**University of São Paulo**

**School of Pharmaceutical Sciences**

**Computational Systems Biology Laboratory**

# Building a biological knowledge graph via Wikidata with a focus on the Human Cell Atlas

PhD's research project presented to FAPESP for scholarship

**Student: Tiago Lubiana Alves**

**Advisor: Prof. Dr. Helder I. Nakaya**

São Paulo, 2019

# Abstract

The Human Cell Atlas is an international effort aiming at characterizing every cell type of the human body. Employing techniques such as single-cell RNA sequencing, mass cytometry, and multiplexed in situ hybridization, it will produce data from virtually all human tissues. This wealth of data can have a significant impact on biomedical research, but only if its content is genuinely available. Wikidata is a knowledge graph database emerging as a FAIR (Findable, Accessible, Interoperable and Reusable) repository for biological knowledge. The formatting and deployment of information from the Human Cell Atlas to Wikidata can increase information availability and impact, by inserting the findings in a network containing multiple associations of concepts of all areas of knowledge (within and outside science). Conceptually defining cell types in a general and applicable concept, formalized into a database-compatible format, is a massive theoretical challenge. This PhD project aims at studying our current understanding of cell types for development a comprehensive ontological model in Wikidata for cell types. We will review the single-cell literature, refining and formalizing concepts for cell type delimitation. Furthermore, we will use Natural Language Processing and Machine Learning tools to automate knowledge extraction from scientific articles in the scope of the Human Cell Atlas. In an advanced step, we will apply concepts of network theory to develop tools for user-friendly querying of the database, making the knowledge ready for the academic community.

# Introduction

## The Human Cell Atlas (HCA) Project

The advent of single-cell technologies has sparked an eagerness in the international scientific community to build a Human Cell Atlas [1]. Since 2017, this project has been running to characterize every cell type in the human body. 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 tropical diseases.

Building a full atlas of human cells comes with multiple challenges. The project includes detection, in single cells, of RNA content (scRNA-Seq), chromatin accessibility (scATAC-Seq), and protein markers (primarily by CYTOF), as well as spatial information on cells with multiplexed FISH approaches (such as MERFISH) and imaging mass cytometry[12]. Every lab will contribute with its expertise, providing samples that are representative of human diversity.

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. However, for HCA to maximize its impact, the distribution of the generated knowledge needs to be FAIR: Findable, Accessible, Interoperable, and Reusable[3].

## Natural language as a limitation for the HCA

HCA reports include not only raw data but the analysis and interpretations, made available as scientific articles written in natural language. The automation of knowledge handling is promising, and Natural Language Processing has shown that information in the literature can be automatically harnessed to generate hypotheses with a high level of confidence[45]. Algorithms based on neural networks, such as Word2Vec [67], find “word embeddings,” multidimensional representations of the context of words in a text. Tshitoyan et al. used[4] a neural network to infer characteristics of materials based on literature abstracts. Another example is the knowledge system IBM Watson, which has predicted genes related to Amyotrophic Lateral Sclerosis[5].

Despite being powerful, these approaches for processing natural language are coarse, due to the difficulties in automatically parsing and understanding natural language. If we want to take advantage of the Human Cell Atlas to the maximum, we need to integrate the data across different domains of knowledge. The format of multi-level ontologies (such as Wikidata) is flexible enough to add biological processes and yet rigorous enough to allow straightforward machine processing.

## Human Cell Atlas and Cell Ontology

The modeling of biomedical concepts is a vast field of research. The  Open Biological and Biomedical Ontology (OBO) Foundry[8] has set a series of principles for openness and standardization, which has guided the creation of interoperable ontologies (structure vocabulary for concepts and relations). One of these, the Cell Ontology, aims to build a controlled vocabulary for cell types in animals, providing working definitions for the cell types and has been built by an active volunteering community since 2005 [910]. As of the latest release (August 2019), the Cell Ontology reports about 2,300 classes for cells (https://bioportal.bioontology.org/ontologies/CL).

The representation of different cell types is already in the scope of the HCA, and a Chan-Zuckerberg Initiative awarded an HCA grant to researchers (https://grants.czi.technology/, cell types) related to the Cell Ontology project. The grantees' mission is "to extend our previous work developing biomedical ontologies(...) to develop a scalable approach for semantically-coherent and statistically-comparable cell type definitions"." The team has produced articles describing challenges and possible ways of formally describing cell types and states in the context of Human Cell Atlas massive data generation [11,12].

## Wikidata

Independent ontologies, however, are domain restricted. That means they do not communicate directly with knowledge outside their specialized community. Wikidata can solve that issue as it is a massive free, open-source tool for multilevel ontology by providing a platform that integrates knowledge across domains. It has an active community developing systems to query, visualize, and use this knowledge. Wikidata, then, would be an outstanding tool to organize, visualize, and integrate the knowledge produced in the Human Cell Atlas (**Figure 1**).

