Manuscript Title

This manuscript ([permalink](#)) was automatically generated from [lubianat/fapesp_report_1@f10dfcf](#) on June 17, 2021.

Authors

- **John Doe**
  ⓘD [XXXX-XXXX-XXXX-XXXX](#) · ⦿ [johndoe](#) · 🐦 [johndoe](#)
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe**
  ⓘD [XXXX-XXXX-XXXX-XXXX](#) · ⦿ [janeroe](#)
  Department of Something, University of Whatever; Department of Whatever, University of Something

UNIVERSIDADE DE SÃO PAULO FACULDADE DE CIÊNCIAS FARMACÊUTICAS Departamento de Análises Clínicas e Toxicológicas

Doctorate Project - Report 1 FAPESP process number: Project term: Period of Scientific report:

Scholarship Holder: Tiago Lubiana

Responsible Researcher: Prof. Dr. Helder Takashi Imoto Nakaya

São Paulo - SP November 2020

Project summary

The advent of single-cell technologies has ignited the desire of a deep knowledge on cells, the building blocks of life [1]. The Human Cell Atlas (HCA) project, has been a major player in the cell knowledge ecosystem, running since 2017 towards the task to characterize every cell type in the human body [2]. The HCA consortium recruited people from all over the world to tackle different parts of the project. In Brazil, Prof. Helder Nakaya (supervisor of this PhD project) is leading the national effort to contribute to HCA, with a focus on the roles of different cell types in the pathological processes of infectious and inflammatory diseases.

HCA is set to revolutionize the biomedical sciences, by creating tools and standards for basic research, as well as allowing better characterization of disease, and thus, ultimately, improving diagnostics and therapy. Its products (data, information, knowledge and wisdom) need to be FAIR: findable, accessible, interoperable and reusable. Data stewardship and data management are growing as core demands of the scientific community, ranging from data management plans [3] to specialized personnel [3].

The Human Cell Atlas has a dedicated team for organizing data: the Data Coordination Platform (DCP) [4] [5]. The DCP is responsible for tracing the plan for computational interoperability, from the data generators to the consumers.[5]. The Human Cell Atlas has its portal for data (https://data.humancellatlas.org/) which composes the data repository landscape with other resources, like the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell) and the Chan-Zuckerberg Biohub Tabula Sapiens (https://tabula-sapiens-portal.ds.czbiohub.org/). In addition to its core team, the HCA is poised to grow by community interaction, and states in its opening paper that "As with the Human Genome Project, a robust plan will best emerge from wide-ranging scientific discussions and careful planning."[2]
Thus, this project inserts itself among the wide-ranging scientific discussions to improve data - and knowledge - interoperability.

The large amount of information generated by HCA creates the need for innovative knowledge management approaches. For the Human Cell Atlas Project (as presented in the chapter) to maximize its benefit for society, its knowledge products will need to be inserted into the main route of automated knowledge discovery .
The field denominated Literature Based Discovery [6] dedicates itself to this challenge: making actual discoveries (or at least very strong hypothesis) using as material plainly the existing literature. [7] The textbook example of Literature Based Discovery is described by Don Swanson's so-called ABC model: If A is related do B, and B is related to C, then A and C are indirectly related [8] In a seminal paper, Swanson showed an hypothesis about using fish oil (A) to treat Raynauld's disease (C), demonstrating that even though the specialized fish-oil (A) literature had shown its association (AB) with a set of blood parameters (B), and the specialized Raynauld's disease literature had show its association (BC) with the same set of parameters (B), the AC link was never made in the literature, despite its seeming obviousness [8].

Modern advancements of literature-based discovery rely on Natural Language Processing, Machine Learning and Knowledge graphs to make inferences on literature knowledge. Word embeddings, for example, are leading inference of properties of compounds based on their shared neighbourhood of words (the words before and after their mentiongs) with known compounds, thus making use of latent knowledge in the body of knowledge. [9] Other, more explicit approaches, rely on extracted relations embedded in knowledge graphs,fo example, the discovery of new RNA-binding proteins related to Amyotrophic Lateral Sclerosis by analysis of the Watson Drug Discovery gene-disease network. [10] Knowledge graphs have a set of characteristics that make then useful for Literature Based Discovery: the power of representing multiple relations, the power of making inferences on top

of those relations, and provide human understandability at every step, allowing for a dialog between expert humans and computing systems. The field of biomedical ontologies explores that direction in depth, and the community is building many solutions, widely applicable for the biomedical sciences.

An ontology, as used here, is a formal computational representation of reality, which tries to represent each concept (and their relations) as precisely as possible. [11] Constructing an ontology is a process of selecting and defining terms of interest, selecting and defining relationships of interest and making statements about reality using terms and relationships. The Gene Ontology is probably the most well known biomedical ontology; it describes (among other things) different classes of biological process, related to each_other by "is_a" and "part_of relations. [12] [13].

