

# 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@a82a02c](#) on December 27, 2020.

## 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 openness 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 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 [1](#) and [2](#)). It also provides 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

Despite its usefulness for the community, the database is on a 3-star category for Linked Open Data [\[3\]](#) as it does not use open standards from W3C (RDF and SPARQL). To make it 5-star, it needs to be also linked to external data via common identifiers.

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.

**hepatocyte** (Q827450) **Item Identifier**

**Item Label**  
liver cell type  
Hepatocyte | hepatocytes

**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
<p><b>instance of</b> <b>cell type</b> <b>Value</b></p> <p><b>Property</b></p> <p>0 references</p> <p><a href="#">edit</a></p> <p><a href="#">+ add reference</a></p> <p><a href="#">+ add value</a></p>

**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 its potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [9]. 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 [10].

Wikidata, however, is a work in progress, and might need extensive improvement. For example, 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"[11] (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>), illustrating the absence of any formal data model.

**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 work has the dual goal of re-releasing PanglaoDB in a 5-star Linked Open Data Format and improving the modelling of the necessary concepts on Wikidata.

## Methodology

### Data acquisition

Gene data from Wikidata was acquired using the Wikidata Query Service [12] - <https://w.wiki/bWc> for *Homo sapiens* genes and <https://w.wiki/bWe> for *Mus musculus* genes.

Data for quality assessment from PanglaoDB was acquired through their metadata database dump repository[13].

The markers dataset was downloaded manually from PanglaoDB’s website ([https://panglaoDB.se/markers/PanglaoDB\\_markers\\_27\\_Mar\\_2020.tsv.gz](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[14] library, with the Seaborn[15] and Matplotlib[16] libraries being used for plotting.

### Reconciliation and matching

The metadata from PanglaoDB on cell types, tissues (including germ layers) and organs was matched to Wikidata items using the reconciler[17] library, further matching was done using a custom stemming function on the item labels, via PorterStemmer from the NLTK library [18]. 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 [14] 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 [19] 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[20] (P594) and Entrez Gene[21] (P351) IDs for genes, Cell Ontology[11] (P7963) IDs for cell types and Uberon[22] (P1554) IDs for organs and tissues.

## Class creation on Wikidata

Different from property creation, class creation on Wikidata does not require community approval, and any user can create new classes and add statements.

Species-neutral cell types were already mostly present on Wikidata. species-specific cell types were created for each human-specific cell type mentioned in PanglaoDB. Class labels and “subclass of” statements (<https://www.wikidata.org/wiki/Q21514624>) were added to a spreadsheet and uploaded to Wikidata via the batch edition tool Quickstatements (<https://quickstatements.toolforge.org/#/>).

## Integration to Wikidata

The reconciled dataset was uploaded to Wikidata via the Wikidata Integrator python package (<https://github.com/SuLab/WikidataIntegrator>), 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](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.

# Results

## Wikidata reconciliation - initial look

---

Entities from PanglaoDB, that is, cell types, genes, tissue types and organs, were matched with Wikidata items, matching summary can be seen on Table 4.

**Table 4:** Summary of the matched entities from PanglaoDB.

	# of total items	# of unique matches	% of total items that were matched
Cells	215	81	37.67%

	# of total items	# of unique matches	% of total items that were matched
Tissues	246	85	34.55%
Organs	29	22	75.86%
Human Genes	58216	35423	60.84%
Mouse Genes	53793	25124	46.70%

