

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

This manuscript ([permalink](#)) was automatically generated from [lubianat/quali\\_phd@8e4db1e](#) on December 8, 2021.

## Authors

- **Tiago Lubiana**

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

School of Pharmaceutical Sciences, University of São Paulo; Ronin Institute · Funded by Grant #2019/26284-1 from the São Paulo Research Foundation (FAPESP).

## Abstract

The Human Cell Atlas (HCA) is an international effort aiming at characterizing every cell type of the human body. By the virtue of techniques such as single-cell RNA sequencing, mass cytometry, and multiplexed in situ hybridization, HCA members are producing cell-level 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 interoperable. While ontologies and semantic technologies have emerged as key players in the data interoperability ecosystem, there are still gaps to cover between the technical possibilities and the practical applications in biomedical research. In addition to ontologies, like the Cell Ontology and the Gene Ontology, large-scale knowledge graphs are growing as a tool for knowledge management. Among those, Wikidata, a sister project of Wikipedia for structured data, is surfacing as a hub in the semantic web for multiple types of information. The formatting and deployment of information from the Human Cell Atlas to Wikidata can increase information availability and impact, connecting the scientific products with the larger knowledge ecosystem. This PhD project aims at studying Wikidata as a platform for representing cell types, addressing theoretical and practical concerns.

We are reviewing the literature on cell types, refining and formalizing concepts for cell type delimitation. At the same time, we are enriching Wikidata with new classes curated from the literature, and with large scale integrations of biomedical databases (e.g. PanglaoDB) into the Wikidata infrastructure. To aid that effort, we are developing Wikidata Bib, a framework for literature management and organized note-taking and recording system for reading the academic literature with high efficiency. Finally, we plan to improve the interplay of Wikidata, the Cell Ontology and software used for single-cell RNA-seq data, inserting Wikidata *de facto* as a tool for the Human Cell Atlas community.

## Preface

Here we present an overview of the different chapters that compose this document, presented as the text for a qualifying exam.

This work is concerned with the conceptual modelling of knowledge about cell types. The introduction contains an overview of the Human Cell Atlas project, and the current state of classification of cells into types. Then, it proceeds to introduce ontologies and knowledge graphs as tools for connecting what we know about cells.

The methodology section is an overview of the core methods used throughout the work. However, as the project contains elements from different scientific traditions, the results chapters might also display particular methods used in the specific branch of the project.

It is worth noticing that the different results shown were not developed chronologically in the order shown. They were actually developed in parallel, with overlapping periods of activity. They have been organized into separate chapters, however, as they tackle different perspectives of the subject matter, and are part of different publications.

The discussion on the concept of cell type is presented first, as it is instrumental for the later steps. It is followed by an account of how PanglaoDB, a database of cell markers, was integrated to Wikidata, based on a notion of species-specific cell type clarified in the preceding chapter.

Then, we present Wikidata Bib, a framework for organized reading of the literature. The framework, although used as a method throughout the PhD project, is presented in the results session. This emphasis as a result was chosen as the technical and theoretical details of the system are part of the intellectual work put into the project. The system evolved into a biocuration platform for the collection of cell types from the literature to Wikidata, and the statistics on this curation are also presented on the section.

To end the results, we discuss how our efforts integrate with the Cell Ontology, the currently leading system for organizing cell types.

Finally, an account of other academic aspects of the project are presented, as part of the qualification requirements. They present an overview of collaborations, participation in events and academic courses taken during the first part of the PhD project.

## Background

### The Human Cell Atlas (HCA) Project

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 gathers people from all over the world to tackle different parts of the project, so to have a diverse and equitable account of the cell type diversity. [3]

Building a full atlas of human cells comes with multiple challenges. The project includes the detection, in single cells, of RNA species (scRNA-Seq), chromatin accessibility (scATAC-Seq), and protein markers (primarily by CYTOF), as well as spatial information on cells with multiplexed *in situ* hybridization (such as MERFISH) and imaging mass cytometry [2,4]. Every lab inside the project 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. 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 [5] to specialized personnel [5].

The Human Cell Atlas has a dedicated team for organizing data: the Data Coordination Platform (DCP) [6] [4]. The DCP is responsible for tracing the plan for computational interoperability, from the data generators to the consumers.[4]. The Human Cell Atlas has its portal for data (<https://dahttps://www.wikidata.org/wiki/Help:Multilingualta.humancellatlas.org/>) which composes the data repository landscape with other resources, like the Broad Institute Single Cell Portal ([https://singlecell.broadinstitute.org/single\\_cell](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 highlight of “knowledge” in the last paragraph is meant to stress that raw data *per se* is not enough to turn the Atlas objectives into reality. There is a long way from raw datasets to commonly agreed scientific knowledge. And, ultimately, this long way is what allows humanity to take advantage of scientific endeavors. Currently, the gap between data and knowledge is mostly targeted via the writing and sharing of scientific manuscripts, the *de facto* currency of exchange of claims about the natural world. The Human Cell Atlas Publication Committee reviews and selects publications that are directly part of the HCA. A set of publications is, thus, one of the major outputs of the whole endeavor.

The challenge that arises, thus, is one of managing a wealth of information and cast it into useful science. Experimental articles that analyze thousands of cells pose an overload of information alone. Ideally, we would like to understand, remember and make use of every statement produced by the HCA. As this goal is humanely impossible, we need to develop tools to make the knowledge interoperable with the aid of computers. At that point, the challenges of the HCA enter in resonance with the challenges of text-mining, biocuration and literature based discovery, which will be discussed in the chapter of this introduction.

## Classification of cells into types

Given that a core goal of the Human Cell Atlas is to advance knowledge about *all* human cell types, [2] the definition of what a cell type is becomes important. Although a number of views exist [1,7,8,9,10,11,12,13,14,15,16,17,18], there is no formal, commonly agreed upon definition of cell type. A 2017 article on the Human Cell Atlas mentions[10]:

“Descriptors such as ‘cell type’ and ‘cell state’ can be difficult to define at the moment. An integrative, systematic effort by many teams of scientists working together and bringing different expertise to the problem could dramatically sharpen our terminology, and revolutionize the way we see our cells, tissues and organs. We invite you to join the effort.” The article highlights both the current gap in knowledge and the need of a community effort to work in that direction, in a direction that justifies the existence of the present work.