The Gene Ontology is part of a much larger effort to formalize concepts across biology: the Open Biomedical and Biological Ontologies (OBO) Foundry. [14] Created in 2007, the OBO Foundry is a hub of biomedical ontologies that sets guidelines for the design and construction of high-quality ontologies. The initial OBO Foundry united several independent ontologies, like the Cell Ontology (CL), the Disease Ontology (DO) and the Protein Ontology (PRO) under a common framework, a great progress towards interoperability. At the same time, the creation of the Relation Ontology (RO) provided a go-to point for relations in biology that could them be reused by different ontologies.

Ontologies are powerful, but require a high degree of tecnichal expertise to get started. Recently, a new approach for formal knowledge representation arose with the dawn of collaborative knowledge graphs. Wikidata, the collaborative knowledge graph of the Wikimedia foundation, allows users to contribute with classes and statements, in the same spirit of Wikipedia and share its "epistemic virtues, like power, speed and availability. [15] It is powerful because of its large community of contributors. With a community of more than 25,000 active editors (https://www.wikidata.org/wiki/Wikidata:Statistics) and growing, it is able to cover a much wider number of concepts than any user individually. It is fast, because one does not need to install any software or ask for permissions to update it: any user can simply do it via a web interface. That speed makes it easier for newcomers to join and contribute, in contrast to OBO Foundry ontologies, which require extensive training on semantics and knowledge of Git/GitHub for contributions. Finally, the information on Wikidata is available via an user interface, via a SPARQL query service and as large, full-size database dumps, providing full extent reusability. The Wikidata model has been so sucessfull that Google decided to migrate its own knowledge base, Freebase, fully into Wikidata.[16]

Several advances towards biological data integration and biological data analysis in Wikidata have been made before, yielding positive results [17] [18] and showcasing its potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [19]. 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 [20].

The aim of this project is to to study current understanding of cell types for development a comprehensive ontological model in Wikidata for cell types. We are reviewing the single-cell literature, refining and formalizing concepts for cell type delimitation and exploring their application in the Wikidata database. At the same time, we are exploring the use of Natural Language Processing tools, in combination with expert annotations tools to automate knowledge extraction from scientific articles in the scope of the Human Cell Atlas. The specific goals outlined in the approved project were:

- Build a data model to capture the main properties to describe a human cell. Provide working definitions of cell types and states and their characteristics.
- Extract and add to Wikidata pieces of information regarding Human Cell Atlas publications to build gold standards. Use this information to develop machine learning tools to extract knowledge from publications.

- Create tools to make data from the underlying knowledge graph accessible employing tools from network theory.

Achievements

In the first year of the project, we addressed 3 different facets of the project.

The first facet was the refinement of the fcoex R/Bioconductor project, written for the detection of cell classes in single-cell RNA-seq data. The second facet dealt with the concept of cell type itself, in the goal of building theoretical models for the formalization of knowledge. The third facet was aimed at the practical application of the concepts. We adapted a public database of cell type markers, Panglao DB, to the Wikidata format, made the information available through Wikidata and perfomed SPARQL (SPARQL Protocol and RDF Query Language) queries to glance at the information on the database.

Fcoex: using coexpression to explore cell type diversity in scRNA-seq data

The fcoex (https://bioconductor.org/packages/fcoex/) was written by us to identify coexpression modules and cell populations in single-cell RNA-seq data. During the first year of this PhD projects, we worked on framing its use in the context of cell type representation and the Human Cell Atlas. The package was refactored and upgraded to conform to the new versions of R (now in 4.1.0) and Bioconductor (now in 3.12). We also developed the narrative and shaped the manuscript for submission.

The fcoex method is designed for application right after a standard scRNA-seq clustering step ([1] A). The cluster assignments convey information about the relations between cells to the algorithm and help to guide feature selection. Then, the package selects global marker genes specific to 1, 2, or more previously defined clusters. It ranks markers by symmetrical uncertainty, a non-linear correlation metric based on classical Shannon entropy. [21]

To find co-expression modules, fcoex inverts the FCBF feature selection algorithm, and instead of removing redundancy, it selects redundant (co-expressed) gene expression patterns. (see Supplementary Data for details). The default gene coexpression modules yielded by the pipeline are small by design (10s of genes per module) to facilitate manual exploration of the coexpression landscape. Each module has one "header" gene, which expression pattern better represents all the genes in the module.