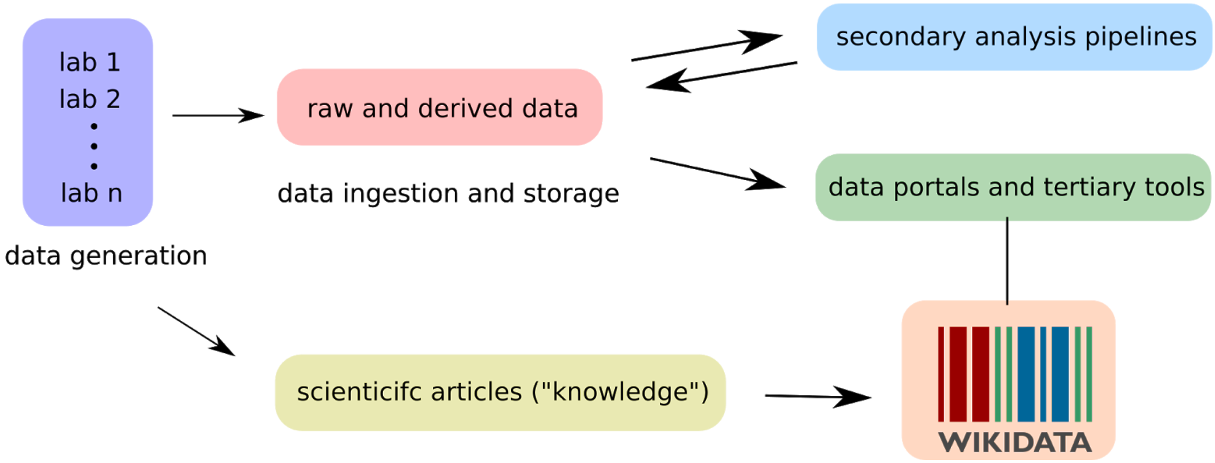
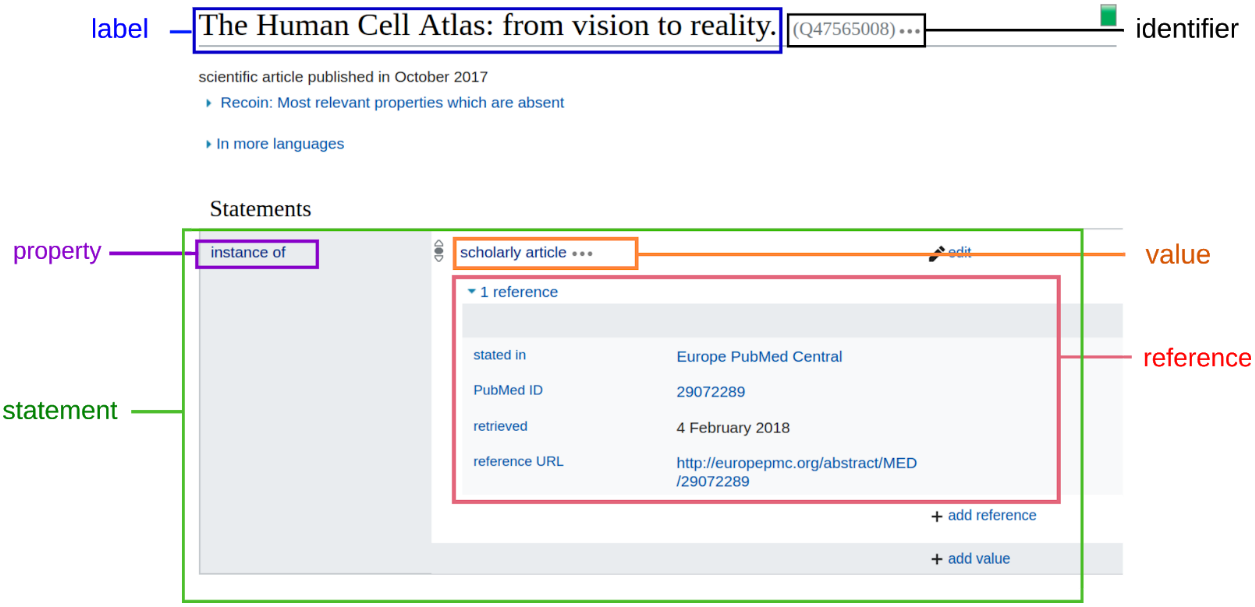

lab 1
lab 2
lab n
data generation
raw and derived data
data ingestion and storage
secondary analysis pipelines
data portals and tertiary tools
scienticifc articles ("knowledge")
WIKIDATA

**Figure 1. Overview of how this project inserts itself into the Human Cell Atlas workflow**. A simplified version of the data curation schematics present at the HCA white paper. Modeling and migration to Wikidata can be a major data portal/tertiary tool and, at the same time, it can receive high-quality information from the interpretation of results published in HCA's articles.

Wikidata is run by the Wikimedia Foundation, the same group that manages Wikipedia. It is a structured database of knowledge, and concepts are represented based on items, properties, and values (**Figure 2**). Items can be present any object or concept, from cell types to scientific articles, and properties are descriptors for items, encoding relevant information. Properties link items to values that can be dates, numbers, words, or even other items.



The Human Cell Atlas: from vision to reality. (Q47565008)

scientific article published in October 2017

Statements

instance of: scholarly article

1 reference

| | |
|---|---|
| stated in | Europe PubMed Central |
| PubMed ID | 29072289 |
| retrieved | 4 February 2018 |
| reference URL | http://europepmc.org/abstract/MED/29072289 |

**Figure 2. Overview of an item in Wikidata**. The label is a tag specific for each language. The identifier is a multilingual tag, unique for each item/concept. Statements are based on properties, which connect an item to a specific value. Values can be items themselves, building a web of knowledge.