One consequence of a lack of a definition is that there is no commonly agreed number of cell types, and not even on an order of magnitude. As of November 2021, the leading answers in the Google Search Engine for the question “How many different cell types are found in the human body?” all point to around 200 different types (<https://askabiologist.asu.edu/questions/human-cell-types>, <https://www.researchgate.net/post/How-many-cell-types-in-a-human-body-How-about-the-number-of-cell-cycles-in-each-species>, <https://www.kenhub.com/en/library/anatomy/types-of-cells-in-the-human-body>), an estimate that is agreed upon by Bionumbers, a database of useful biological numbers [19] (<https://bionumbers.hms.harvard.edu/bionumber.aspx?id=103626>). A list of cell types in the adult human body on Wikipedia also amounts to around a couple hundred cell types [20], [=List\\_of\\_distinct\\_cell\\_types\\_in\\_the\\_adult\\_human\\_body&oldid=1044853788](#). However, the Cell Ontology has so far had catalogued 2,311 cell types of interest for the Human Cell Atlas as of June 2021 [21], increasing the estimate by at least one order of magnitude. Additionally, with an estimate of 37 trillion cells on average per human body [22] and an ever-increasing report of new cell types/clusters in single-cell transcriptomics ([23]), a precise estimate is not reasonable. In fact, the Human Cell Atlas project itself does not commit to any estimates of numbers of cell types, due to the sheer difficulty of estimating a number given current knowledge. (Aviv Regev; reply to question in the HCA conference)

Even though there is no agreement, different views on cell types are maturing. One core line of thought to define “cell type” is based on the cell type as an evolutionary unit defined by a Core Regulatory Complex (CoRC) of transcription factors. That definition enables the drawing of parallels, from the evolution of other biological entities (such as genes, proteins, and species) to cell types’ evolution. Models of how multicellular life works greatly benefit from concepts such as “sister types” (cell types that diverged from a single ancestor), “cell type homology” (cell types in different species that share a common evolutionary origin), and “cell type convergence” (cell types that execute similar functions but which are not directly evolutionarily related) [24,25]

Another direction is based on the notion of attractors: regions of dynamical stability in a feature space, which might have different qualities. [26,27] In this theory, “basins of attraction” direct cell phenotypes, providing points in, say, a gene expression space towards which different cells “move” their expression programs. This dynamic view see each cell type corresponding to “a self-stabilizing regulatory program, which acts to maintain and restore the cell type-specific program of gene expression.” [28] It aligns itself with dynamic systems theory, and some authors go as far as to say that “Lacking the idea of attractors we have no clear idea of what a cell type is.” [29]

As much as different concepts of species coexist [30], our quest to define cell types may take various forms. The challenge of representing cell types in the context of evolution is conceptually different from representing cell types in biomedical experimentation. In that second direction, the groundwork of the Cell Ontology [31,32,33] and CELDA [34] and the contributions of the International Workshop on Cells in Experimental Life Sciences series [35,36] are notable.

Even though many sources of knowledge contribute to our knowledge about cell types [37], arguably single-cell transcriptomics is the workhorse for current efforts of the HUMAN cell Atlas, with an increasing amount of published studies using the methodology and of cells per study. [37] Current scRNA-seq data analyses often rely on unsupervised clustering of cells followed by assignment of cell-type labels to clusters. For the clustering, bioinformaticians tailor parameter sets to a target resolution, i.e., the level of detail used to detect cell identities. [38] [2] When the clustering is finished, the groups of cells are annotated with class labels, representing the underlying biology in a language we can understand. [39]

Instead of assigning expression gates from pre-defined markers, as is the standard for flow-cytometry analysis, single-cell RNA-seq analysis pipelines usually start from *de novo* clustering of cells followed by cluster annotation. [38] While it is clear that clusters and cell types are different concepts [38], often cluster labels are treated as cell types. There are a number of ways to cluster cells to find groups of similarity, but arguably the current default is derived on the methodology proposed by PhenoGraph. [40] The protocol is to calculate the distances between cells in a reduced PCA space (with the number of dimensions chosen by the experimenters), followed by constructing a k-nearest-neighbours network, in which each cell is a node connected by *k* (another parameter) edges to other cells. Once the network is build, network modules (i.e. cell clusters) are commonly found using the Louvain algorithm, published in 2008 by researchers of the Université catholique de Louvain, in Belgium. [41] The cell clusters found by the PhenoGraph (or any other) algorithm are then labeled by domain experts, often based on genes differentially expressed on each cluster, so-called “markers”. [38]

While it is possible to manually investigate the identities of which clusters, automatic methods have been developed to aid on the task. [39] One approach (“marker-based automatic annotation”) bases itself on crossing clusters markers in the analyzed dataset with previous knowledge from databases like PanglaoDB [42] and CellMarker [43] [39]. Another approach (reference-based automatic cell annotation) relies on base, expert-annotated

datasets as references from which labels are transferred to the dataset of interest. [39] Other methods bypass the clustering step and focus on labelling the individual cells, which avoids lumping dissimilar cells together, but require a high amount of reads per individual cells for it to be efficient. [39] A recent review and tutorial by Clarke et al [39] provides an extensive account of current techniques.

Of note, even though a range of methods is available, the vast majority of techniques and publications do not use standard identifiers for cell types. This is in contradiction with the acknowledgement by the community of the advantages of using identifiers the ad using standard identifiers, such as those provided by the Cell Ontology. [39] [44] [21] [45] [46] [47]. Nevertheless, projects that use Cell Ontology identifiers for single-cell RNA-seq data are appearing [48], including python and R packages (e.g. Besca [49], OnClass [50] and ontoProc[51]), data management projects and reference datasets, (e.g. Tabula Muris [52/] and Tabula Sapiens [53] Azimuth map [54/] and HubMap's ASCT+B Tables [55]) and annotation platforms (e.g. the Cell Annotation Platform [56] and CellTypist [57].

As elegantly put by Meehan et al [58] the Cell Ontology is a “manually constructed computer readable resource that links cell types by different relationships”. it was first described in 2005 by Jonathan Bard, Seung Y Rhee† and Michael Ashburner [31] and was oriented at creating an “organism-independent classification of cells”, following criteria that included function, histology, lineage and ploidy and providing “Cell-type unique identifiers (ID) that can be incorporated into any database holding cell-type-associated knowledge.” It also had a didactic goal in itself, as the authors mention [31]: “It is designed to be useful in the sense that a researcher should be able to find, in a rapid and intuitive way, any cell type in any of the major model organisms and, having found it, learn a considerable amount about that cell type and its relationships to other biological objects.” The collaborative project gradually evolved and changed its design and scope to fit new needs. By 2011, for example, a need for computable definitions for hematopoietic cell types lead to a sizeable advance in the number and quality of immune cell types represented in CL. [59] It also included the addition of species-specific cell types to better handle marker-based definitions, which are usually given at the species level. [59] Further developments over the years included both technical improvements as well of the addition of new cell types, and by the time of the last official CL publication, in 2016, it contained approximately 2,200 classes. [45]

The Cell Ontology, currently, is growing as a resource for the Human Cell Atlas and in providing identifiers for cell types [48].

