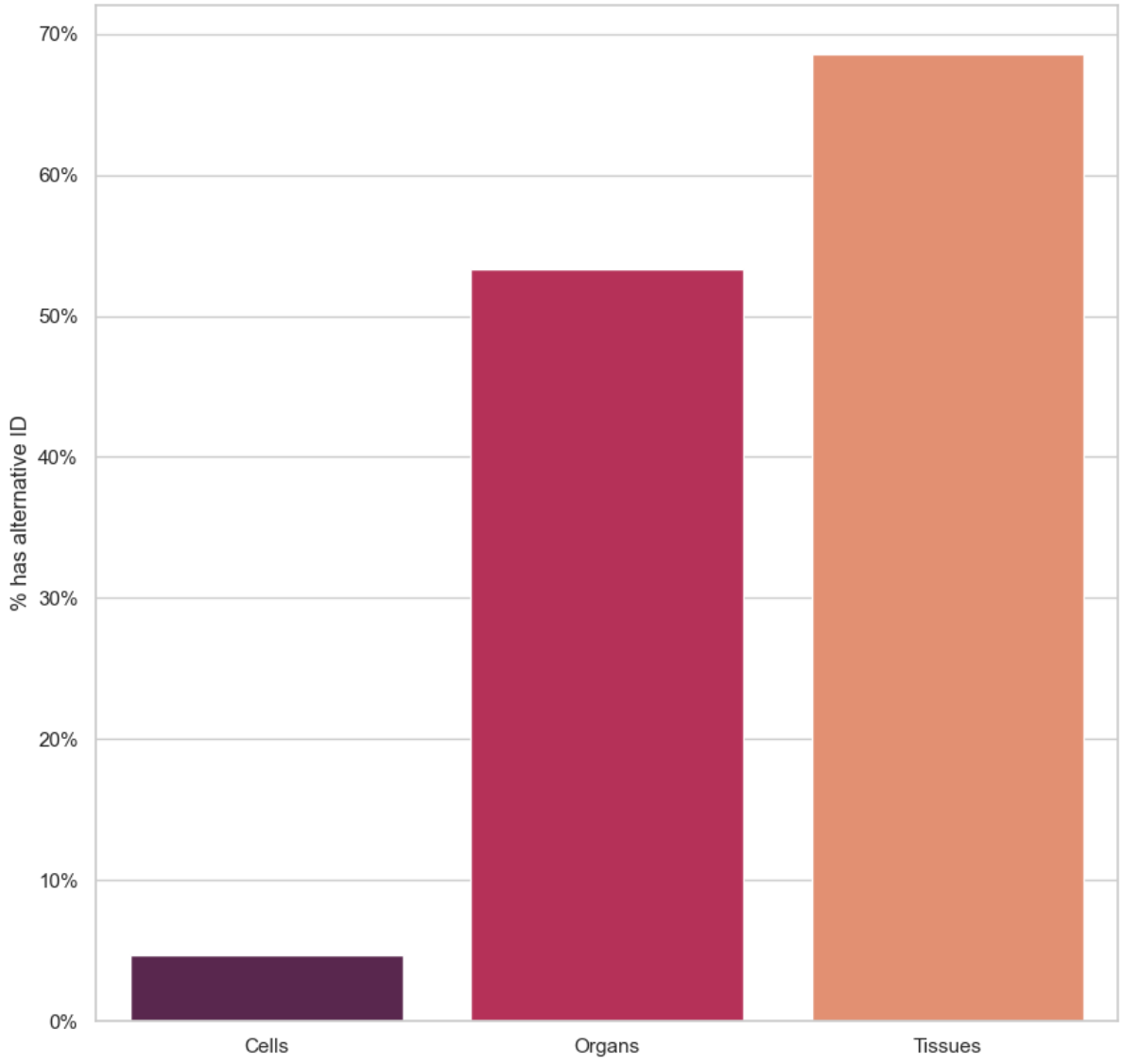

Figure 4: Percentage of matched histological items that had alternative identifiers, UBERON IDs for Tissues and Organs, Cell Ontology IDs for Cell types.

