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

This manuscript (permalink) was automatically generated from jvfe/paper_wdt_panglao@b961eac on September 22, 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

Methods

Histological entities from PanglaoDB, that is, cell types, tissue types and organs, were reconciled with Wikidata items using the reconciler[1] python package and an associated Python stemming function from NLTK [2] 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
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).

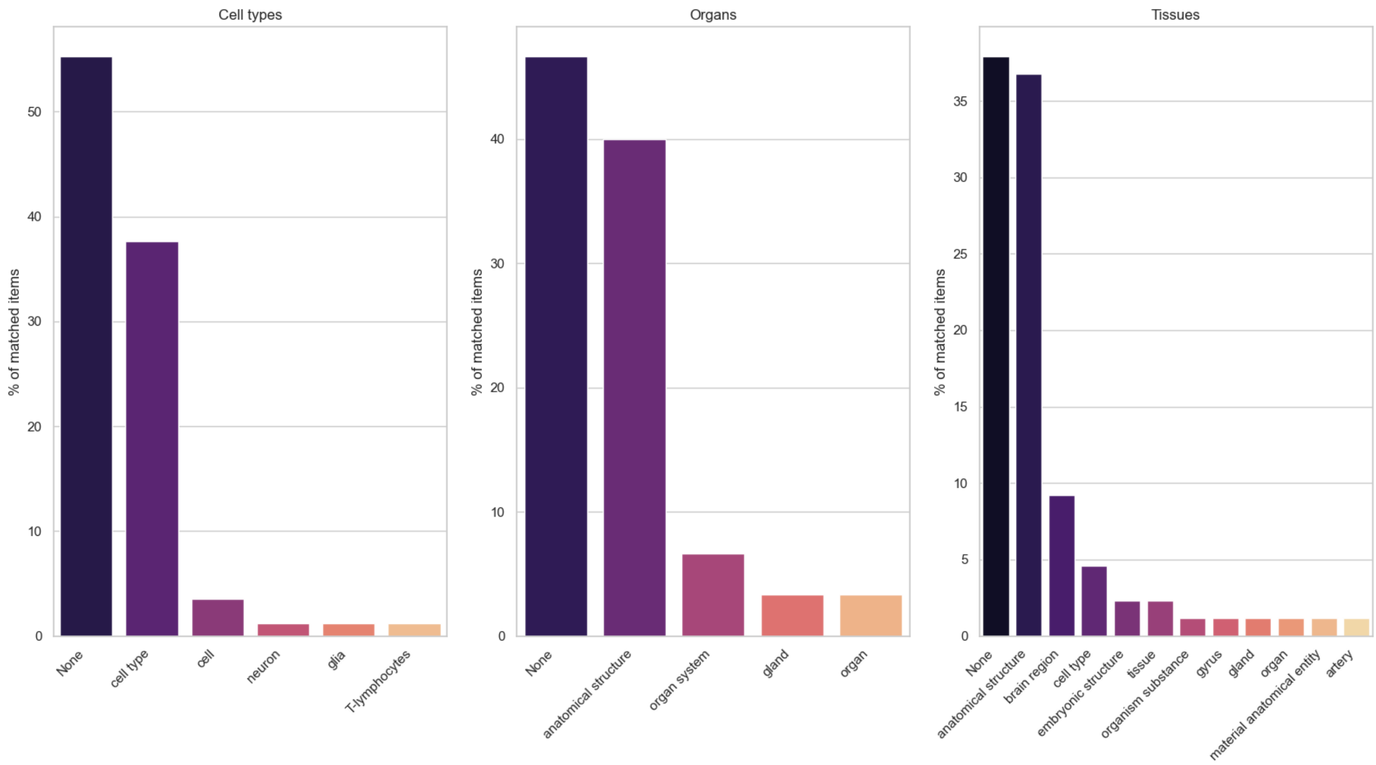

Figure 1: Percentage of reconciled entities, divided by which item type they belong to. Most reconciled items don't count with a "instance of" property.

General Ideas

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

med2rdf[3] 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 [4]

References

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

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

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

4.:**{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