Wikidata contains more than 1 billion edits, and 60 million items on the public domain, it can be updated by anyone and it is run by an active, decentralized, international community. Wikidata is one of the powerhouses of Google's Knowledge Graph, and Google has migrated its base, Freebase, to Wikidata in 2016[13]. Wikidata provides integration with different domains of knowledge, creating the opportunity of inserting biological knowledge in an interdisciplinary context. It has query systems, and active developers building visualizations and edition tools. Finally, there are packages in major programming languages dedicated to interfacing Wikidata (WikidataR in R (https://github.com/Ironholds/WikidataR) and Wikidata integrator in Python (https://github.com/SuLab/WikidataIntegrator).

A few projects have worked on adding biomedical information to Wikidata. The WikiGene team [14] has added information about genes and proteins in Wikidata and created two Wiki projects (efforts to improve a topic) on molecular biology and microbiology. A database of gene annotations from the Wellcome Sanger Institute, GeneDB, has been fully integrated into Wikidata[15]. WikiProject Medicine(www.wikidata.org/wiki/Wikidata:WikiProject_Medicine) has also worked towards structuring information about medical knowledge, modeling clinical trials, adding identifications from disease databases, and creating properties to describe anatomical parts. Besides ongoing efforts, recent reviews have highlighted the possibilities of Wikidata for managing medical knowledge[16] and for making biological knowledge more FAIR 15.

However, Wikidata still lacks much information for cell biology, and even major cell types lack basic descriptions. For example, as of September of 2019, “astrocyte” has 18 statements in Wikidata, and the movie “Dumb and Dumber” has 109 statements. Currently, Wikidata hosts two hundred forty-two items correspond to “cell types” - and we added more than one hundred of these as preliminary work for this project. The Wikidata community has yet to create representations of crucial features, such as marker genes and possible cell states, a task which requires domain expertise (in this case, about cell biology) and mastery of the inner workings of the database. [17] [18]

## The concept of cell type

For Wikidata to be fully applicable to the Human Cell Atlas endeavor, we need to refine the concepts of “cell types” and “cell states.” The challenge is clearly stated in the publication by Regev et al. defining the goals of the Human Cell Atlas[1]: “we lack a rigorous definition of what we mean by the intuitive terms ’cell type’ and ’cell state.’ (...) The boundaries between these concepts can be blurred because cells change over time in ways that are far from fully understood. Ultimately, data-driven approaches will likely refine our concepts.”

The concept of cell type in mature organisms is actively debated by prominent researchers, suggesting different ways of integrating cell function, localization, lineage, and expression programs, even calling it a “paradigm shift on the issue of cell type”[19]. A recent review on single-cell workflows avoids the definition altogether, and chooses, due to the definition issues, to use the term “cell identity” instead[20]. It is clear that the refinement of the concept of “cell type” can be of great value to the scientific community.

The challenge is similar to the one faced in the definition of species concepts[21,22]. Nevertheless, we have a wealth of theoretical material, and many key discussion points are already catalogued[21]. The reproductive isolation criterion, for example, and the idea of the hierarchy of genus, family, order, and phylum are already incredibly useful for describing and teaching biology. Meanwhile, the definition of cell types stands in a “pre-Linnaean” era, where we lack even a standardized nomenclature for cell types.

Our goal will not be to find a ground truth of what a cell type is. We aim at providing operational definitions of cell type and cell states in a way that we can satisfactorily enter this information in the Wikidata knowledge base. The concept needs to be recognizable by the community, useful for cataloging information, and rigorous enough to solve significant ambiguities. The iteration with Wikidata allows further refinements by integrating feedback from multidisciplinary backgrounds. The quality of the data model and the quality of the concepts.

## Objectives

This Ph.D. project aims at formalizing our understanding of cell biology in a knowledge graph, making the contents of the Human Cell Atlas interoperable. We will develop a model on Wikidata for primary cell types, modeling features as locations, morphologies, and genetic markers produced within the scope of HCA. Scientific and philosophical challenges built in this process includes questions as:

- “Is the Wikidata sufficient for describing the core information about cells? Would we need different formal representations of the complex network of knowledge?”
- “What is a cell type? What is a cell state? What is a cell marker? Are precise definitions of these terms possible?”

### Specific goals

- 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, in order 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.

## Methodology

The general description of our proposed methodology is depicted in the roadmap in **Figure 3** and each step is detailed below.