In conclusion, the advancement of our *formal* classification of cell types, such as in the Cell Ontology, represents a tangible goal of current cell-oriented large scale projects. While purely theoretical developments have their value, refining the cell type theory in the context of knowledge management arguably will have a influence directly on how the products of the Human Cell Atlas will impact modern science. One reason is that formal systems enable automation of knowledge integration, and can feed intelligent systems that aid current research practices. In the following chapter, it will be discussed how computer-based knowledge processing can influence life-sciences research, as well as discuss techniques and platforms to advance the frontier.

## Ontologies

---

The classification of biological concepts is at the core of biology. At least since the Aristotelian endeavours to group classes of animals, a good part of the scientific work is to capture concepts into knowledge systems [60]. Linnaeus' binomial system for naming species and Mendeleiev's periodic table are likely the two most famous classification systems, but are part of a much larger ecosystem of structuring scientific knowledge.

On the 20th century, the development of the analytical philosophy of Russel and Wittgenstein and their search for formalizations [61] gradually layed the foundations for the the logic of scientific descriptions. Karl Popper and his “The Logic of Scientific Discovery”[62] was heavily influenced by analytical philosophy, and the field is at the foundation of the “falseability” system of Popper. Less known among life scientists, Tarski's inquiries on what can be considered to be “true” [63] were also

The whole movement for formalization of knowledge progressed on the computational end, and at the late 20th century were at the root of the functioning of the World Wide Web, the advent of computational ontologies and large scale knowledge graphs. In this chapter, I will provide an overview of ontologies and knowledge graphs and their use in today's biomedical sciences, alongside its future prospects.

## The OBO Foundry and biomedical ontologies

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. [60]

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. [64] [65].

The Gene Ontology is part of a much larger effort to formalize concepts across biology: the Open Biomedical and Biological Ontologies (OBO) Foundry. [66] 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 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.

## OWL and ontology languages

One of the OBO Principles for its ontologies is that they should be resolvable as a “syntactically valid OWL file using the RDF/XML syntax.” (http://www.obofoundry.org/principles/fp-002-format.html). The OWL Web Ontology Language was introduced as a standard by the W3C consortium in 2004. OWL is not a programming language, as it does not instruct computers to perform actions, but an ontology language, which allows computerizable descriptions of the world. Furthermore, it is an umbrella ontology language that includes several languages with varying levels of expressivity. Generally, more expressive languages can represent more complex ideas, but make computations harder.

Regardless of the sublanguage used by ontology it must be resolvable to an RDF/XML file. RDF stands for Resource Description Framework, another W3C standard built around a graph-based data model (https://www.w3.org/TR/rdf11-concepts/). Statements in RDF are triples consisting of 2 nodes (a subject and an object) and an edge (a predicate) connecting the nodes. All nodes and edges are represented in RDFs by International Resource Identifiers (IRIs), and there are many ways to lay out those IRIs on a text file to represent triples. One of those layouts is the RDF/XML syntax, inspired by the XML markup language. Arguably, other syntaxes (interchangeable with RDF/XML) are easier to read for human. As an example of an RDF triple, here is how one would represent in the Turtle RDF Syntax, the notion that plasmacytoid dendritic cells are a type of dendritic cells:

```
http://purl.obolibrary.org/obo/CL_0000784  http://www.w3.org/2000/01/rdf-schema#subClassOf
http://purl.obolibrary.org/obo/CL_0000451 .
```

Where [http://purl.obolibrary.org/obo/CL\\_0000784](http://purl.obolibrary.org/obo/CL_0000784) and [http://purl.obolibrary.org/obo/CL\\_0000451](http://purl.obolibrary.org/obo/CL_0000451) are the unique IDs in the Cell Ontology for “plasmacytoid dendritic cells” and dendritic cells, respectively, and <http://www.w3.org/2000/01/rdf-schema#subClassOf> is the identifier for the “subclass of” relation as defined by the RDF schema.

A longer explanation of the details of OWL and RDF is outside the scope of this work. This brief introduction has a dual goal of introducing the architecture of formal representations and of demonstrating the complexity of the system. There is a high energy barrier to acquire the knowledge and the technical skills to engage in ontology building. That complexity might be one of the reasons why a very small fraction of the biomedical communities represents data with ontologies and an even smaller fraction engages with ontology building.

## Wikidata

---

Even though the Semantic Web (which ontologies are a part of) spawned with promises of a revolution in the way knowledge is shared, it is still to be widely known outside the semantic engineering. Two recent projects are playing a particularly important role in bringing the Semantic Web to a wider audience has been receiving a boost of attention recently powered by two large projects: the Google Knowledge Graph and Wikidata.

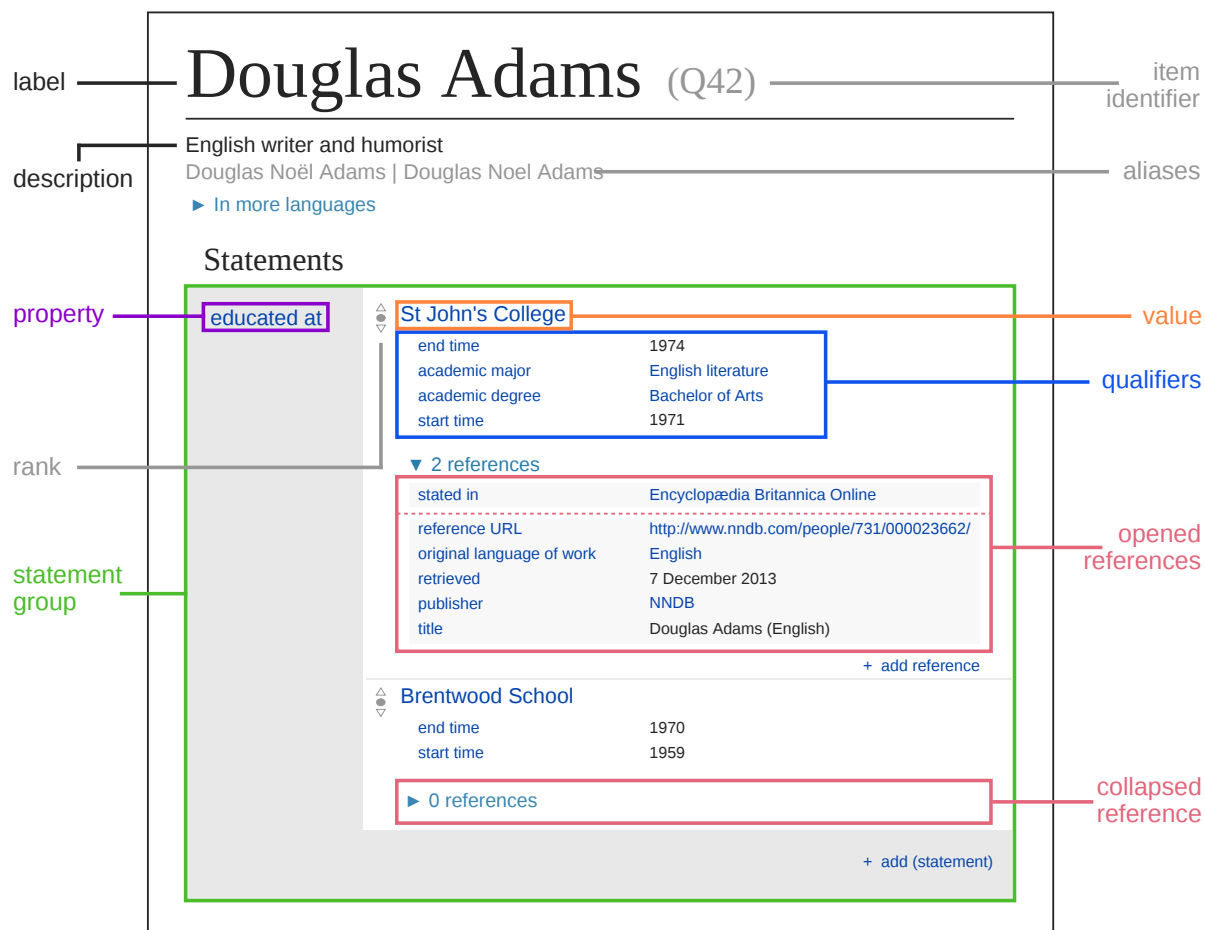
The Google Knowledge Graph introduced the Semantic Web *de facto* in the daily life of users of Google. [67]. Its underlying structure is similar to the triples in an ontology, but it is less concerned with being logically coherent, and does have strict semantics of a representation. In that way, Google Knowledge Graphs can feed on a variety of sources and not crash if there is some data modelling that, rigorously, could be inconsistent. Even though there is not a rigorous boundary between ontologies and knowledge graphs, one reasonable interpretation is that a knowledge graph may not be perfectly coherent, as long as it still can provide enough knowledge and reasoning for the approach of interest. While the lack of formal semantics limits reasoning and inference, the knowledge graphs are arguably easier to use, edit and understand, and so provide an user friendly alternative for computable information with a lower entry barrier.