Fcoex treats each module as a gene set to find cell populations, using only their expression to re-cluster cells. The new classifications are based, thus, on the genes (and functions) captured by each co-expression module. The multiple module-based clusters serve as a platform for exploring the diversity of the dataset and identifying upper cell classes, grouping cells by common functions.
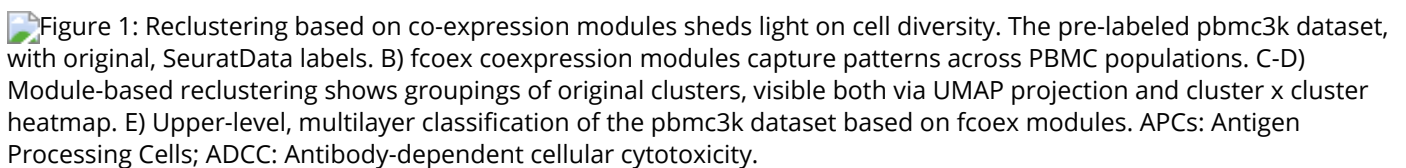

Figure 1: Reclustering based on co-expression modules sheds light on cell diversity. The pre-labeled pbmc3k dataset, with original, SeuratData labels. B) fcoex coexpression modules capture patterns across PBMC populations. C-D) Module-based reclustering shows groupings of original clusters, visible both via UMAP projection and cluster x cluster heatmap. E) Upper-level, multilayer classification of the pbmc3k dataset based on fcoex modules. APCs: Antigen Processing Cells; ADCC: Antibody-dependent cellular cytotoxicity.

Figure 1: Reclustering based on co-expression modules sheds light on cell diversity. The pre-labeled pbmc3k dataset, with original, SeuratData labels. B) fcoex coexpression modules capture patterns across PBMC populations. C-D) Module-based reclustering shows groupings of original clusters, visible both via UMAP projection and cluster x cluster heatmap. E) Upper-level, multilayer classification of the pbmc3k dataset based on fcoex modules. APCs: Antigen Processing Cells; ADCC: Antibody-dependent cellular cytotoxicity.

Multi-hierarchies of blood types

To validate the fcoex pipeline, we selected the well-known pbmc3k dataset from the SeuratData R package, which contains around 2700 peripheral blood mononuclear cells (PBMC) with previously-defined cluster labels.

The standard fcoex pipeline detected nine modules that capture different parts of the cellular diversity in the dataset. For example, module M8 contains cytotoxicity genes, as PRF1 and GZMA, and splits the dataset into cytotoxic (NK and CD8) and non-cytotoxic cells. M2 (CD3D) splits the dataset in T-cells and non-T-cells. M5 (HLA-DRB1) groups together monocytes, B cells, and dendritic cells, all known antigen-presenting cells (APC). (1 B-E) The classifications provided by fcoex are easily reintegrated to Seurat to power visualizations and get differentially expressed markers, providing more genes for the analysis, if desired.

In general, fcoex clusters combined biologically similar cell types of the original dataset. The clusterings help to explore and classify cells by function (1 E). Even in a well-studied dataset, fcoex provided a new light on the shared functionality of some NK cells and macrophages: they both markedly express the CD16-coding gene FCGR3A, whose product is a key player in antibody-dependent cellular cytotoxicity (ADCC). [22] Thus, a complete functional classification of cells might want to include a "professional ADCC cells" class

To sum up, main goal of the fcoex pipeline is to use the modules to find biologically relevant populations, which then can be represented in ontologies, like the Cell Ontology [23], and by knowledge graphs, like Wikidata. By doing so, fcoex offers new avenues to explore data-driven classifications of cells, aligning itself with the challenges of the Human Cell Atlas and building catalogs of cell types in the single-cell era.

Concept of cell types

During the development of *fcoex* we assumed the common notions of cell type, as a group of cells that share common features. The definition of the concept of "cell type," however, is currently a topic of debate by the biomedical community.[1,24,25,26,27,28,29,30,31,32,33,34]. In an opinion article published in Cell Systems in 2017, a series of researchers presented their views on the conceptual definition of 'cell type' in the context of a mature organism [24]. The opinions were varied, and no consensus was achieved.

Before we proceeded with the knowledge-graph formalizations via Wikidata, we dedicated time for a theoretical research on the concept of "cell type" in the context of knowledge representation. This line of research aligns itself with the groundwork of the Cell Ontology [35,36,37] and CELDA [38] and the contributions of the International Workshop on Cells in Experimental Life Sciences series [39,40].

In this period, we targeted the question: which cell type definition allows crafting coherent biological statements? The goal was to not say what cell types *are*, but what they can be for a consistent representation on a knowledge graph, like Wikidata. Furthermore, for the practical purpose adopted here, we avoid the dissection of the differences between persistent classes of cells (often called "cell types") or the transient, fugacious classes of cells (often called "cell states") (see "Definition of cell identity" section in [41] for an example). Even though such a distinction is an essential topic for theoretical research, it is not required to represent formally biomedical experiments.