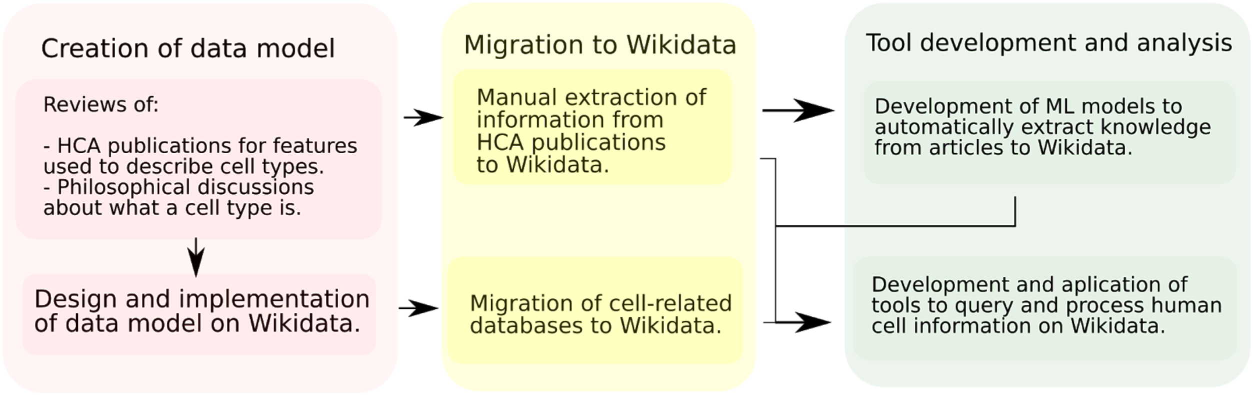

Creation of data model
Reviews of:
- HCA publications for features used to describe cell types.
- Philosophical discussions about what a cell type is.
Design and implementation of data model on Wikidata.
Migration to Wikidata
Manual extraction of information from HCA publications to Wikidata.
Migration of cell-related databases to Wikidata.
Tool development and analysis
Development of ML models to automatically extract knowledge from articles to Wikidata.
Development and aplication of tools to query and process human cell information on Wikidata.

**Figure 3. Methodology roadmap**. Conceptual map of the methodological tasks included in this project. Arrows indicate steps that depend on the others. HCA: Human Cell Atlas. ML: Machine Learning.

## Building a data model to describe cell types

To provide an accurate description of what a cell type is, we will perform two reviews: i) screen publications of the Human Cell Atlas for parameters of cell types and, ii) a review of the conceptual discussions about what a cell type is (example in ref. 18[19]). We present examples of data models in the results section.

The data model will consist of a set of Wikidata properties and their constraints necessary and sufficient to describe a cell type. It will be based on the two reviews described above, and documented openly on Wikidata, proposing the modeled concepts as new properties (https://www.wikidata.org/wiki/Wikidata:Property_proposal). All proposed properties will be reviewed by the Wikidata community before being approved.

## Migration of databases

Databases of cell markers (PanglaoDB (https://panglaodb.se/)) and cell types (http://www.obofoundry.org/ontology/cl.html) will be filtered for stringent relations and matched to Wikidata items. Database processing and reconciliation/matching to Wikidata be done in

Python 3 based on WikidataIntegrator https://github.com/SuLab/WikidataIntegrator) and with the aid of the OpenRefine framework(https://openrefine.org/).

## Manual curation of Human Cell Atlas publications

Publications of the Human Cell Atlas (https://www.humancellatlas.org/publications/) will be manually checked for mentions of cell types. Any mention of a previously unindexed cell type will be added to Wikidata, with its respective features, and referencing the article where the existence of the cell type is stated. This curation will provide a set of “ground truth” cell types and features that will be used for the automation of the process.

## Natural Language Processing and knowledge extraction.

Automated extraction of cell types and properties from articles will rely on the SPIED (Stanford Pattern-based Information Extraction and Diagnostics)[23], the Python Spacy NLP module (spacy.io/) and Word2Vec[6] neural networks using open cell biology articles. The number of articles to be processed will depend on project progress. As a rule-of-thumb, we aim to process the top 100 most relevant (by PubMed’s algorithm) articles in a search for ‘ "single-cell RNA-sequencing" AND "human"’.  A user interface will be developed for selecting candidate hits from the automated tools. A similar open-source interface is https://appstract.pub/.

## Analysis and exploration of integrated data

Tools to analyze the knowledge network derived from our work will include the native SPARQL query service (www.wikidata.org/wiki/Wikidata:SPARQL_query_service/) and general-purpose visualization tools (www.wikidata.org/wiki/Wikidata:Tools/Visualize_data). In addition, we will implement a question and answer tool similar to QAnswer (https://qanswer-frontend.univ-st-etienne.fr/about) targeted to biologists. The details of the implementation will be defined during the execution of the project.

# Preliminary results

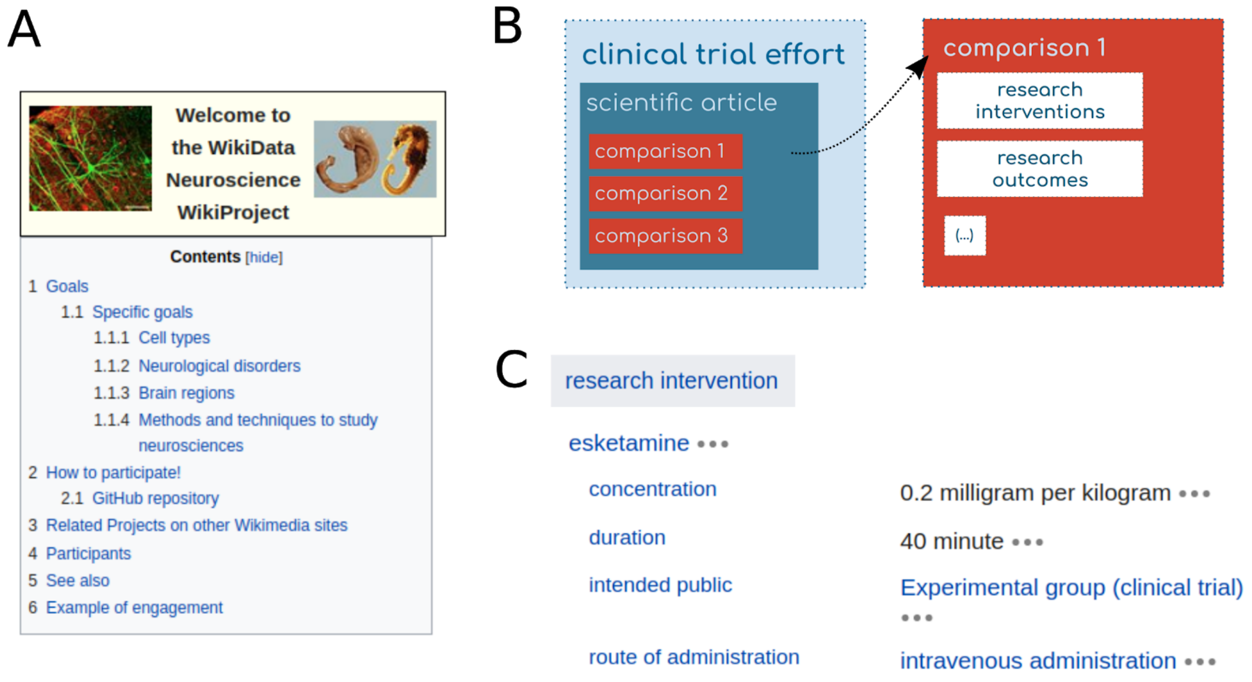
## Pilot projects on Wikidata for biology