While the Google Knowledge Graph is widely used as a source of knowledge, it does not allow independent users to contribute with information. On the other hand, 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. [68] Its power is derived of its large community of contributors, closely linked to the hugely successful Wikipedia. With a community of more than 20,000 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 successful that Google decided to migrate its own knowledge base, Freebase, fully into Wikidata.[69]

## The inner workings of Wikidata

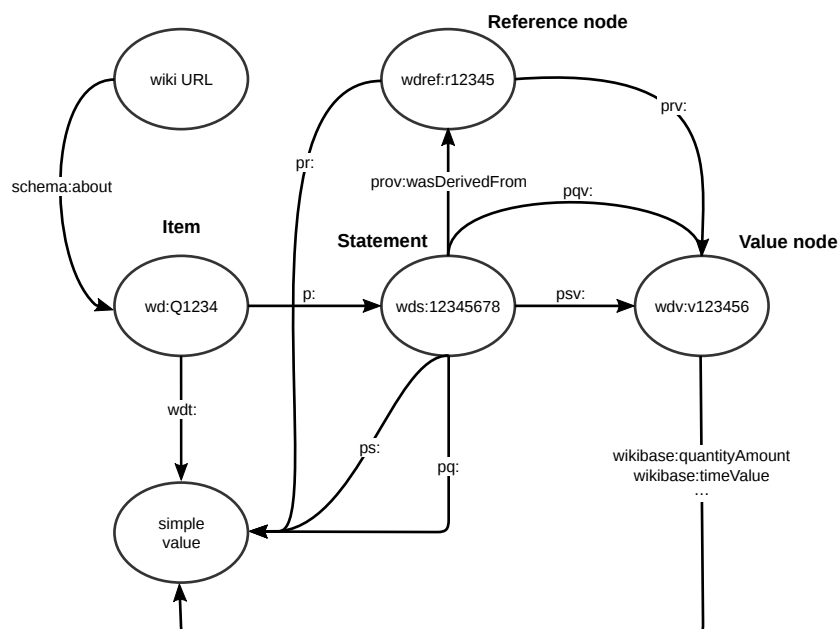
Wikidata uses the same framework (RDF) that powers ontologies, and its model represents statements about the world in triples containing a subjects, a property and an object. [70] Its data model is serialized both in JSON and RDF. The data model contains 17 different datatypes, including, for example “Item”, any entry on Wikidata that refers to “o a real-world object, concept, or event that is given an identifier in Wikidata” and “String”, a “sequence of freely chosen characters interpreted as text”. [71]. Knowledge is stored on Wikidata upon basic triples containing a subject (of type “Item”), a property and a value (which can be of any of the 17 types). As of November 2021, Wikidata contains more than 90 million data items [72] and more than 9000 properties that link them to values. As values often are other items, the database acquires a network format with labeled edges.

As can be seen in the example in 1, each the items in the database contain an item identifiers (Q followed by numbers). They also contain a label, a description and a list of aliases, which can be recorded in any of the more than 200 hundred languages, thereby making it a multilingual project. [73] Each item is decorated with statements, comprising of property-value pairs. These pairs can be further specified via qualifiers and references, which treats the full triple as the subject, adding metadata to it (a process called reification [74/#reification]). Qualifiers provide ways to extend the information on the triple, while references provide provenance, enabling users to judge the validity of the claims in the database.



**Figure 1:** Wikidata's model for describing an item. Image released in Public Domain by Charlie Kritschmar.

All the information is available on a user interface, but its data is also available programatically in diverse formats, including as full JSON and RDF dumps, the MediaWiki API and a SPARQL endpoint. [75] A number of wrappers of such services are available in languages such as R [76] and python [77/]. A scheme of the data can be seen in 2, where each item is connected to a statement node via a property in the "p:" namespace, from which references and qualifiers are accessible. To facilitate basic usage, the namespace "wdt:" connects items to values directly, simplifying, for example, the writing of SPARQL queries.