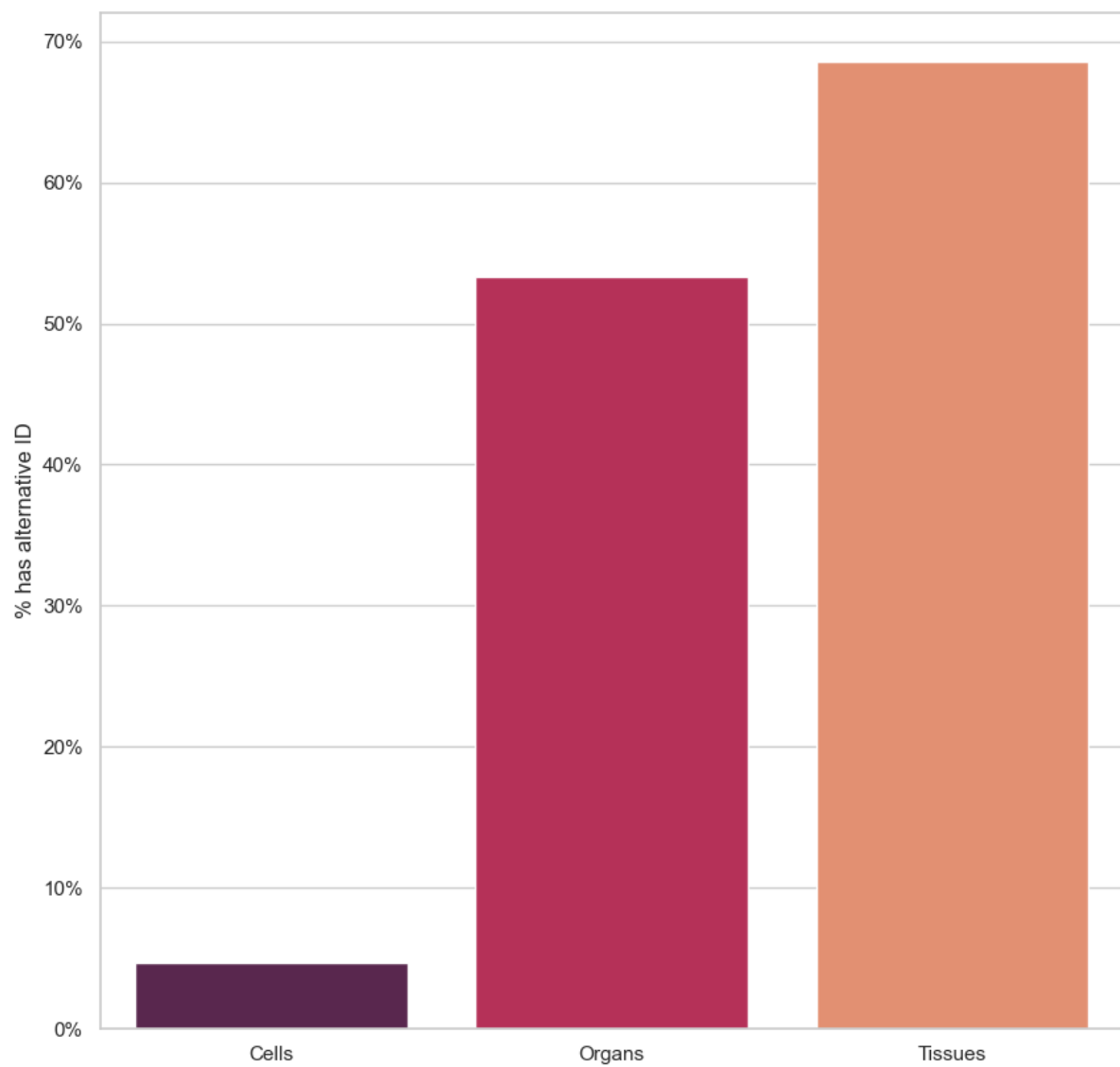
## Analysis of item quality - initial look

Only *Homo sapiens* genes and Organs reconciled more than 50%. In the case of genes, this is probably due to the Gene Wiki initiative [23], a long-running project to improve biological information in Wikipedia and its sister-projects, including Wikidata.