In the preparation for this PhD project, we pitched the idea of integrating biomedical knowledge to Wikidata for two different hackathons, receiving travel awards for developing it in Rio (Brazil) and Cambridge (United Kingdom). In this session, we will describe the previous work on both these events. The practical experience acquired in both events has shaped the idea of the project and shared goals and methods with the ones proposed here.

# No-Budget Science Hack Week

The No-Budget Science Hack Week was a week-long event organized by the Brazilian Reproducibility Initiative team[24,25] in Rio de Janeiro. At the event, we worked in a five people group, and created a WikiProject Neuroscience on Wikidata. We adapted and improved a model for Clinical Trials, creating ontological models for how to represent this type of research.

In our model, we define a Clinical Trial as a research effort that produces one or more scientific articles. A scientific article is composed of one or more comparisons, each of them can be described in semantic format. As a proof-of-concept, we worked with the trial “A Study of the Efficacy of Intravenous Esketamine in Adult Patients With Treatment-Resistant Depression “ (www.wikidata.org/wiki/Q63842675; clinicaltrials.gov/ct2/show/NCT01640080), reported in the article: “Intravenous Esketamine in Adult Treatment-Resistant Depression: A Double-Blind, Double-Randomization, Placebo-Controlled Study.“ (www.wikidata.org/wiki/Q34506620[26]). One of the comparisons in the article is the “Comparison of .2 mg/kg esketamine to placebo” (/www.wikidata.org/wiki/Q65950239). The details of the experimental group of the comparison are reported in **Figure 4.**



Welcome to the WikiData Neuroscience WikiProject

Contents [hide]

1 Goals
1.1 Specific goals
1.1.1 Cell types
1.1.2 Neurological disorders
1.1.3 Brain regions
1.1.4 Methods and techniques to study neurosciences
2 How to participate!
2.1 GitHub repository
3 Related Projects on other Wikimedia sites
4 Participants
5 See also
6 Example of engagement

research intervention

esketamine

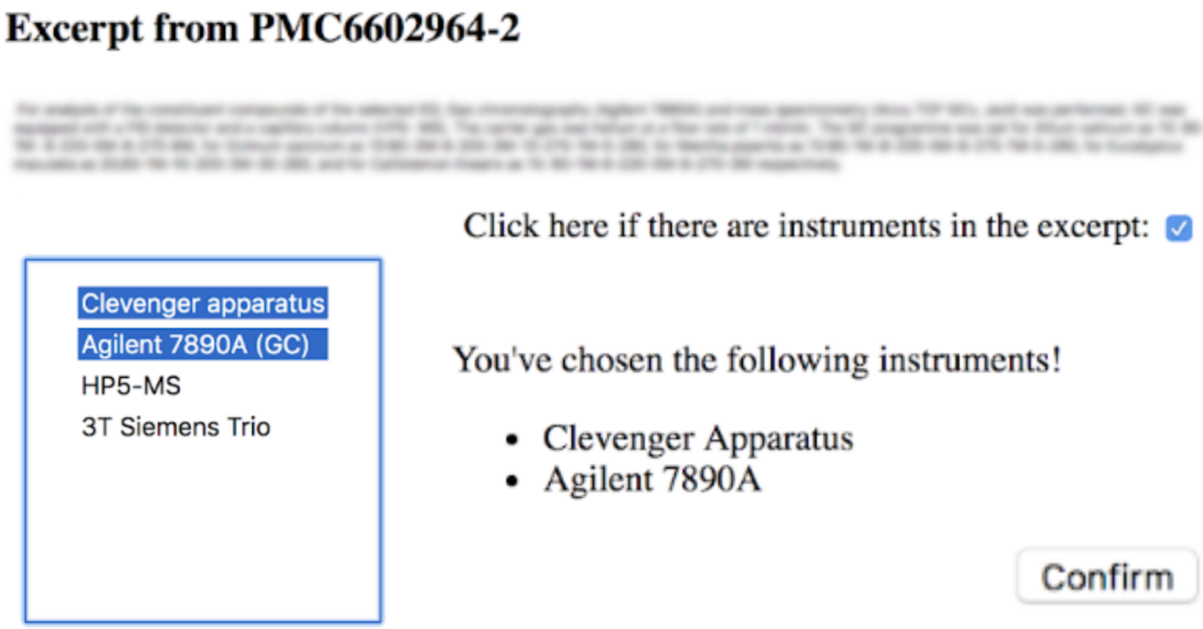
| | |
|---|---|
| concentration | 0.2 milligram per kilogram |
| duration | 40 minute |
| intended public | Experimental group (clinical trial) |
| route of administration | intravenous administration |

**Figure 4. Work by the NeuroWiki group at No-Budget Science Hack Week 2019**. A) Frontpage of the Wikidata Neuroscience Wikiproject created at the hack week. B) Schematics of how core concepts relate to each other. The clinical trial endeavor produces articles, which include comparisons. A comparison has a specific research intervention. D) Research intervention in comparison inside the article Q34506620 [26].