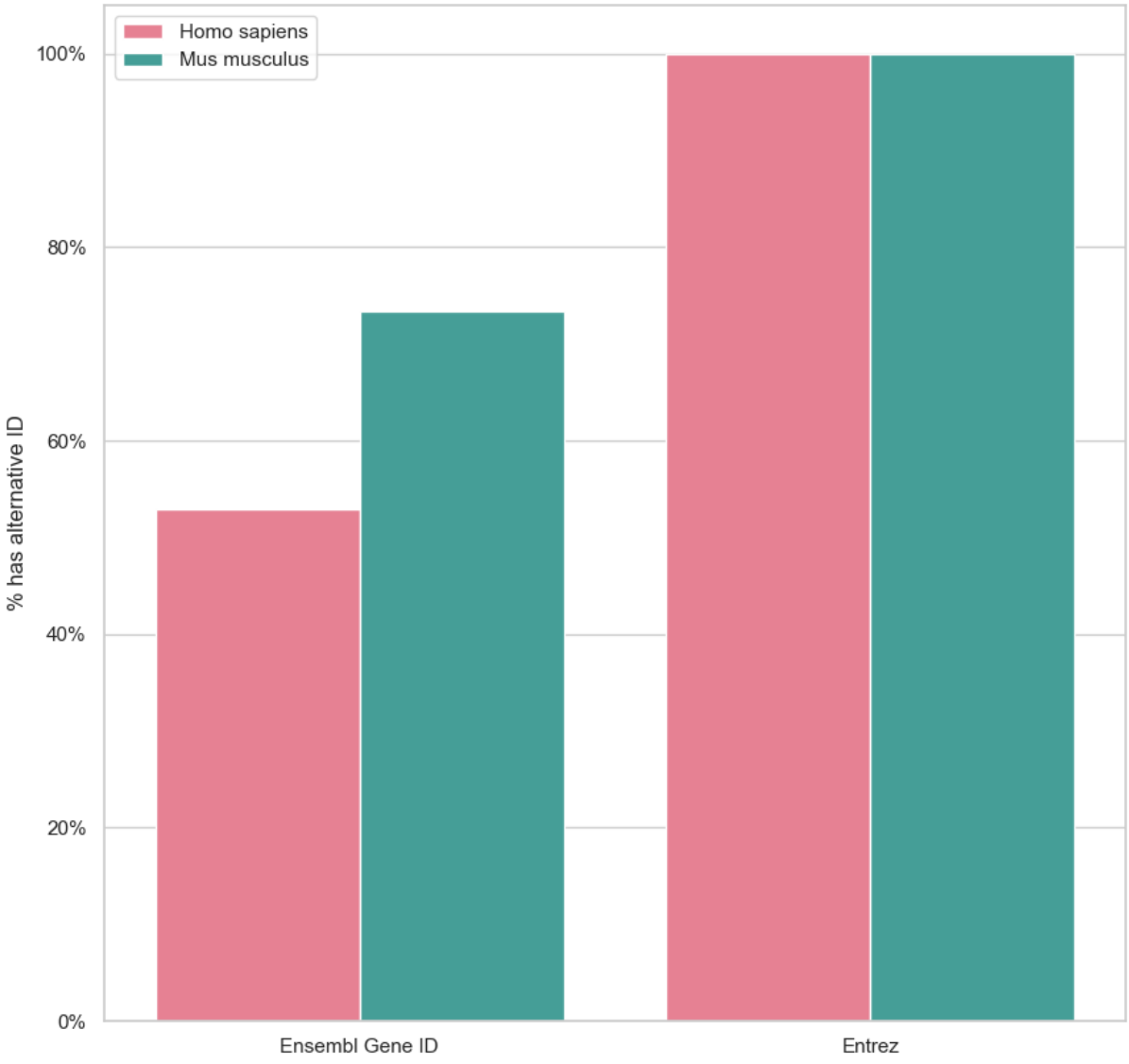

Figure 5: Percentage of matched gene items that had alternative identifiers, Entrez ID and Ensembl Gene ID, divided by species.