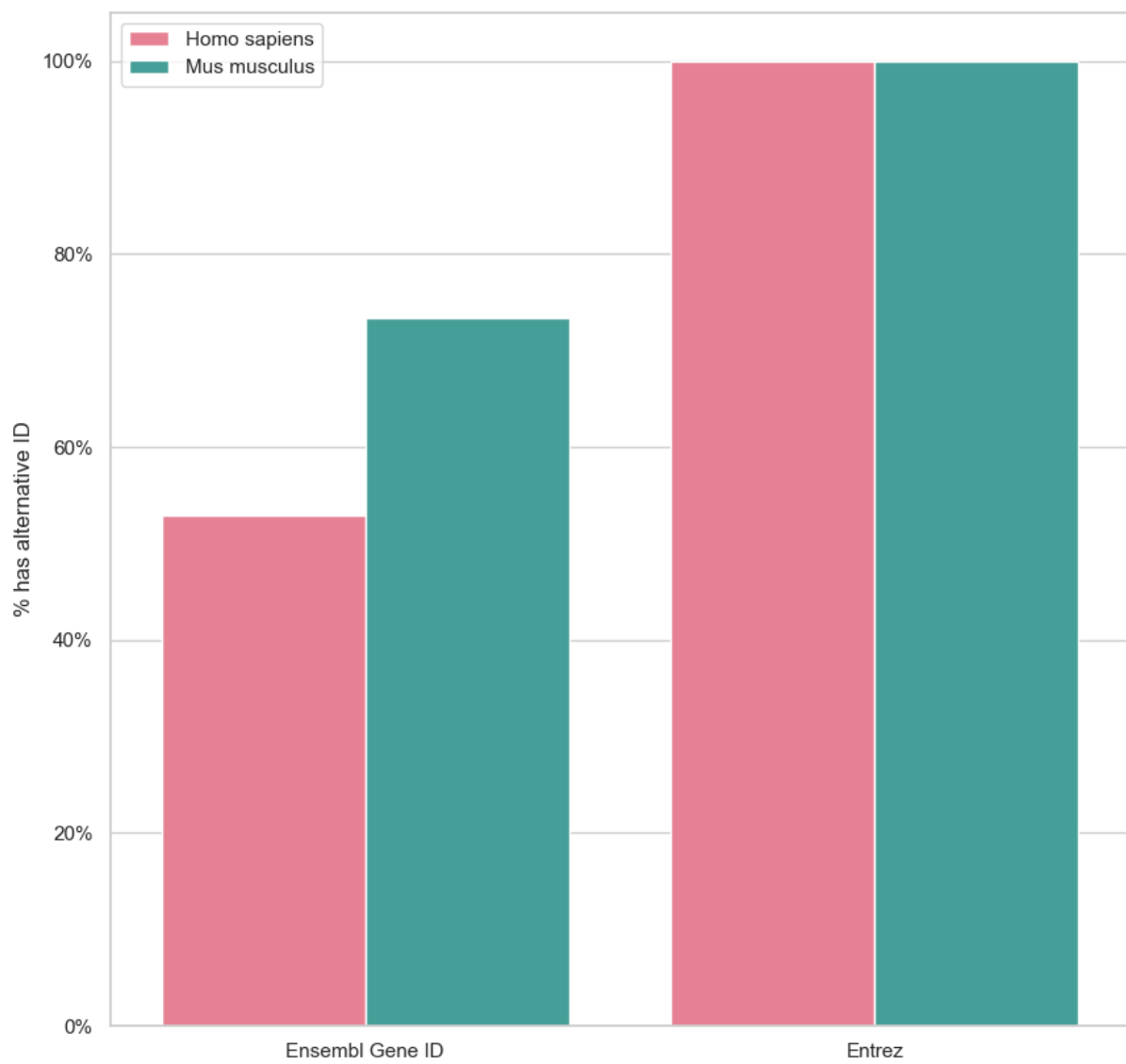
This is further illustrated by Figure 3, in which we can see that all *Mus musculus* gene items - and nearly all *Homo sapiens* items - analysed had the Entrez ID alternative identifier present - Most of the data from the Gene Wiki project came from NCBI, creator and maintainer of Entrez. Nevertheless, 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 2).