**Figure 2:** Wikidata's data model, scheme released under the CC-BY 4.0 license by Michael F. Schönitzer. It outlines the basic representation of statements, qualifiers and values in the Wikidata database

Information on Wikidata is released under a CC0 license, which enables full reuse of the data. [78] One of the major points of access and reuse of the information is the Wikidata Query Service [79/], a core resource of the community which enables live querying in the SPARQL language. [80] A number of services make use of embedded queries from the Wikidata Query Service [79/] to create interactive, live dashboards, for example Scholia [81/] and the SARS-CoV-2 Query Book [82/]



Wikidata is not only accessible in different ways, but also writable in many ways. It provides a user-friendly, point-and-click interface for modifying the database, providing a low entry barrier for newcomers. It is also possible to semi-automatically reconcile spreadsheets to Wikidata items and use batch tools such as Open Refine [83] and Quickstatements [84], which enable batches on the magnitude of thousands of edits. For larger amounts of edits, it is possible to ask for bot permissions [85] and deploy systems that integrate big data sources. Bot edits are made via the Wikimedia API and are predominantly written via Python wrappers, such as Pywikibot [86] and the Wikidata Integrator. [87]

## Wikidata as a knowledge graph for the life sciences

Due to its privileged position inside the linked data ecosystem and its ease of write and query, Wikidata has been growing as a hub for interoperable data for the life sciences community. [88] [89] Even though Wikidata was created in 2013, the demand for a community-cured life sciences knowledge graph is clear at least since 2008 [90] [91]. The Wikidata-like project proposed at the time was eventually discontinued, an example of the challenge of maintaining independent biomedical databases. [92] As Wikidata has a very large community, has stable funding and is at the core of modern technologies, like the Google Knowledge Graph [69] and Amazon's Alexa, [93/] it is virtually guaranteed that data in Wikidata will remain accessible for a long time, regardless of local funding schemes.

The Gene Wiki project [94] was likely the first large scale biomedical project to rely directly on the Wikipedia infrastructure for community curation. It provided a direct connection between the generalist community of Wikipedia and domain experts. The interplay of both communities is a topic of discussion and the opportunities and challenges were already discussed in NAR in 2012. [95]

Notably, Wikidata appeared chronologically after those efforts.

Notwithstanding, the Gene Wiki research group has embraced the Wikidata environment for community biocuration and data interoperability [96][97] [88] [98]. The information on Wikidata is still integrated to Wikipedias across multiple languages, often as source of information in Wikipedia's infoboxes.

Other projects outside the Gene Wiki initiative also started using Wikidata as a platform for knowledge integration. A list of several projects that use Wikidata as part of their service to their community is given on table 1. There is movement exploring how Wikidata can be employed to the advance of Computational Biology, and how it can be integrated to current publication status quo. [99] In that direction, Wikidata is being developed as a platform for scholarly linked open data, particularly via the Scholia platform [100] [101], (<https://scholia.toolforge.org/>) which provides profiles of pre-templated SPARQL queries for entities like particular authors and articles (e.g. Scholia profile on Prof. Helder Nakaya: <https://scholia.toolforge.org/author/Q42614737>).

## Table 1

During the COVID-19 pandemic, Wikidata has spawned as a hotspot for modelling information about the virus and the pandemic in real time. [102] [[wikidata:99196713?](#)] The general scope of the database allowed representation in a shared system of molecular, epidemiologic and socio-economic aspects of the pandemic. [102][103] Information curated in Wikidata was immediately available, feeding live dashboards and other applications based on SPARQL queries. [104] [105] [106] Additionally, as the information presented on Wikidata is multilingual and collaboratively edited, it presented itself as a resource for constructing structured vocabularies in non-english languages. [107]

In addition to its value as a structured database, Wikidata is tightly connected to Wikipedia. The gene identifiers in the context of Gene Wiki [96] are now fed to Wikipedias across languages, benefitting users directly. Additionally, gene expression information from the Bgee database [108] was added to Wikidata and connected to Wikipedia, which lead to a sizeable increase of the Bgee database. Currently, Wikipedia is one of the top 3 sources from which people access Bgee (personal communication with Tarcisio Farias, <https://scholar.google.fr/citations?hl=fr&user=sB87J-cAAAAJ>), thus leading to direct recognition for integrated bases. More generally, the connections of Wikidata and Wikipedia make it unique in the power of flowing knowledge back to human-accessed interfaces. In the words of Matthias Samwald [109] and colleagues "Wikidata could emerge as a community-backed and highly visible structured knowledge base of medical and biological information, bringing concepts and methodologies such as controlled taxonomies, Semantic Web / semantic technologies and ontologies into mainstream use."

In conclusion, Wikidata's unique position, robustness and guarantee of long term stability, prompts the need of works exploring new ways of integrating it to current knowledge management. Given the speed and breadth of the Human Cell Atlas, and the challenges of knowledge representation on cell types, this PhD work plans on discovering and addressing knowledge gaps on how Wikidata can play a role in organizing and disseminating the discoveries about all human cell types.

## Objectives

- Study and refine theories of classes of cells within the constraints of ontologies and knowledge bases
- Provide a comprehensive list of currently described cell types on Wikidata
  - Develop a biocuration framework for the task of sharing information on Wikidata
  - Catalog as many cell types as possible, as a groundwork for future applications
- Devise ways to connect the Human Cell Atlas and other life-sciences products to Wikidata:
  - Craft Wikidata relations ("properties") for making cell-type-related assertions
  - Write bots and scripts to reconcile data sources to Wikidata
- Provide proofs-of-concepts of how Wikidata integration can benefit the advancement of HCA

## Methodology

This project's methodology resembles practical research-action practices [110]. Its goals of improving interoperability of cell-type data implies a combination of action and research. Action in the form of active contributions to ontologies and knowledge-graphs, by getting involved and contributing to ongoing projects in the context of the Human Cell Atlas and knowledge management. Research in the 3 forms: - Philosophical investigation on the nature of knowledge representations of cell types, both in formal logic settings and in current academic practice - Applied investigations of database integration and data quality in the context of Wikidata and biomedical ontologies - Data-driven biomedical research targeted at hypothesis generation and literature-based discovery using knowledge at the level of cell-type

All the research forms are intertwined with the improvement of knowledge management in biomedical sciences, with a focus on the Human Cell Atlas. The methods included the development and application of a framework for organized reading of the scientific literature, aimed at providing contact with the different facets of biocuration and Human Cell Atlas-related research.

## Organized reading

---

To handle the literature reading necessary for this project, a framework was developed for reading and is described in details in the results section. The framework is based on GitHub and includes Python scripts, a file organizing the reading list, and another documenting the reading history in RDF. Notes and additional information are saved in a GitHub repository, and the structured information powers a live website with analytics on the users recent readings. The source code for Wikidata Bib is available at [https://github.com/lubianat/wikidata\\_bib/tree/template](https://github.com/lubianat/wikidata_bib/tree/template) and notes on my readings can currently be accessed at [https://lubianat.github.io/wikidata\\_bib/](https://lubianat.github.io/wikidata_bib/).

