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

Authors

- **João Vitor Ferreira Cavalcante**
  (ID) [0000-0001-7513-7376](#) · ⬡ [jvfe](#)
  Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte

- **Tiago Lubiana**
  (ID) [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, its 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 types 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 [1] is an open, freely editable, knowledge graph database within the semantic web [2] 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 [3] [4] and showcasing it's potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [4]. 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 [5]. 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"[6] (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 [7] [8] 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

Methodology

Data acquisition

Data from Wikidata will be acquired using the Wikidata Query Service [9] and associated wrapper packages in Python, such as WikidataIntegrator[10] and wikidata2df [11].

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[12], Seaborn[13] and Jupyter[14].

Reconciliation

The metadata from PanglaoDB on cell types, genes, tissues (including germ layers) and organs will be matched to Wikidata items using the reconciler[15] Python package, which is itself a wrapper around the well known OpenRefine[16] 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 [17] and Wikidata's own query service. And also by the presence of external identifiers, such as Ensembl Gene[18] and Entrez Gene[19] IDs for genes, Cell Ontology[6] IDs for cell types and Uberon[20] IDs for organs and tissues.

Results

Histological entities from PanglaoDB, that is, cell types, tissue types and organs, were reconciled with Wikidata items using the reconciler[15] python package and an associated Python stemming function from NLTK [21] for string matching.

After reconciliation, items were manually checked for false matches - items with the same name but that don't represent the same concept. Finally, we obtained the reconciled csv data, and it's summary can be seen on Table 1.

Depicted in Table 1 are also matches for gene data, which were acquired using manual intersection of both sources, via a Pandas inner merge. This data didn't go through reconciliation because of it's size, but since both ends are dealing with identifiers (Gene Symbol), the item matching should work much better than with histological entities, which are described in natural language.

Table 1: Summary of the matched entities from PanglaoDB.

	# of unique matches	# of matched items	% of total items that were matched	% of matches that were perfect	% of matches that don't have P31
Cells	81	85	37.6744	38.8235	55.2941
Tissues	79	87	32.1138	62.069	37.931
Organs	22	30	75.8621	53.3333	46.6667
Human Genes	35423	35427	60.847533	NA	NA

	# of unique matches	# of matched items	% of total items that were matched	% of matches that were perfect	% of matches that don't have P31
Mouse Genes	25124	25127	46.704962	NA	NA

Afterwards, we analysed which types these reconciled items belonged to in Wikidata, which is indicated by their "instance of" (P31) property. Most items were missing this information (Figure 1).
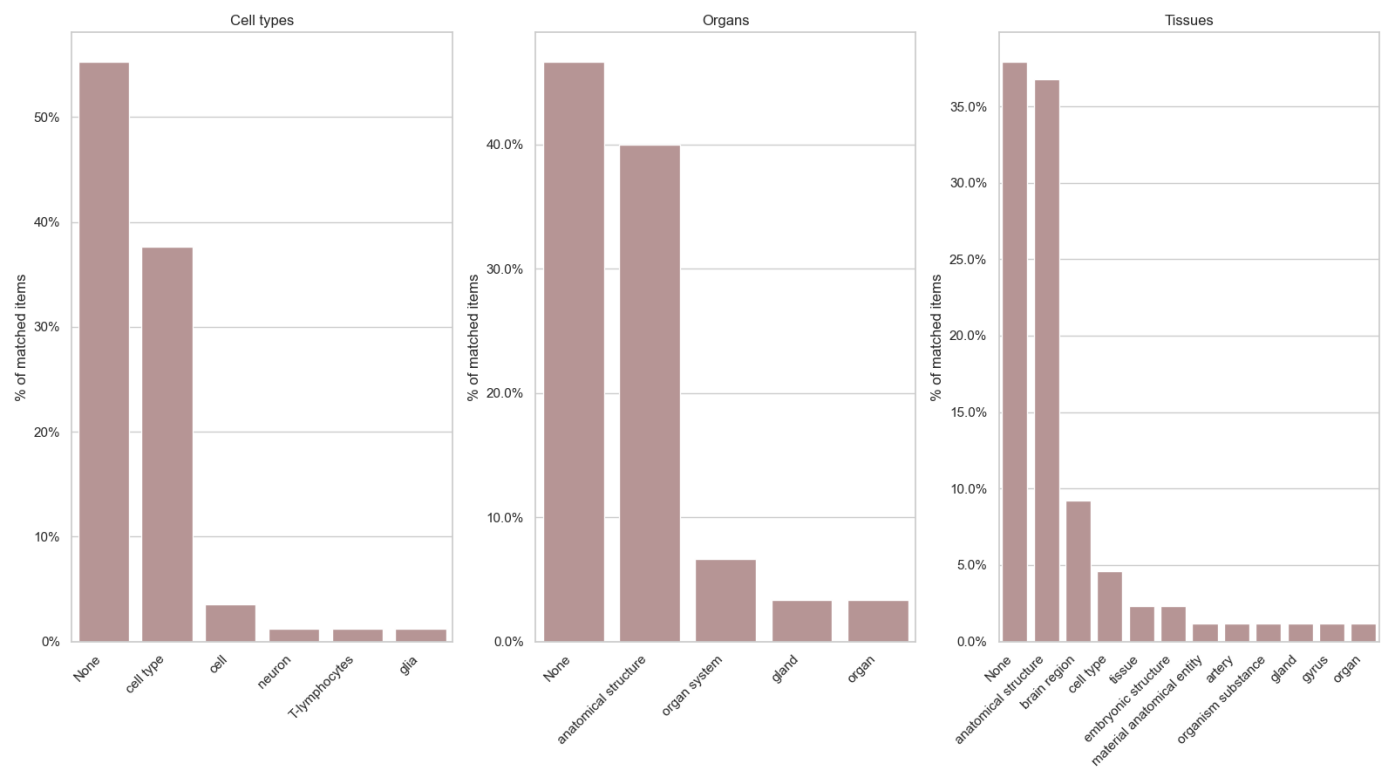
Figure 1: Percentage of reconciled entities, divided by which item type they belong to. Most reconciled items don't count with a "instance of" property.

General Ideas

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

med2rdf[22] 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 [23]

References

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

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

3. **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

4. **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

5. **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

6. **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

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

8. **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

9. https://query.wikidata.org/

10. **wikidataintegrator: Python package for reading and writing to/from Wikidata**
    Andra Waagmeester Burgstaller-Muehlbacher Greg Stupp, Sebastian
    https://github.com/sulab/WikidataIntegrator

11. **wikidata2df: Utility package for easily turning a SPARQL query into a dataframe**
    João Vitor F. Cavalcante
    https://github.com/jvfe/wikidata2df

12. **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

13. **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

14. **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

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

16. **OpenRefine** https://openrefine.org/

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

18. **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

19. **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

20. **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

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

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

23. **:{unav)**
    Jonathan Bard, Seung Y Rhee, Michael Ashburner