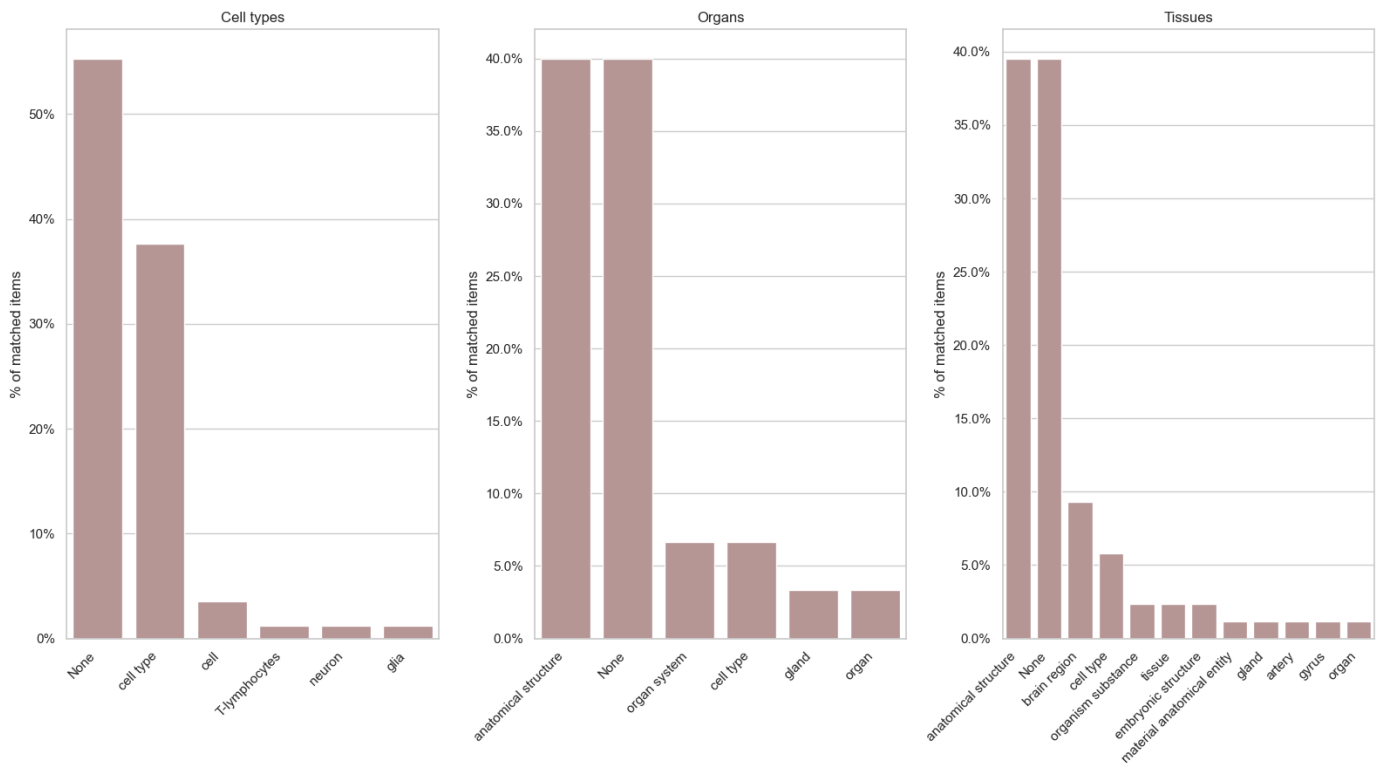


**Figure 2:** Percentage of matched histological items that had alternative identifiers, UBERON IDs for Tissues and Organs, Cell Ontology IDs for Cell types.



**Figure 3:** Percentage of matched gene items that had alternative identifiers, Entrez ID and Ensembl Gene ID, divided by species.