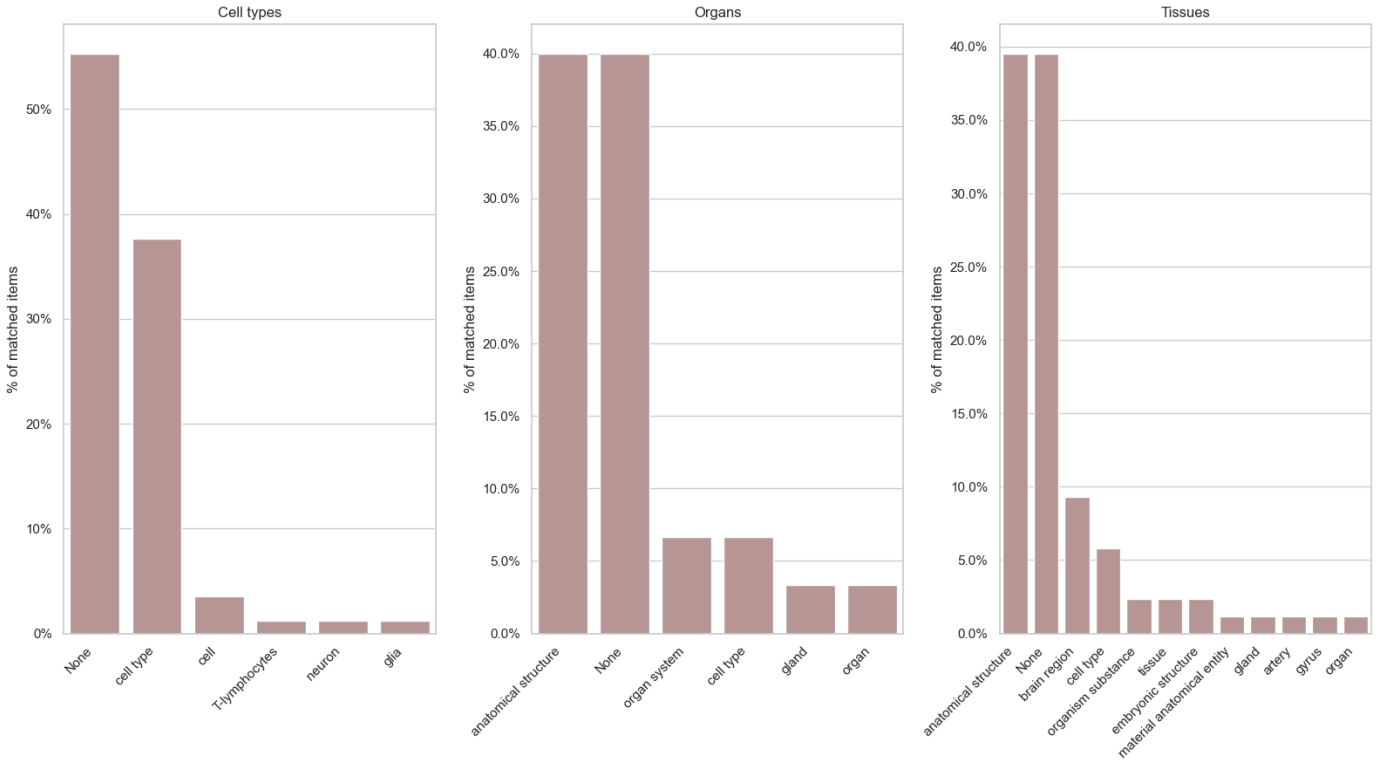

Figure 6: Percentage of reconciled entities, divided by which item type they belong to. Most reconciled items don't count with the P31 property.

A significant proportion of the matches we could acquire for histological data didn't contain in their data model an "instance of" (P31) property, this illustrates an extremely concerning fact: Although we could still match around 30 percent of the data - in the case of Cell types and Tissues - this data was probably "low-quality", that is, hard to find and even harder to obtain insights from, we can affirm this since the P31 property is the basis for most items in Wikidata, it's the most intuitive way to perform queries against their database and to annotate their items.