## eLife Innovation Sprint 2019

After the No-Budget Science Hack Week, we attended the eLife Innovation Sprint [27], a hackathon organized by the eLife platform for research communication. During the two-day event in Cambridge, United Kingdom, we worked (in a 6-people group) to extract information about which scientific equipment was used in a dataset of 1000 articles related to essential oils.

From a list of instruments derived from the eagle-i system[28], we built to predict which words in a text document could be scientific instruments, given their contexts. Then, we uploaded equipment data to Wikidata, connecting articles to the equipment they use.



Excerpt from PMC6602964-2

[illegible]

Click here if there are instruments in the excerpt: ☑

Clevenger apparatus
Agilent 7890A (GC)
HP5-MS
3T Siemens Trio

You've chosen the following instruments!

- Clevenger Apparatus
- Agilent 7890A

**Figure 5**. **User interface by the instruMinetal group at eLife's 2019 sprint**. A user interface to select of names of scientific equipment out of the terms ranked high by the natural language processing step.

Our team used Word2Vec[6] and regular expression to identify putative scientific instruments. The candidate list was made to be validated by crowdsourcing in a user interface (**Figure 5)**. The final output included statements as follows: a specific article [29] (Q28817266) *describes a project that uses* (P4510) the Shimadzu QP-5000 equipment (Q67146252). The code written for all three parts is available at GitHub(github.com/caffiendFrog/elife2019).

## Human Cell Atlas examples: the liver and the brain

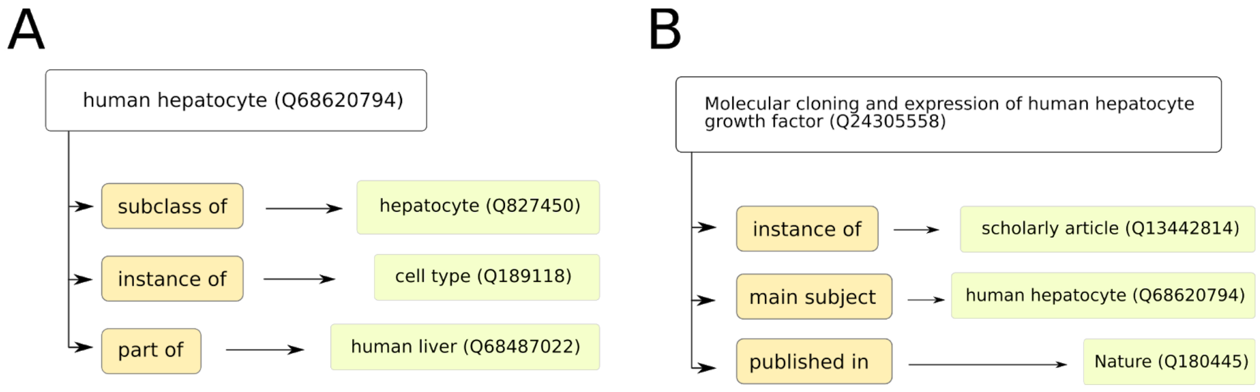
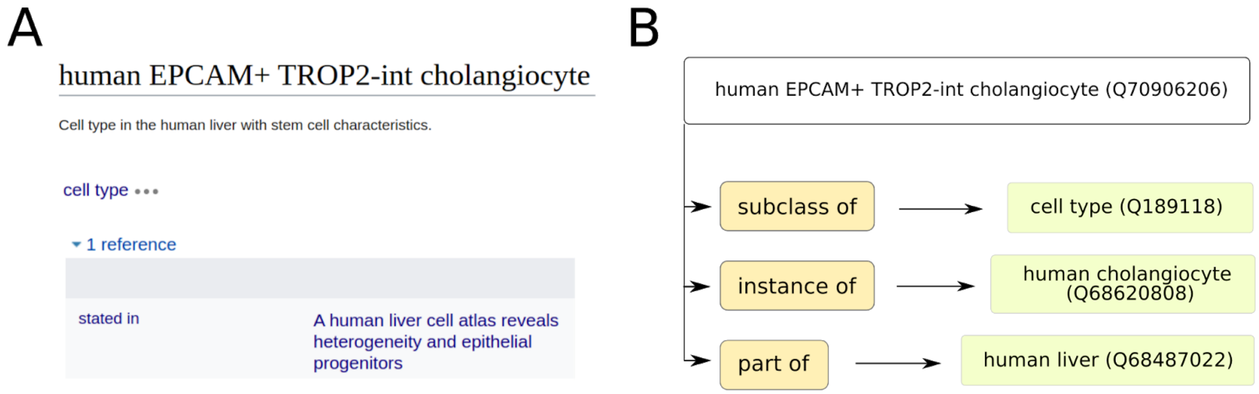

A
human hepatocyte (Q68620794)
subclass of
hepatocyte (Q827450)
instance of
cell type (Q189118)
part of
human liver (Q68487022)
B
Molecular cloning and expression of human hepatocyte growth factor (Q24305558)
instance of
scholarly article (Q13442814)
main subject
human hepatocyte (Q68620794)
published in
Nature (Q180445)

**Figure 7**: **Structuring information about cell types in Wikidata.** A) Interaction graph for the item human hepatocyte. B) Interaction graph for an article published in Nature linked to human hepatocyte by the *main subject* property (https://www.wikidata.org/wiki/Property:P921) .