To facilitate communication among life scientists, in a preprint derived from this PHD project[42], we proposed, among other theoretical advancements, naming conventions for different cell types classes. Much of the literature mixes cell types in one species (e.g., when dealing with a cell type as an evolutionary unit) or multiple species (e.g., in the Cell Ontology). It is useful to distill these different concepts into names. Given the importance of the species' concept in biological classification [43], we derive a species-centric view on the naming of classes of cell types. The four classes (Figure 2) we propose are as follows:
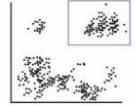
- archetypes, for when the taxonomic scope of the type is beyond the level of species; for example, "mammal neutrophils."
- *sensu stricto* cell types, for when the taxonomic scope of the type corresponds to a single species; for example, *Mus musculus* neutrophils."
- infratypes, for when the taxonomic scope is below the level of species; for example, considering the mouse strain "C57BL/6J," "neutrophils from C57BL/6J mice."
- technotypes, for specific, experimentally defined cell types that harbor in their definition the precise conditions of the cells sampled; "2-month-old male C57BL/6J, Ly-6G$^+$ CD11b$^+$ M-CSF R$^-$ CD244$^-$ neutrophils."

Figure 2: Names for classes of cell types.

The 4 different categories of cell types help us to better organize the knowledge about cell types. Even though individual articles and databases often have species-neutral names, the information often comes from experiments with a single strain of a single species. Two articles might call by the same name cells that come from different animals, or were selected by different protocols. Large scale knowledge management requires an organized way of representing those details.

The division between archetypes and *sensu stricto* cell types is of special importance in the context of biocuration and annotation of data. Associations like the HUGO Gene Nomenclature Committee and UniProt organize names and identifiers for genes and proteins in single species. Thus, if we want to annotate marker genes, for example, we need to associate them to a species-specific cell type (a *sensu stricto* cell type) instead of the more vague association to a species-neutral type. That might seem obvious, but current standards still use identifiers that are species-neutral (e.g. in the reference HuBMAP app; https://azimuth.hubmapconsortium.org/references/)

The ontological discussion on the classes of cell types, thus, extends the current state-of-the-art and introduce new ways to organize our knowledge about cells. Notably, the technotype and the infratype are, currently, mostly theoretical constructs and almost no resources deal with cell types at the level of strains or below. The division of archetypes and *sensu stricto* cell types, on the other hand, was already instrumental for the integration of the Panglao database of cell markers to Wikidata, described in the next session.

PanglaoDB integration to Wikidata

The process of making the Human Cell Atlas more useful via Wikidata also includes the connection of related databases.

PanglaoDB [44] [45] 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 and cell type markers, obtained both via automatic and manual methods. It also displays a rich web user interface for easy data acquisition, including database dumps for bulk downloads.

As of 17 June 2021, the article describing PanglaoDB had been cited 147 times. Despite its use by the community, the database is on a 3-star category for Linked Open Data [46] as it does not use the open semantic standards from W3C (RDF and SPARQL) needed for a 4-star rank, neither the links to external data via standard identifiers that make datasets 5-star. Improving the data format toward W3C's gold standards is a valuable step in making biological knowledge FAIR (Findable, Accessible, Interoperable, and Reusable). Thus, we aimed to provide a case study of making the core information of PandlaoDB available in a 5-star Linked Open Data Format while improving the modeling of the necessary concepts on Wikidata.

As of August 2020, Wikidata had 264 items being categorized as a "cell type" (Wikidata ID Q189118), considerably less than in specialized cell catalogs, which count over two thousand cell types [23,47]. Strikingly, there were also 23 items categorized as "instances of cell (Q7868)" . This classification is imprecise, as an instance of cell would be an individual named cell from a single named individual.

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").

We diligently fixed and improved information on cell types on Wikidata. As of 17 June 2021, the Wikidata database contains 2102 instances of "cell type" (see current status at https://w.wiki/b2t) and 0 instances of "cell" (https://w.wiki/b2q) highlighting the improvements in both quantity and quality.

Methodology

After obtaining approval from the owners of the database, we matched genes and cell types to Wikidata, and performed Wikidata queries to demonstrate the value of the approach. An overview of the process is shown in ??. Wikidata SPARQL queries bring to light hidden biomedical knowledge