Furthermore, there is a significant disparity between histological data and gene data: while we could only match around 37% of Cell types from PanglaoDB, and of those 55% didn't have P31, we matched 60% of *Homo sapiens* genes, and all of them had P31. This disparity is not clearly shown when looking exclusively at the number of statements for these items (Figures 7 and 8), but it shows there is still a great amount of missing information for biological data, in particular in regards to cell types.

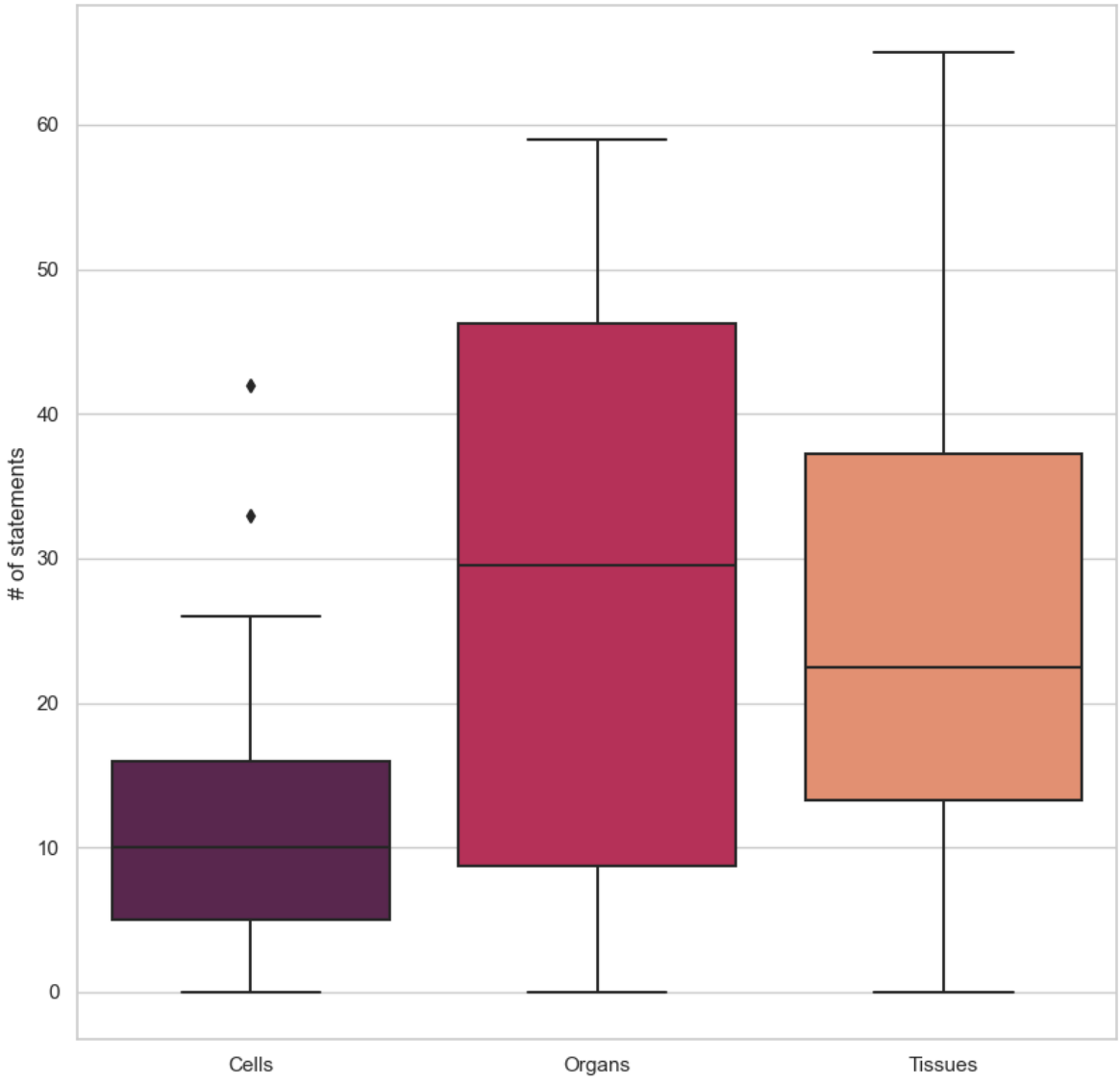

Figure 7: The distribution of the number of statements of the matched histological entities. Cell types performed the lowest.