In order to access the feasibility of the specific project of using Wikidata for representing information of the Human Cell Atlas (HCA), we took the information present on the descriptive figure in HCA's white paper describing the goals of the project. The liver and the brain were the organs of choice, as the most recent publication of the HCA project targeted liver[30] and a recent paper on brain cells included an explicit ontology as supplementary figure[31]. The data model used for organs, cells and similar entities was developed as the example below.

The term "Human Liver" (Q68487022) describes all livers of Homo sapiens. The liver of one specific human being is an instance of a human liver. Steve Jobs' transplanted liver (Q72303384), for example, would be an **instance of** Q68487022. The item "Human Liver" itself is a subclass of "Liver" (which describes this organ in a broader manner, for all species). This same idea can be applied to cells and related entities. We added major liver entities to Wikidata in the same fashion as "human hepatocyte" (Q68620704, **figure 7).**



human EPCAM+ TROP2-int cholangiocyte

Cell type in the human liver with stem cell characteristics.

cell type

1 reference

| | |
|---|---|
| stated in | A human liver cell atlas reveals heterogeneity and epithelial progenitors |

human EPCAM+ TROP2-int cholangiocyte (Q70906206)
subclass of
cell type (Q189118)
instance of
human cholangiocyte (Q68620808)
part of
human liver (Q68487022)

**Figure 8 Structuring HCA's discoveries.** A) EPCAM+ TROP2-int cholangiocyte (Q70906206) item, referenced to the paper which describes it. B) Interaction graph for the item in A.

We used Python scripts combined with SPARQL queries and identified 66 articles containing the words "human hepatocyte" in the title. Then, we used the batch edition tool "Quickstatements" (**Figure 7B**) to edit items describing articles about human hepatocytes via the property **main subject** (P921). An example can be seen in **figure 7D**.

We then added cells in the HCA publication paper by Aizarani et al [30]. They report EPCAM+ TROP2-int cholangiocytes as stem cells important for regeneration in the liver, and we created an item for this cell type, framed as an instance of cell type and a subclass of stem cell (**figure 8A-B)**. Noticeably, this is the only cell type in Wikidata which is labeled as a stem cell in Wikidata, denoting that the database still lacks core information (**Figure 8C**).

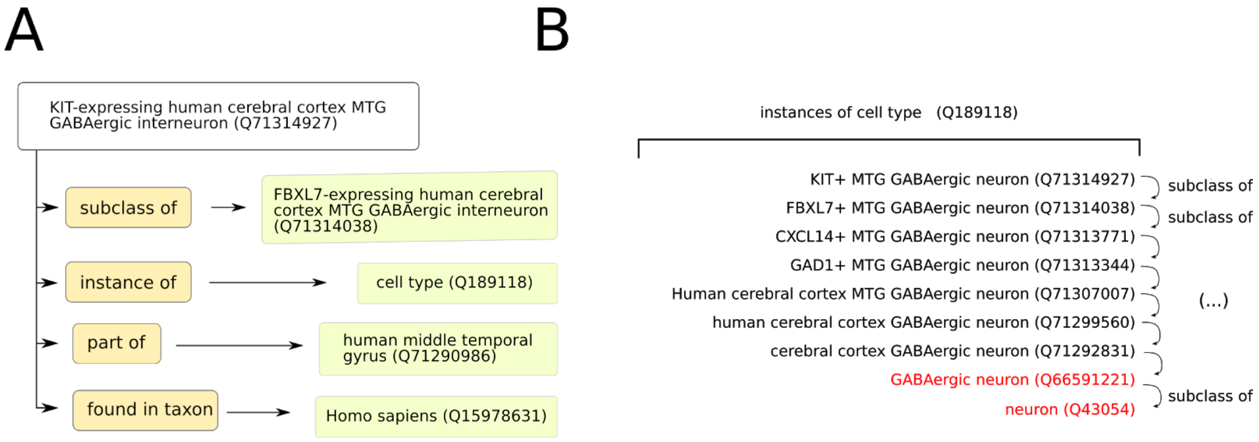

A
B
KIT-expressing human cerebral cortex MTG GABAergic interneuron (Q71314927)
subclass of
FBXL7-expressing human cerebral cortex MTG GABAergic interneuron (Q71314038)
instance of
cell type (Q189118)
part of
human middle temporal gyrus (Q71290986)
found in taxon
Homo sapiens (Q15978631)
instances of cell type (Q189118)
KIT+ MTG GABAergic neuron (Q71314927)
FBXL7+ MTG GABAergic neuron (Q71314038)
CXCL14+ MTG GABAergic neuron (Q71313771)
GAD1+ MTG GABAergic neuron (Q71313344)
Human cerebral cortex MTG GABAergic neuron (Q71307007)
human cerebral cortex GABAergic neuron (Q71299560)
cerebral cortex GABAergic neuron (Q71292831)
GABAergic neuron (Q66591221)
neuron (Q43054)
subclass of
subclass of
(...)
subclass of

**Figure 9. migration of cell ontology from human medial temporal gyrus.** A) Example of relations assigned to one tip of the ontology tree (KIT-positive human cerebral cortex GABAergic interneuron. B) Taxonomic relations of each cell type in accordance to Hodge et al. Red color denotes cell types already present in Wikidata.