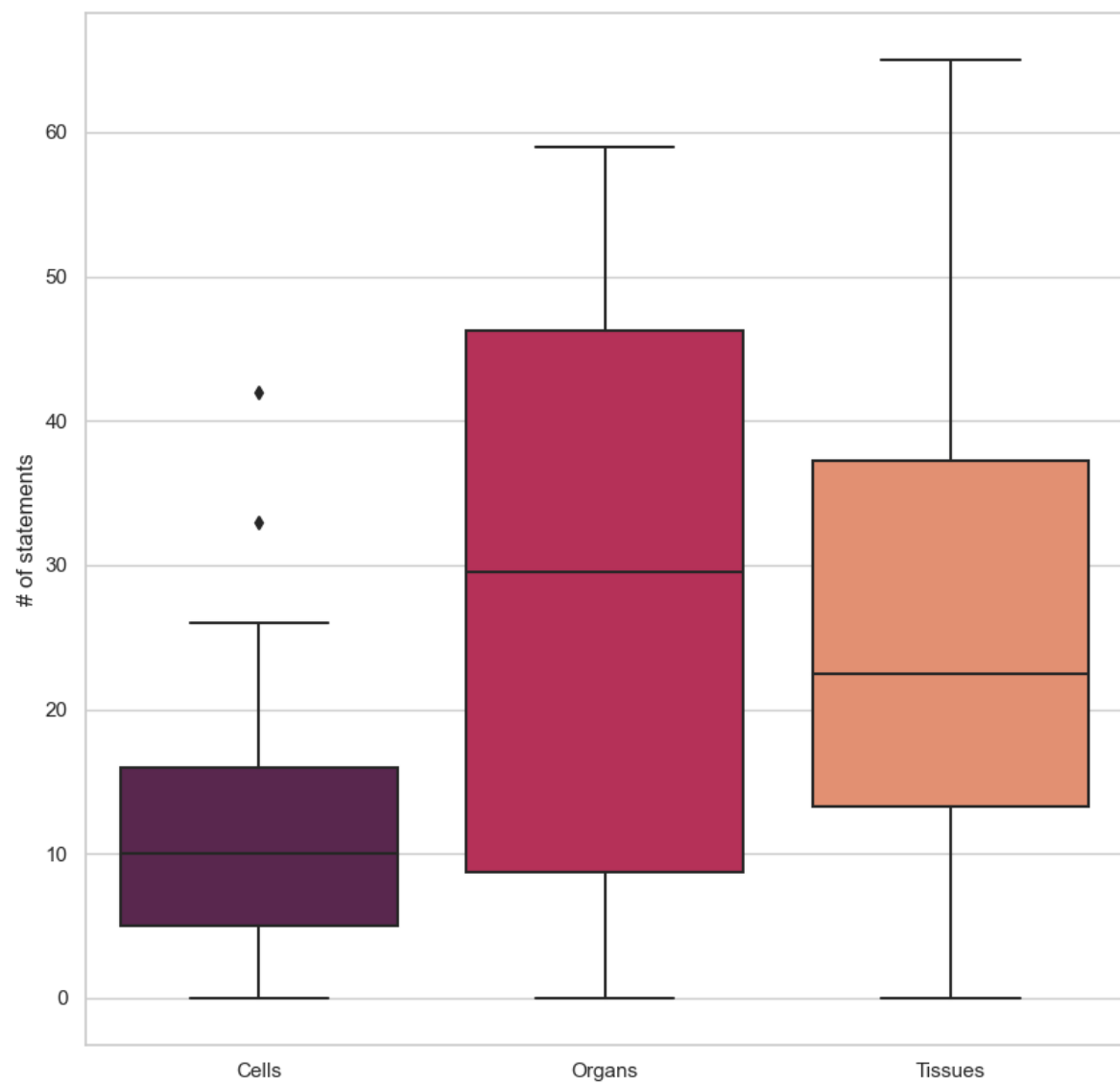




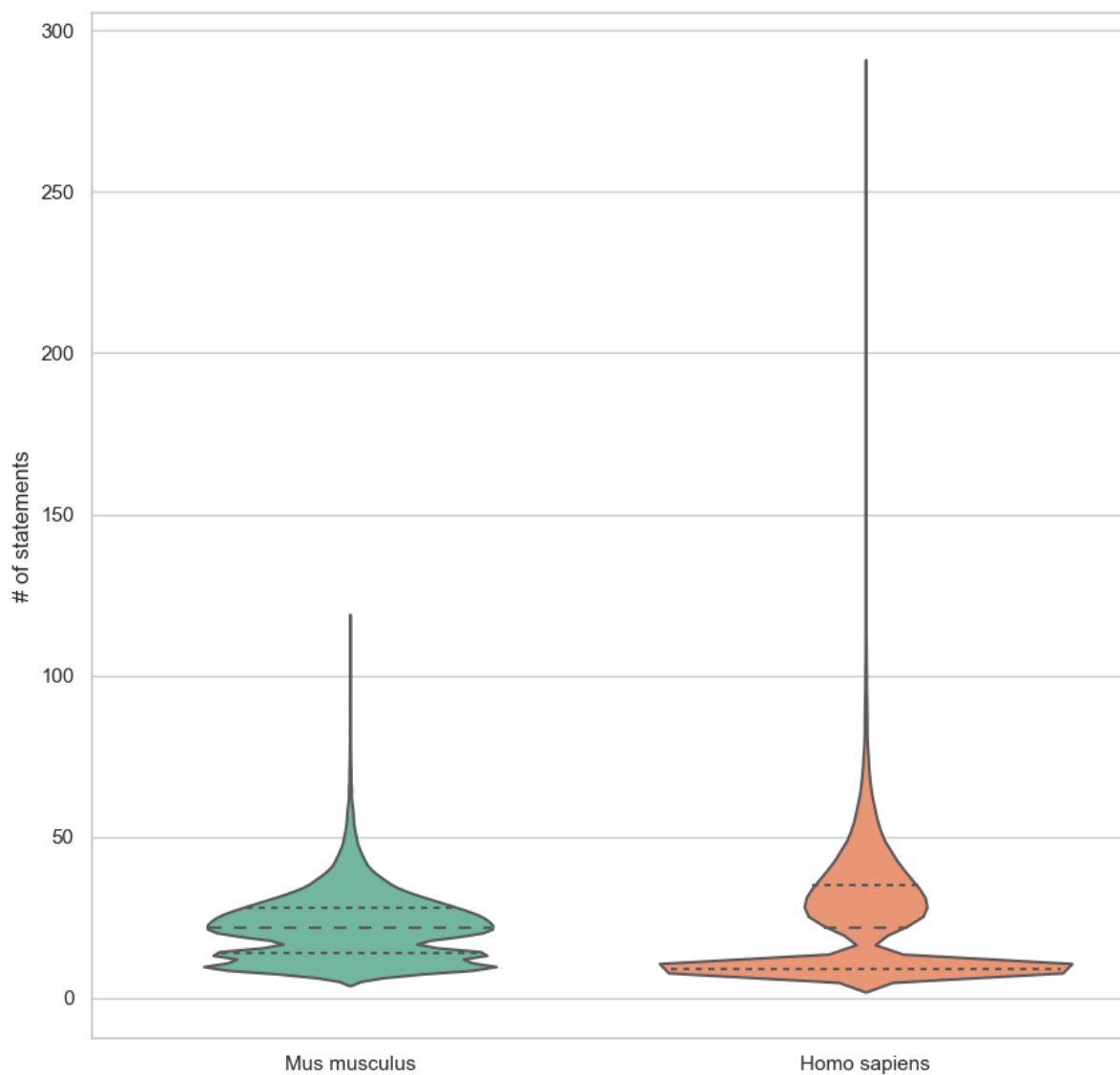
**Figure 4:** Percentage of reconciled entities, divided by which item type they belong to. Most reconciled items don't count with the P31 property.

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 5 and 6), but it shows there is still a great amount of missing information for biological data, in particular in regards to cell types.



**Figure 5:** The distribution of the number of statements of the matched histological entities. Cell types performed the lowest.



**Figure 6:** The distribution of the number of statements for matched gene items, divided by species.

## Improving Wikidata

---

- Adding species specific terms

### Adding a new property

### Adding missing items

- TBD

### Improving interoperability

- TBD

# Wikidata reconciliation - final look

After the aforementioned improvements were made, data from PanglaoDB was reconciled once again, now most cell types had appropriate matches (Table 5).