Class creation on Wikidata

Classes corresponding to species-neutral classes were retrieved from Wikidata manually using Wikidata's Graphic User Interface. A manually-curated dictionary matching terms in PanglaoDB to Wikidata identifiers was assembled and used for integration. Cell types that 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 and 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 [38].
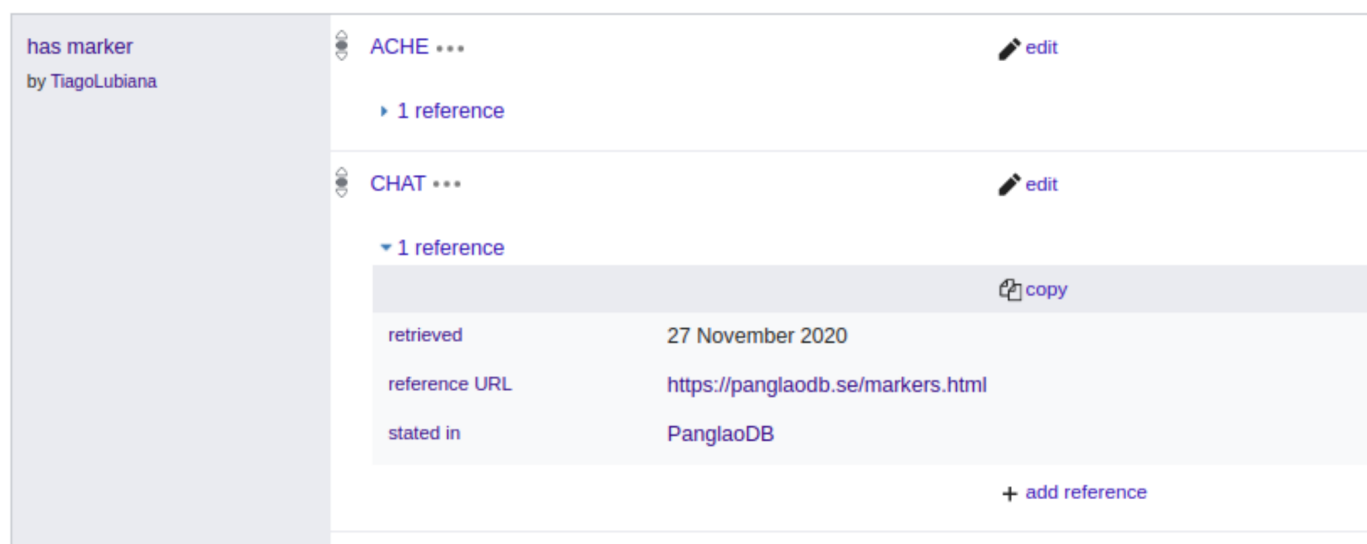
Integration of PanglaoDB to Wikidata

After receiving authorization by e-mail from the PanglaoDB developer, Oscar Franzen, the PanglaoDB markers dataset was downloaded manually from PanglaoDB's website (https://panglaodb.se/markers/PanglaoDB_markers_27_Mar_2020.tsv.gz) for integration. It contains 15 columns and 8256 rows. Only the columns `species`, `official gene symbol`, and `cell type` were used for the reconciliation. The reconciled dataset was uploaded to Wikidata via the WikidataIntegrator Python package [48], a wrapper for the Wikidata Application Programming Interface.

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

Cell Marker information on Wikidata

Adding marker information on Wikidata was not possible before this study and became possible after we proposed and got community approval of the property "has marker" (P8872). Figure 3 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 3: Subset of the marker genes for item Q101405051 (human cholinergic neuron)

Now that we re-formatted the markers on PanglaoDB as Linked Open Data, we can make queries that were not possible before, including federated queries with other biological databases, such as Uniprot [49] and Wikipathways [50]. Due to previous similar reconciliation projects, Wikidata already contains information about genes, including their relations to Gene Ontology (GO) terms.

PanglaoDB's integration to the Wikidata ecosystem allows us to ask a variety of questions (figure 4).
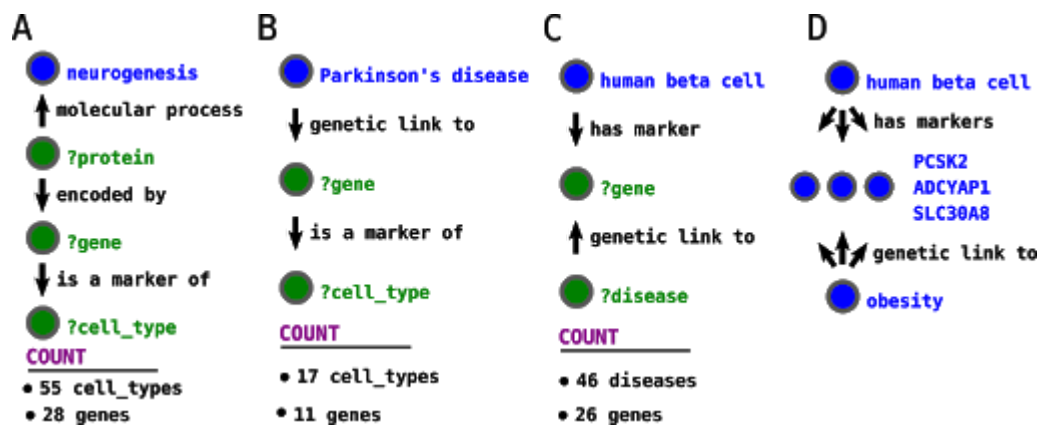
Figure 4: SPARQL queries in Wikidata now harness information from Panglao DB. Queries with the above design were run on Wikidata. Results might change in real time with Wikidata updates by contributors A-C) Graphical representation of feasible SPARQL queries (https://w.wiki/yQ6, https://w.wiki/yQD and https://w.wiki/3HjX,). D) Sample result from query C.