Additionally, the methodology included a discipline of reading that entails the daily reading of 2 articles, one about “cell types” and another about “biocuration”. The articles are obtained by a mixed manual and automatic approach, including a la carte selection of articles to read alongside Wikidata queries for Cell, Nature, Science and eLife papers about single cell transcriptomics (query: <https://w.wiki/4LHr>) and for papers on biocuration (query: <https://w.wiki/4LHi>). )

## Biocuration of cell classes for Wikidata

For each article about cell types read, cell types previously absent on Wikidata are added via a combination of curation in a Google Spreadsheet and a custom Python script ([https://github.com/lubianat/wikidata\\_markers/tree/master/curation\\_of\\_classes](https://github.com/lubianat/wikidata_markers/tree/master/curation_of_classes)).

## Annotation of Human Cell Atlas articles

Human Cell Atlas publications (<https://www.humancellatlas.org/publications>) were selected and abstracts were annotated as richfully as possible with Wikidata IDs using the hypothes.is annotation system (<https://web.hypothes.is/>). One article [2], describing the complete Human Cell Atlas project, was annotated in full. Annotations were retrieved via the hypothes.is API and processed with custom Python and R scripts ([https://github.com/lubianat/ann/tree/main/hypothesis\\_parsing](https://github.com/lubianat/ann/tree/main/hypothesis_parsing)).

## Wikidata updates

---

Wikidata is similar to a graph database, and is flexible enough to add new relations without need to change the underlying infrastructure.

Creation of new entities was done either manually in the Graphical User Interface (<https://www.wikidata.org/wiki/Special:NewItem>) or via custom python scripts combined with the Quickstatements tool (<https://quickstatements.toolforge.org/#/>) or the Wikidata Integrator python library (<https://github.com/SuLab/WikidataIntegrator>).

Properties, which link items to values, cannot be created at will and need to undergo community approval. Under the scope of this PhD project, we have gotten the community approval for a number of properties:

- entry receptor (<https://www.wikidata.org/wiki/Property:P8339>) used to link pathogens to their cellular entry receptors.
- Cell Ontology ID (<https://www.wikidata.org/wiki/Property:P7963>) used to link cell types to their IDs in the Cell Ontology
- has marker (<https://www.wikidata.org/wiki/Property:P8872>) used to link cell types to genes and proteins considered their markers
- derived from organism type (<https://www.wikidata.org/wiki/Property:P9072>) used to link cell lines to the taxon of the organism from which it was derived.

The property acceptance cycle takes at least one week and is completely open for opinions by any Wikidata user. All the information regarding the property proposal is available at [https://www.wikidata.org/wiki/Wikidata:Property\\_proposal](https://www.wikidata.org/wiki/Wikidata:Property_proposal).

## Cell Ontology participation

---

As part of the research-action process, I have joined the Cell Ontology working group. I participate in the monthly meetings and sporadic workshops, learning and contributing to the discussions. Additionally, I contribute to the ontology development, actively engaging in the Cell Ontology GitHub repository (<https://github.com/obophenotype/cell-ontology>) and contributing with new terms and assertions. I edit the ontology with the software for ontology editing Protégé v. 5.5.0 (<https://protege.stanford.edu/>).

## Status of cell type info on Wikidata and the Cell Ontology

Status of cell type information on Wikidata was accessed via SPARQL queries combined with processing in python and is available at [https://colab.research.google.com/drive/1GvQXOs51\\_U8icdGMtKXMeLOXKM8pXWet#scrollTo=szvBWl9zr\\_AA](https://colab.research.google.com/drive/1GvQXOs51_U8icdGMtKXMeLOXKM8pXWet#scrollTo=szvBWl9zr_AA).

Counts of cell classes in the Cell Ontology were performed via regex matching on Cell Ontology releases following the code available at [https://github.com/lubianat/cell\\_ontology\\_count](https://github.com/lubianat/cell_ontology_count).

# Preliminary Results

# Concept of cell types

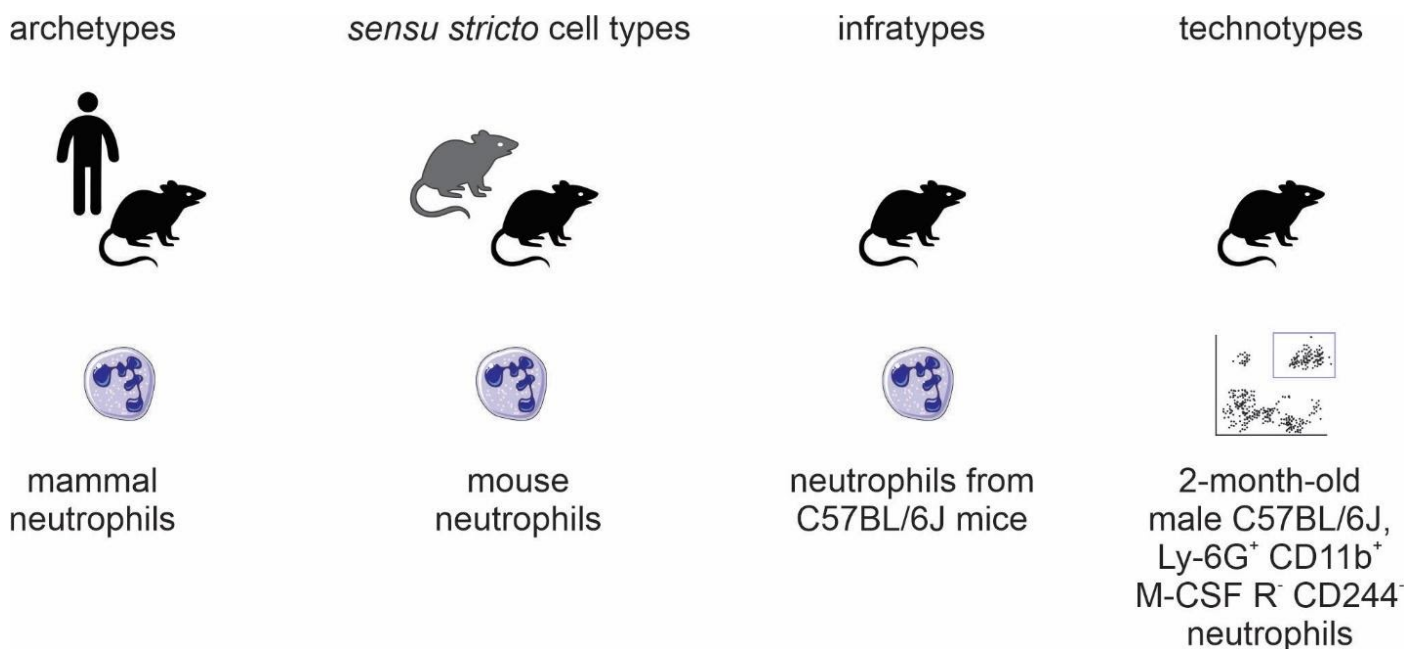
## General work on the concept of cell type

As an initial step of this PhD project, we decided to investigate the definition of “cell type” and how to shape a definition for knowledge management on Wikidata. The definition of the concept of “cell type” is currently a topic of debate by the biomedical community.[1,7,8,9,10,11,12,13,14,15,16,17]. 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 [31,32,33] and CELDA [34] and the contributions of the International Workshop on Cells in Experimental Life Sciences series [35,36].