**Table 5:** Summary of matched PanglaoDB entities after improvements were made.

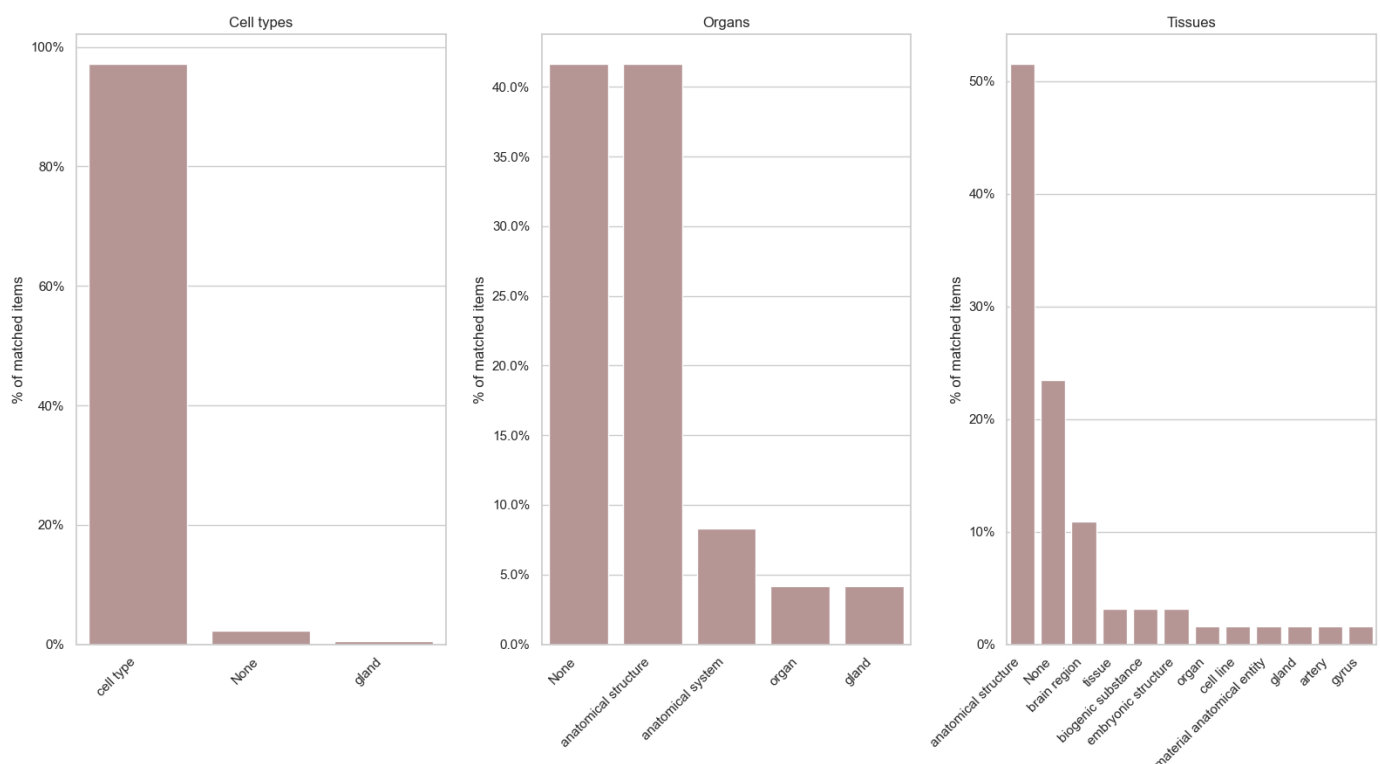
	# of total items	# of unique matches	% of total items that were matched
Cells	215	173	80.4651
Tissues	246	63	25.6097
Organs	29	18	62.0689
Human Genes	58216	35423	60.8475
Mouse Genes	53793	25124	46.705

While it may seem that the information for other entity types may have decreased, such as tissues and organs, this is difficult to ascertain, as different items could have been merged for clarity or were reclassified as belonging to different types not covered by this study.

## Analysis of item quality - final look

As can be gathered from Figure 7, 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 7:** Percentage of reconciled entities gathered during the second and final reconciliation, divided by which item type they belong to.

# 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. Due to previous similar reconciliation projects, Wikidata contains already information about genes, including their relations to Gene Ontology terms. The PanglaoDB 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
OMP	human purkinje neuron
OMP	human olfactory epithelial cell
OMP	human neuron
RET	human enteric neuron
EPHB1	human oligodendrocyte
EPHB1	human osteoclast
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

**“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

**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, TGFB3
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

- Reminder of what we did.
- Limitations of our work
  - Details of the PanglaoDB which were not included
- Future prospects related to markers
  - Integration of other resources, like CellMarkerDB