"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 also retrieved non-neural cell types such as the "human loop of henle cell, a kidney cell type, and"human osteoclast. These seemingly unrelated cell types markedly express genes involved in neurogenesis, but that does not mean that they are involved with this process. The seemingly confusing results reinforce 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. SPARQL allows us to seamlessly compare Gene Ontology processes with cell marker data, providing a sandbox to generate hypotheses and explore the biomedical knowledge landscape.

Table 1: Top 10 cell types related to neurogenesis via markers (07/02/2020, full query on https://w.wiki/yQ6).

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

"Which cell types express markers associated with 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 to detail 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 marked by genes genetically associated with Parkinson's disease. Even considering the false positives, the overview can aid domain experts in coming up with novel hypotheses.

Table 2: Top 5 cell types related to Parkinson's disease via markers (07/02/2020, full query on https://w.wiki/yQD).

geneLabel	diseaseLabel	cellTypeLabel
BST1	Parkinson's disease	human b cell
BST1	Parkinson's disease	human neutrophil
RIT2	Parkinson's disease	human neuron
SH3GL2	Parkinson's disease	human alpha cell
SH3GL2	Parkinson's disease	human beta cell

Discussion and conclusion

In this part of the PhD project, 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. Based on the theoretical considerations on the concept of cell type, we added species-specific terms to Wikidata for cell types of *Homo sapiens* and *Mus musculus* described in the PanglaoDB database.

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/). It is not first case study of biomedical data integration to Wikidata (see [51] for example. Nevertheless, the differences among the articles in style and scope contribute to a richer ecosystem for possible contributor.]) The work also paves the way for Wikidata reconciling of other databases for cell-type markers, such as CellMarker [52], labome [53], CellFinder [47] and SHOGoiN/CELLPEDIA [54]/]) (if proper authorization are given by the owners). 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.

Next steps

After this groundwork, the next steps of this PhD project will be geared towards increased alignment with the representation of knowledge of the Human Cell Atlas.

- Set up the semantic infrastructure on Wikidata for handling knowledge about cell types

    - Refine the theories of types/states/classes of cells within the constraints of ontologies and knowledge bases
    - Investigate the types of statements done about cell types
        - On Wikidata
        - On OBO Foundry ontologies
        - Freely on the biomedical literature
    - Craft wikidata relations ("properties") for making cell-type-related assertions (like "has marker" or "is the progenitor of")

- Devise ways to connect the Human Cell Atlas products to Wikidata and the Linked Open Data cloud

    - Write bots and scripts to reconcile data sources to Wikidata
    - Create tools for biocuration of Human Cell Atlas products combining text mining and expert curation
    - Project software for reuse of HCA-related knowledge integrated into common bioinformatics workflows.

- Provide proofs-of-concepts of how Wikidata integration can benefit the advancement of HCA

Chronogram

ADAPT FROM THE ORIGINAL PROJECT

Preprints and articles submitted for publication

- Fcoex
- Technotype

References

1. **An era of single-cell genomics consortia**
   Yoshinari Ando, Andrew T Kwon, Jay W Shin
   Experimental and Molecular Medicine (2020-09-15) https://www.wikidata.org/wiki/Q99418649
   DOI: 10.1038/s12276-020-0409-x

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

3. **Everyone needs a data-management plan**
   Nature
   (2018-03-15) https://www.wikidata.org/wiki/Q56524391
   DOI: 10.1038/d41586-018-03065-z

4. **About the Data Coordination Platform**
   HCA Data Portal
   https://data.humancellatlas.org/about/

5. **The Human Cell Atlas White Paper**
   Aviv Regev, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael JT Stubbington, Kristin Ardlie,
   Amir Giladi, Paola Arlotta, Gary D Bader, Christophe Benoist, Moshe Biton, … Human Cell Atlas
   Organizing Committee
   (2018-10-11) https://www.wikidata.org/wiki/Q104450645

6. **Literature-based discovery**
   Wikipedia
   (2021-01-24) https://en.wikipedia.org/w/index.php?title=Literature-
   based_discovery&oldid=1002467308

7. **Literature Based Discovery: models, methods, and trends.**
   MSSam Henry, Bridget T McInnes
   Journal of Biomedical Informatics (2017-08-21) https://www.wikidata.org/wiki/Q38371706
   DOI: 10.1016/j.jbi.2017.08.011

8. **Online tools to support literature-based discovery in the life sciences.**
   Marc Weeber, Marc Weeber, Jan A Kors, Jan A Kors, Barend Mons
   Briefings in Bioinformatics (2005-09-01) https://www.wikidata.org/wiki/Q36280460
   DOI: 10.1093/bib/6.3.277

9. **Unsupervised word embeddings capture latent knowledge from materials science
   literature**
   Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova,
   Kristin A Persson, Gerbrand Ceder, Anubhav Jain
   Nature (2019-07-03) https://www.wikidata.org/wiki/Q91595456
   DOI: 10.1038/s41586-019-1335-8