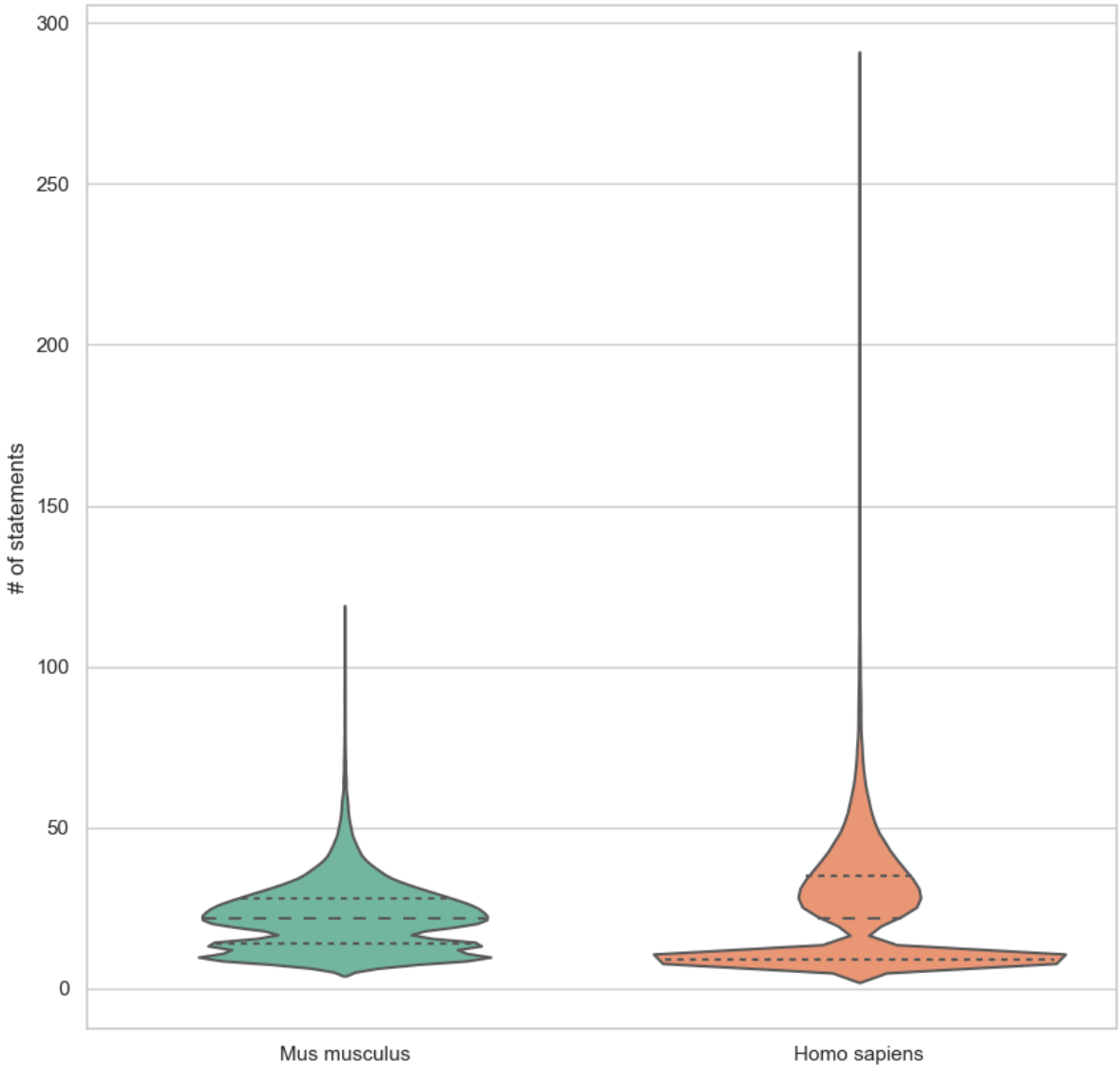

Figure 8: The distribution of the number of statements for matched gene items, divided by species.