Linking biological with Wikidata allows out-of-the-box integrative SPARQL queries, as many biomedical ontologies and datasets have been already integrated to Wikidata, and are available in Wikidata's graph. Besides the well-known advantages of having data linked to the Linked Open Data cloud, the Wikidata integration provides user-friendly interfaces for the data. That includes both navigable html pages of classes and properties (e.g. <https://www.wikidata.org/wiki/Q67801129>) as well as an SPARQL Query Service with user-friendly modifications to ease queries for beginners (<https://query.wikidata.org/>) with helper pages for learning SPARQL ([https://www.wikidata.org/wiki/Wikidata:SPARQL\\_tutorial](https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial)) or even requesting queries ([https://www.wikidata.org/wiki/Wikidata:Request\\_a\\_query](https://www.wikidata.org/wiki/Wikidata:Request_a_query)).

In addition to user-friendly data access systems, Wikidata makes it easy for users to contribute. This user-friendliness is specially important in the case of the biomedical sciences, where database curation is becoming increasingly challenging with the growth of scientific publications. Wikidata allows editions directly in the Graphical User Interface, which makes it accessible for domain experts with little to no experience with programming and formal ontological representations. The Wikidata community has developed wrappers for the API in web applications that further facilitate contribution, such as the Quickstatements tool (<https://quickstatements.toolforge.org/#/>) for general purpose statements. The python module Wikidata Integrator facilitates for python users to reconcile databases to Wikidata, and it has been used to build bots for several different biological databases [wikidata:Q87830400].

This work exemplifies the power of releasing Linked Open Data via Wikidata, and provides the biomedical community with the first (to my knowledge) semantically accessible, 5-star LOD dataset. I hope that community will keep improving marker content on Wikidata, and that the interlinked marker information will be useful for researchers all over the world.

## General Ideas

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

med2rdf[[24](#)] 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 [[25](#)]



# 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](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/)
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](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 Mitranka, 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](https://doi.org/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 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](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/)
10. **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/)
11. **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/)
12. <https://query.wikidata.org/>
13. **oscar-franzen/PanglaoDB**  
Oscar Franzén

(2020-09-02) <https://github.com/oscar-franzen/PanglaoDB>

14. **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)

15. **mwaskom/seaborn: v0.11.0 (September 2020)**

Michael Waskom, Olga Botvinnik, Maoz Gelbart, Joel Ostblom, Paul Hobson, Saulius Lukauskas, David C Gempertline, Tom Augspurger, Yaroslav Halchenko, Jordi Warmenhoven, ... Thomas Brunner  
*Zenodo* (2020-09-08) <https://doi.org/ghcq2j>  
DOI: [10.5281/zenodo.4019146](https://doi.org/10.5281/zenodo.4019146)

16. **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](https://doi.org/10.5281/zenodo.4030140)

17. **reconciler: Python utility to reconcile Pandas DataFrames**

João Vitor F. Cavalcante  
<https://github.com/jvfe/reconciler>

18. **Natural language processing with Python**

Steven Bird, Ewan Klein, Edward Loper  
*O'Reilly* (2009)  
ISBN: [9780596516499](https://www.isbn.org/9780596516499)

19. **API:REST API - MediaWiki** [https://www.mediawiki.org/wiki/API:REST\\_API](https://www.mediawiki.org/wiki/API:REST_API)

20. **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](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/)

21. **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/)

22. **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/)

23. **Wikidata as a semantic framework for the Gene Wiki initiative**

Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mittraka, 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](https://doi.org/10.1093/database/baw015) · PMID: [26989148](https://pubmed.ncbi.nlm.nih.gov/26989148/) · PMCID: [PMC4795929](https://pubmed.ncbi.nlm.nih.gov/PMC4795929/)

24. **MED2RDF**

website

<http://med2rdf.org/>

25.:**(unav)**

Jonathan Bard, Seung Y Rhee, Michael Ashburner

*Genome Biology* (2005) <https://doi.org/dfxc74>

DOI: [10.1186/gb-2005-6-2-r21](https://doi.org/10.1186/gb-2005-6-2-r21) · PMID: [15693950](https://pubmed.ncbi.nlm.nih.gov/15693950/) · PMCID: [PMC551541](https://pubmed.ncbi.nlm.nih.gov/PMC551541/)