10. **Artificial intelligence in neurodegenerative disease research: use of IBM Watson to
    identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis.**

Nadine Bakkar, Tina Kovalik, Ileana Lorenzini, Scott Spangler, Alix Lacoste, Kyle Sponaugle, Philip Ferrante, Elenee Argentinis, Rita Sattler, Robert Bowser
Acta Neuropathologica (2017-11-13) https://www.wikidata.org/wiki/Q47406275
DOI: 10.1007/s00401-017-1785-8

11. **Ontologies for the life sciences**
Steffen Schulze-Kremer, Barry Smith
(2005-11-15) https://www.wikidata.org/wiki/Q105870680
DOI: 10.1002/047001153x.g408213

12. **The Gene Ontology resource: enriching a GOld mine**
Gene Ontology Consortium
Nucleic Acids Research (2020-12-08) https://www.wikidata.org/wiki/Q104130127
DOI: 10.1093/nar/gkaa1113

13. **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**
M Ashburner, CA Ball, Judith A Blake, David Botstein, H Butler, JMichael Cherry, AP Davis, K Dolinski, Selina S Dwight, JT Eppig, … Gavin Sherlock
Nature Genetics (2000-05-01) https://www.wikidata.org/wiki/Q23781406
DOI: 10.1038/75556

14. **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**
Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, … Suzanna Lewis
Nature Biotechnology (2007-11-01) https://www.wikidata.org/wiki/Q19671692
DOI: 10.1038/nbt1346

15. **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

16. **From Freebase to Wikidata: The Great Migration**
Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, Lydia Pintscher
Proceedings of the 25th International Conference on World Wide Web (2016-01-01)
https://www.wikidata.org/wiki/Q24074986
DOI: 10.1145/2872427.2874809

17. **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

18. **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

19. **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

20. **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

21. **Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution**
Lei Yu, Huan Liu
https://www.wikidata.org/wiki/Q106704674

22. **CD16 is indispensable for antibody-dependent cellular cytotoxicity by human monocytes**
Wei Hseun Yeap, Kok Loon Wong, Noriko Shimasaki, Esmeralda Chi-yuan Teo, Jeffrey Kim Siang Quek, Hao Xiang Yong, Colin Phipps Diong, Antonio Bertoletti, Yeh Ching Linn, Siew Cheng Wong
Scientific Reports (2016-09-27) https://www.wikidata.org/wiki/Q27341786
DOI: 10.1038/srep34310

23. **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.wikidata.org/wiki/Q36067763
DOI: 10.1186/s13326-016-0088-7

24. **What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?**
Paul Blainey, Hans Clevers, Cole Trapnell, Ed Lein, Emma Lundberg, Alfonso Martinez Arias, Joshua R Sanes, Jay Shendure, James Eberwine, Junhyong Kim, … Mathias Uhlén
Cell systems (2017-03-01) https://www.wikidata.org/wiki/Q87649649
DOI: 10.1016/j.cels.2017.03.006

25. **A periodic table of cell types**
Bo Xia, Itai Yanai
Development (2019-06-15) https://doi.org/ggctwf
DOI: 10.1242/dev.169854 · PMID: 31249003 · PMCID: PMC6602355

26. **Exciting times to study the identity and evolution of cell types**
Maria Sachkova, Pawel Burkhardt
Development (2019-09-15) https://doi.org/ghdb9v
DOI: 10.1242/dev.178996 · PMID: 31537583

27. **The Human Cell Atlas: from vision to reality.**
Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, Sarah Teichmann
Nature (2017-10-01) https://www.wikidata.org/wiki/Q47565008
DOI: 10.1038/550451a

28. **Human Cell Atlas and cell-type authentication for regenerative medicine**
Yulia Panina, Peter Karagiannis, Andreas Kurtz, Glyn N Stacey, Wataru Fujibuchi
Experimental and Molecular Medicine (2020-09-15) https://www.wikidata.org/wiki/Q99418657

DOI: [10.1038/s12276-020-0421-1](https://doi.org/10.1038/s12276-020-0421-1)

29. **A community-based transcriptomics classification and nomenclature of neocortical cell types**
Rafael Yuste, Michael J Hawrylycz, Nadia Aalling, Argel Aguilar-Valles, Detlev Arendt, Rubén Armañanzas, Giorgio A Ascoli, Concha Bielza, Vahid Bokharaie, Tobias B Bergmann, … Ed S Lein
Nature Neuroscience (2020-08-24) <https://www.wikidata.org/wiki/Q98665291>
DOI: [10.1038/s41593-020-0685-8](https://doi.org/10.1038/s41593-020-0685-8)

30. **The evolving concept of cell identity in the single cell era**
Samantha A Morris
Development (2019-06-27) <https://www.wikidata.org/wiki/Q93086971>
DOI: [10.1242/dev.169748](https://doi.org/10.1242/dev.169748)