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 an ontology or a knowledge graph, like Wikidata. We avoided 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 [111] 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[112], we proposed, among other theoretical novelties, 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 [113], we derive a species-centric view on the naming of classes of cell types. The four classes (Figure 3) 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<sup>+</sup> CD11b<sup>+</sup> M-CSF R<sup>-</sup> CD244<sup>-</sup> neutrophils”.



**Figure 3:** 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 a future session.

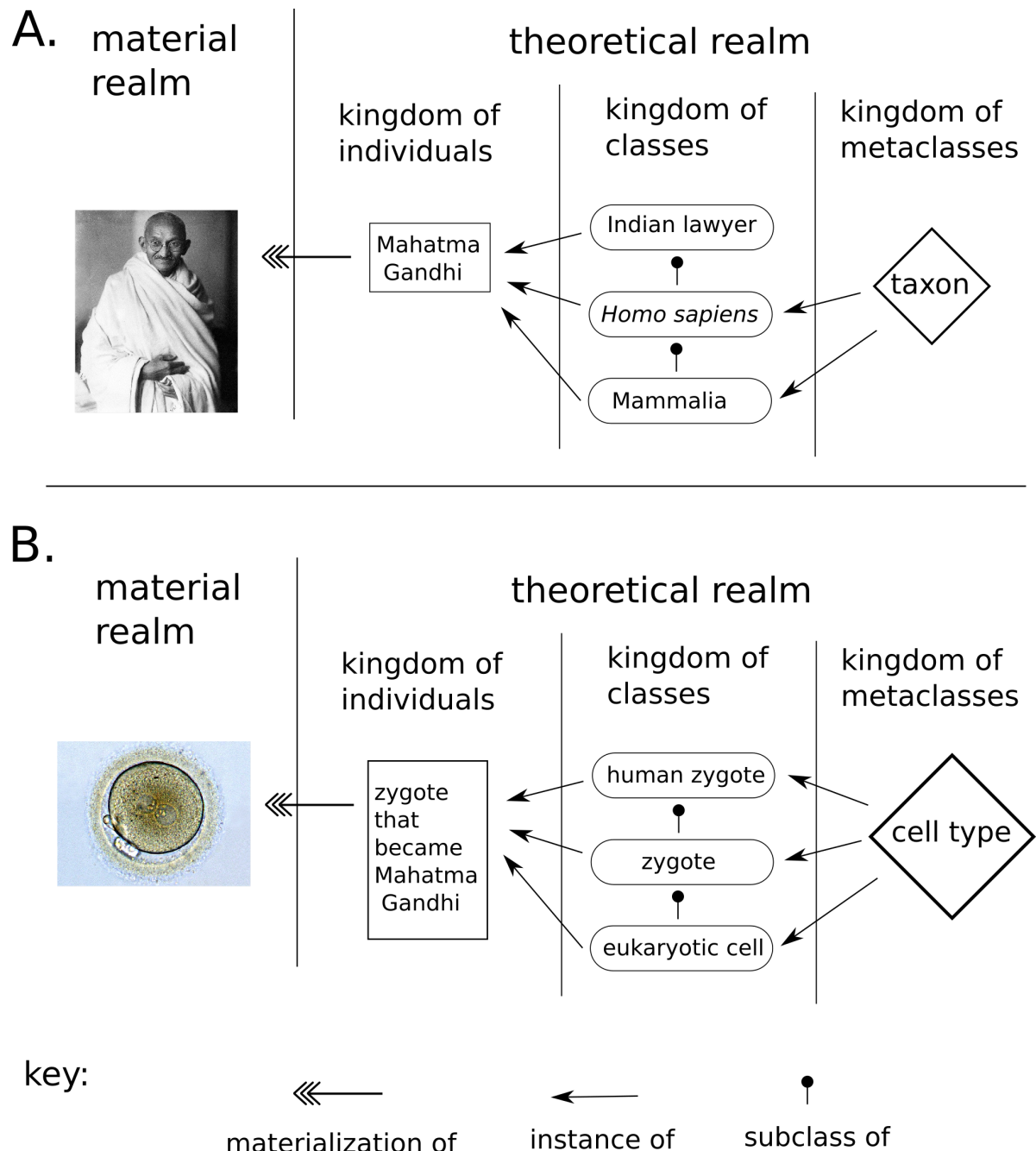
## A simplified definition

Refining the different concepts around the notion of “cell type” is important, but will require decades before a reasonable consensus. Here we adopt a liberal view of cell type, defining, for our purposes, a cell type as any class of cells described by a domain expert with evidence of reality of its instances. The requirement of evidence of existence in reality is based on the Principle of instantiation of ontological realism [114]. Barry Smith and Werner Ceuster



states that “A term should be included in a reference ontology only if there is experimental evidence that instances to which that term refers exist in reality.”(‘Exists’ here should be understood in a tenseless sense in order to accommodate, for example, universals pertaining to extinct species as well as universals such as swarm or hurricane which are instantiated only intermittently.)” Thus, in this work one minimum requirement for a cell type to be catalogued is a public description by a researcher of the class, with evidence for existence of instances of the class in reality.

By “class” we mean an abstract entity in the sense intended by the multilevel theory (MLT) of conceptual modelling (referred as “*types*” by Carvalho et al. ) [115] Figure 4 displays a simplified version of MLT adopted throughout this project. In this framework, real-world entities are materializations of *individuals*. *Individuals* are theoretical constructs which are (1) thought to exist or have existed, as per the principle of instantiation, and (2) refer to only one (01) material entity at any point in time. For example, Wikidata has entries for people, e.g. “Helder Nakaya (Q42614737)” and “Charles Darwin (Q1035)” which are considered *individuals* by Multi Level Theory. Other examples of *individuals* include “Albert Einstein’s brain (Q2464312)” and the “Christ the Redeemer statue (Q79961)”.



**Figure 4:** Multileveltheory for cell types

Figure 4 A multilevel theory (MLT) can divide the theoretical realm into different kingdoms. A) A representation of people in the MLT framework adopted in this work. The theoretical-realm entity “Mahatma Gandhi” is materialized by the material-realm Mahatma Gandhi. The theoretical *individual* is considered an instance of multiple *classes* such as “Indian lawyer” and “*Homo sapiens*”, which are related to each other via subclass relations. The classes themselves are instances of *metaclasses*, like “taxon”, a first order metaclass. B) An analogous representation of the MLT framework, but applied to cells and cell types.