Finally, to test the feasibility of migrating databases, we added 78 cell types of the human cortex described by Hodge et al [30]. The authors - which included the grantee of a Chan Zuckerberg Grant for developing ontologies - provide an ontology of cell types as a supplementary table, but without support for Wikidata. We migrated all cell types and relations in the table to the Wikidata. An example of how the items were structured it is integrated is detailed in **Figure 9**. All the code is available at https://github.com/lubianat/wikidata_hca.

## Deliverables:

The development of a data model to define cell types is the primary deliverable, and arguably the most demanding. It will require domain-specific knowledge (via reviews outlined in the Methodology section) and active contact with the Wikidata community, which evaluates and approves data models. It is modular, however, and can be scaled depending on time constraints (e.g. adding more or fewer descriptors for cell types), mitigating the risk associated with this fundamental step.

Manual migration of HCA articles and cell database information is a curation effort that may be time-consuming, but the annotation is important for the later parts of the project: automatic extraction of knowledge and querying of the Wikidata database. The development of tools presents programming challenges, but are within reach considering the laboratory expertise.

**Table 1 - Chronogram of Activities. Columns represent semesters.**

| **Activities** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| Create a Wikidata data model for cell types | P | P | s | s | | | | |
| Structure and migrate  items in HCA publications | s | P | P | | | | | |
| Migrate Cell Related databases to Wikidata | | s | P | | | | | |
| Write article on data model and migrations | | | s | P | | | | |
| Develop and apply tools for automatic extraction | | | s | s | P | P | | |
| Develop and apply tools to query the Wikidata graph | | | s | s | P | P | | |
| Write a summary article | | | | s | P | P | P | |
| Write a thesis | | | | | s | P | P | P |

P: primary activity, s: secondary activity.

# Bibliography

1. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).
2. Regev, A. *et al.* The Human Cell Atlas White Paper. (2018).
3. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
4. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
5. Bakkar, N. *et al.* Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis.

*Acta Neuropathol.* **135**, 227–247 (2018).

6. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. (2013).
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. (2013).
8. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
9. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, 1–5 (2005).
10. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
11. Bakken, T. *et al.* Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* **18**, 559 (2017).
12. Aevermann, B. D. *et al.* Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* **27**, R40–R47 (2018).
13. Tanon, T. P., Vrandečić, D., Schaffert, S., Steiner, T. & Pintscher, L. From Freebase to Wikidata. *Proceedings of the 25th International Conference on World Wide Web - WWW '16* (2016) doi:10.1145/2872427.2874809.
14. Burgstaller-Muehlbacher, S. *et al.* Wikidata as a semantic framework for the Gene Wiki initiative. *Database* **2016**, (2016).
15. Manske, M., Böhme, U., Püthe, C. & Berriman, M. GeneDB and Wikidata. *Wellcome Open Res* **4**, 114 (2019).
16. Turki, H. *et al.* Wikidata: A large-scale collaborative ontological medical database. *J. Biomed. Inform.* **99**, 103292 (2019).
17. Atkinson, C. & Kühne, T. The Essence of Multilevel Metamodeling. *≪UML≫ 2001 — The Unified Modeling Language. Modeling Languages, Concepts, and Tools* 19–33 (2001) doi:10.1007/3-540-45441-1_3.
18. Carvalho, V. A. & Almeida, J. P. A. Toward a well-founded theory for multi-level conceptual modeling. *Software & Systems Modeling* vol. 17 205–231 (2018).
19. What Is Your Conceptual Definition of 'Cell Type' in the Context of a Mature Organism? *Cell Syst* **4**, 255–259 (2017).
20. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

21. De Queiroz, K. Species Concepts and Species Delimitation. *Systematic Biology* vol. 56 879–886 (2007).

22. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).

23. Gupta, S. & Manning, C. SPIED: Stanford Pattern based Information Extraction and Diagnostics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (2014) doi:10.3115/v1/w14-3106.

24. de Oliveira Andrade, R. Brazilian biomedical science faces reproducibility test. *Nature* **569**, 318–319 (2019).

25. Amaral, O. B., Neves, K., Wasilewska-Sampaio, A. P. & Carneiro, C. F. The Brazilian Reproducibility Initiative. *Elife* **8**, (2019).

26. Singh, J. B. *et al.* Intravenous Esketamine in Adult Treatment-Resistant Depression: A Double-Blind, Double-Randomization, Placebo-Controlled Study. *Biol. Psychiatry* **80**, 424–431 (2016).

27. Innovation Sprint 2019: Project roundup. *elifesciences.org* (2019).

28. Vasilevsky, N. *et al.* Research resources: curating the new eagle-i discovery system. *Database* **2012**, bar067 (2012).

29. da Silva Ramos, R. *et al.* Chemical Composition and Antioxidant, Cytotoxic, Antimicrobial, and Larvicidal Activities of the Essential Oil of L. (Lamiaceae). *ScientificWorldJournal* **2017**, 4927214 (2017).

30. Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).

31. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).