



Releasing PanglaoDB human cell-type markers to Wikidata

This manuscript ([permalink](#)) was automatically generated from [lubianat/semantic web course report@3dedc09](#) on December 7, 2020.

Authors

- **Tiago Lubiana**

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

Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil

Abstract

Introduction

PanglaoDB [1] [2] is a public database that contains data and metadata on different types of cells. The types of cells are associated with marker genes, which are used to identify the classes that best fit cells observed in biomedical experiments. PanglaoDB, specifically, derives its marker genes from the curation of several single-cell RNA sequencing experience.

The database is used for scientists when analyzing RNA-sequencing data to help in identification of the cells in a sample. Despite its usefulness for the community, the database is only 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].

In this work, I created classes on Wikidata for human-specific cell types, as well as an object property for linking cell type classes to gene classes. Then, I proceeded to reconciled the human cell-type / marker relations on PanglaoDB to Wikidata , and uploaded the PanglaoDB dataset as Linked Open Data directly to Wikidata via its Application Programming Interface. Finally, I show how this upload now enables SPARQL queries to Wikidata's endpoint that extend the usefulness of the Panglao database.

Data source


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 scope of this work, only the columns `species`, `official gene symbol` and `cell type` were used.

Property creation on Wikidata

To represent the cell-type -> marker property, I proposed a property for the Wikidata community. In Wikidata:Property proposal (https://www.wikidata.org/wiki/Wikidata:Property_proposal/has_positive_marker), I posted a message in 17th of November presenting the property, domain and range constraints, as well as additional comments. The html of the property proposal is reproduced below:

-

 Done: [has marker](#) (P8872) ([Talk and documentation](#))

Description	a gene or a protein published as a marker of a species-specific cell type
--------------------	---

Representations	Marker gene (Q2776413) (partially)
Data type	Item
Domain	?subject instance of (P31) cell type (Q189118) . ?subject found in taxon (P703) ?taxon . ?taxon taxon rank (P105) species (Q7432) .
Allowed values	{?object instance of (P31) protein (Q8054) .} UNION {?object instance of (P31) gene (Q7187) .} UNION {?object instance of (P31) macromolecular complex (Q22325163) .}
Example 1	human astrocyte (Q67801129) → GFAP (Q14864879) referenced by: <ul style="list-style-type: none"> • stated in (P248) PanglaoDB (Q99936939) • retrieved (P813) 08 of November 2020 • reference URL (P854) https://panglaoDB.se/markers.html
Example 2	human cytotoxic t cell (Q101423166) → CD8 [plasma membrane] (Q50260473) referenced by: <ul style="list-style-type: none"> • retrieved (P813) 09 of November 2020 • reference URL (P854) https://www.niaid.nih.gov/research/immune-cells
Example 3	human t helper cell (Q101423298) → CD4 molecule (Q412587) referenced by: <ul style="list-style-type: none"> • retrieved (P813) 09 of November 2020 • reference URL (P854) https://www.niaid.nih.gov/research/immune-cells
Source	<ul style="list-style-type: none"> • PanglaoDB marker database: PanglaoDB (Q99936939) and PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data (Q63664483) <ul style="list-style-type: none"> ◦ CellMarker marker database: CellMarker (Q64371987) and CellMarker: a manually curated resource of cell markers in human and mouse (Q56984510) • CellPedia-SHOGoin marker database: CELLPEDIA: a repository for human cell information for cell studies and differentiation analyses. (Q35482351)
Planned use	Reconcile knowledge from the PanglaoDB marker database to Wikidata. In the future, expand to other trusted sources of cell type marker information.

It 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"

In the same day, the property received 3 "Support" tags by different members of the community:

- **Support** — [Jvcavv](#) ([talk](#)) 14:51, 17 November 2020 (UTC)
- **Support** [Bmeldal](#) ([talk](#)) 20:00, 17 November 2020 (UTC)
- **Support** – *The preceding [unsigned](#) comment was added by [Tinker Bell](#) ([talk](#) • [contribs](#)) at 20:07, November 17, 2020 (UTC).*

As there were 3 approvals and no objections, 10 days later one Wikidata Property Administrator created the property:

@[TiagoLubiana](#), [Jvcavv](#), [Bmeldal](#), [Tinker Bell](#):  **Done** [has marker \(P8872\)](#) [Pamputt \(A\)](#) ([talk](#)) 07:24, 27 November 2020 (UTC)

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. Human-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. 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 of the integration

As explained in the method session

SPARQL query for cell types related to neurogenesis

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) or even requesting queries (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.

Acknowledgements

This work has been done within the scope of a slightly larger ongoing project in a collaboration with João Vitor Ferreira Cavalcante available at https://github.com/jvfe/wikidata_panglaodb/.

This work has been supported by grant #2019/26284-1, São Paulo Research Foundation (FAPESP).

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>