31. **Implications of Epigenetic Variability within a Cell Population for "Cell Type" Classification**
Inna Tabansky, Joel Stern, Donald W Pfaff
Frontiers in Behavioral Neuroscience (2015-12-16) <https://www.wikidata.org/wiki/Q26770736>
DOI: [10.3389/fnbeh.2015.00342](https://doi.org/10.3389/fnbeh.2015.00342)

32. **Geometry of the Gene Expression Space of Individual Cells**
Yael Korem, Pablo Szekely, Yuval Hart, Hila Sheftel, Jean Hausser, Avi Mayo, Michael E Rothenberg, Tomer Kalisky, Uri Alon
PLOS Computational Biology (2015-07-10) <https://www.wikidata.org/wiki/Q35688096>
DOI: [10.1371/journal.pcbi.1004224](https://doi.org/10.1371/journal.pcbi.1004224)

33. **Evolution of Cellular Differentiation: From Hypotheses to Models**
Pedro Márquez-Zacarías, Rozenn M Pineau, Marcella Gomez, Alan Veliz-Cuba, David Murrugarra, William C Ratcliff, Karl J Niklas
Trends in Ecology & Evolution (2020-08-20) <https://www.wikidata.org/wiki/Q98633613>
DOI: [10.1016/j.tree.2020.07.013](https://doi.org/10.1016/j.tree.2020.07.013)

34. **Inferring cell type innovations by phylogenetic methods-concepts, methods, and limitations**
Koryu Kin, Koryu Kin
Journal of Experimental Zoology. Part B: Molecular and Developmental Evolution (2015-10-14) <https://www.wikidata.org/wiki/Q40436539>
DOI: [10.1002/jez.b.22657](https://doi.org/10.1002/jez.b.22657)

35. **An ontology for cell types**
Jonathan Bard, Sue Rhee, Michael Ashburner
Genome Biology (2005-01-01) <https://www.wikidata.org/wiki/Q21184168>
DOI: [10.1186/gb-2005-6-2-r21](https://doi.org/10.1186/gb-2005-6-2-r21)

36. **Logical Development of the Cell Ontology**
Terrence F Meehan, Anna Maria Masci, Amina Abdulla, Lindsay G Cowell, Judith A Blake, Christopher J Mungall, Alexander D Diehl
BMC Bioinformatics (2011-01-05) <https://doi.org/c7kw6x>
DOI: [10.1186/1471-2105-12-6](https://doi.org/10.1186/1471-2105-12-6) · PMID: [21208450](https://pubmed.ncbi.nlm.nih.gov/21208450) · PMCID: [PMC3024222](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024222)

37. **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://doi.org/gg99b9>

DOI: [10.1186/s13326-016-0088-7](#) · PMID: [27377652](#) · PMCID: [PMC4932724](#)

38. **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](#)

39. **Cells in experimental life sciences - challenges and solution to the rapid evolution of knowledge**
Sirarat Sarntivijai, Alexander D Diehl, Yongqun He
BMC Bioinformatics (2017-12-21) https://doi.org/gg99b7
DOI: [10.1186/s12859-017-1976-2](#) · PMID: [29322916](#) · PMCID: [PMC5763506](#)

40. **Cells in ExperimentaL Life Sciences (CELLS-2018): capturing the knowledge of normal and diseased cells with ontologies**
Sirarat Sarntivijai, Yongqun He, Alexander D Diehl
BMC Bioinformatics (2019-04-25) https://doi.org/gg99b8
DOI: [10.1186/s12859-019-2721-9](#) · PMID: [31272374](#) · PMCID: [PMC6509796](#)

41. **The Human Cell Atlas: Technical approaches and challenges.**
Chung Chau Hon, Jay W Shin, Piero Carninci, Michael JT Stubbington
Briefings in functional genomics (2017-10-28) https://www.wikidata.org/wiki/Q48563763
DOI: [10.1093/bfgp/elx029](#)

42. **Towards a pragmatic definition of cell type**
Tiago Lubiana, Helder I Nakaya
Authorea, Inc. (2021-01-04) https://doi.org/ghrxwf
DOI: [10.22541/au.160979530.02627436/v1](#)

43. **PhyloCode** https://www.wikidata.org/wiki/Q1189395

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

45. **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](#)

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

47. **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](#)

48. **SuLab/WikidataIntegrator**
Su Lab
(2021-06-16) https://github.com/SuLab/WikidataIntegrator

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

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

51. **A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses**
Andra Waagmeester, Egon Willighagen, Andrew I Su, Martina Summer-Kutmon, José Emilio Labra Gayo, Daniel Fernández-Álvarez, Quentin Groom, Peter J Schaap, Lisa M Verhagen, Jasper Koehorst
BMC Biology (2021-01-22) https://www.wikidata.org/wiki/Q105037759
DOI: 10.1186/s12915-020-00940-y

52. **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

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

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