In MLT, *individuals* are instances of some *classes*. For example, both “Helder Nakaya (Q42614737)” and “Charles Darwin (Q1035)” could be represented as instances of the class “*Homo sapiens* (Q15978631)” on Wikidata. “*Homo sapiens* (Q15978631)” is only one of the classes that those individuals belong to. Another one is “animal (Q729)”. As all instances of “*Homo sapiens* (Q15978631)” are also instances of “animal (Q729)”, “*Homo sapiens* (Q15978631)” is a subclass of “animal (Q729)”. It is possible to continue the hierarchy of subclasses, as “animal (Q729)” is a subclass of “organism (Q7239)”, until the root case, which in the case of wikidata is the class “entity (Q35120)”

Classes, however, can themselves behave as individuals. For example, both “*Homo sapiens* (Q15978631)” and “animal (Q729)” are instances of “taxon (Q16521)”. “Taxon (Q16521)”, thus, is a *metaclass*, or, more precisely, a *1st-order metaclass*. Other examples of metaclasses are “*species* (Q7432)” and

“phylum ([Q38348](#))”. These, in turn, are instances of “taxonomic rank ([Q427626](#))”, a *2nd-order metaclass*.

In the Figure 4 B there is a proposal of this version of MLT for cell types. As individual cells are rarely named, for the sake of example, we can consider the “zygote of Mahatman Gandhi” as an *individual* in the theoretical system, an instance of the class “zygote ([Q170145](#))”, which is itself an instance of the metaclass “cell type ([Q189118](#))”. A more concrete example stems from RNA-sequencing datasets with barcodes for each single cell in a particular sample. Each barcode can be thought as an identifier for an *individual*. Thus, labeling single-cells is a process of identification, where each *individual* is connected to a *class* of interest.

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 [\[111\]](#) for an example). We also consider only the cell as it was observed in an experiment, not necessarily the future conditions of any cell (i.e., the “cell fate”). [\[27\]](#) Even though such a distinction is an important topic for theoretical research, it is outside the initial scope of this work.

Another logical consequence of the definition is that the concept of subtype becomes redundant with the concept of cell type. The notion of subtype, then, only makes sense when discussing classes with different degrees of universality. Thus, claims to discover new cell “subtypes” or “types” differ only stylistically and can be considered indistinguishable from the perspective of research synthesis.

We also note that we made a judgment call to use the term “cell type” to emphasize the focus on types as classes (or “kinds”) in contrast to real-world objects. The term “cell class” is also used in the literature and is a suitable synonym for our notion of cell type. We opted to frame our work around the term “cell type” due to its historical usage and familiarity for the life sciences community. Other related terms a “cell set,” “cell population,” and “cell cluster,” can also reminisce of a specific, countable group of cells, frequently from the same experiment. The term “cell identity” has also been suggested for avoiding the cell type/cell state dilemma [\[38\]](#), but we avoid it to emphasize a nominalistic perspective (in the Popperian sense [\[116\]](#)). In doing so, we reinforce the intent on represent on what cell types are *reported to exist*, instead of trying to state bluntly which cell types *exist* or, even worse, are *essential* of human beings.

The employment of MLT as described before, and the notion of species-specific cell types are fundamental for the next chapters of this work. In the chapter about the PanglaoDB integration, we describe how we used the notion of species-specific types to add marker information to Wikidata, and how we cleaned up conceptual disarrays that broke MLT.

# PanglaoDB integration to Wikidata

## Introduction

The process of making the Human Cell Atlas more useful via Wikidata also includes the connection of related databases. PanglaoDB [117] [118] 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 8 December 2021, the article describing PanglaoDB had been cited 230 times. Despite its use by the community, the database is on a 3-star category for Linked Open Data [119] 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 PanglaoDB 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", considerably less than in specialized cell catalogs, which count over two thousand cell types [45,120]. 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, an example of conceptual disarray that often occurs on Wikidata. [121]

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 8 December 2021, the Wikidata database contains 2834 instances of "cell type" (see current status at <https://w.wiki/b2t>) and 0 instances of "cell" (<https://w.wiki/4XAg>) highlighting the improvements in both quantity and quality. This increase stems both from the PanglaoDB initiative (around 430 new types) and from the Wikidata Bib curation, described in a later chapter.

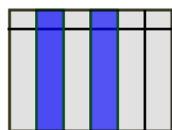
## Methodology for PanglaoDB integration

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 ??.

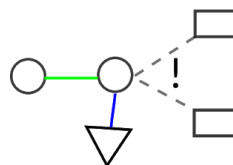
get permission  
from developers



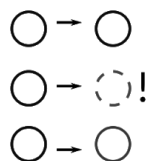
select fields to keep



choose/create  
Wikidata properties



match/create  
Wikidata classes



add to Wikidata  
with citations



make integrative  
SPARQL queries



## 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 [34].

## 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](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 [87], 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 5 shows 2 of the current markers of “human cholinergic neuron”(Q101405051), CHAT and ACHE, as they are 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.

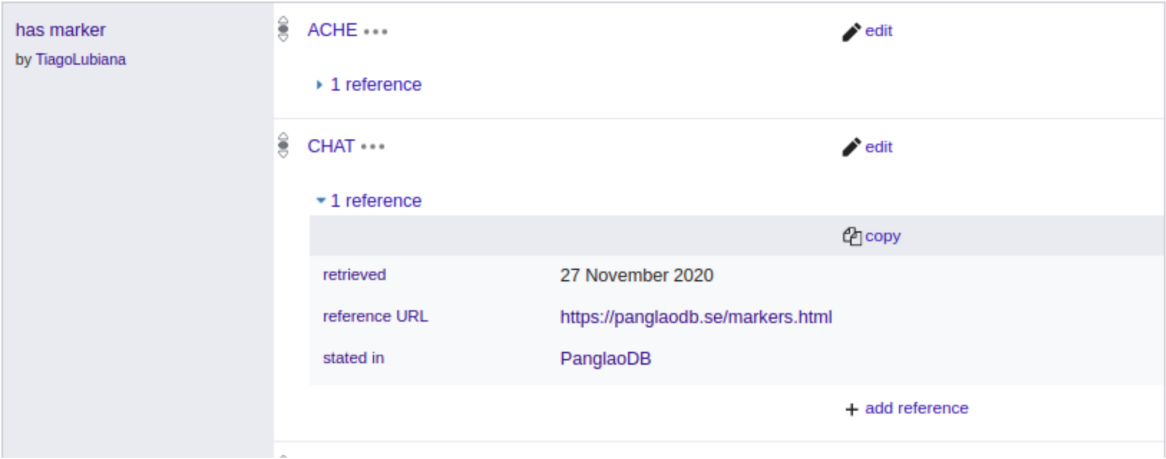


Figure 5: Subset of the marker genes for item Q101405051 (human cholinergic neuron)

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 [122] and Wikipathways [123]. 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 6).

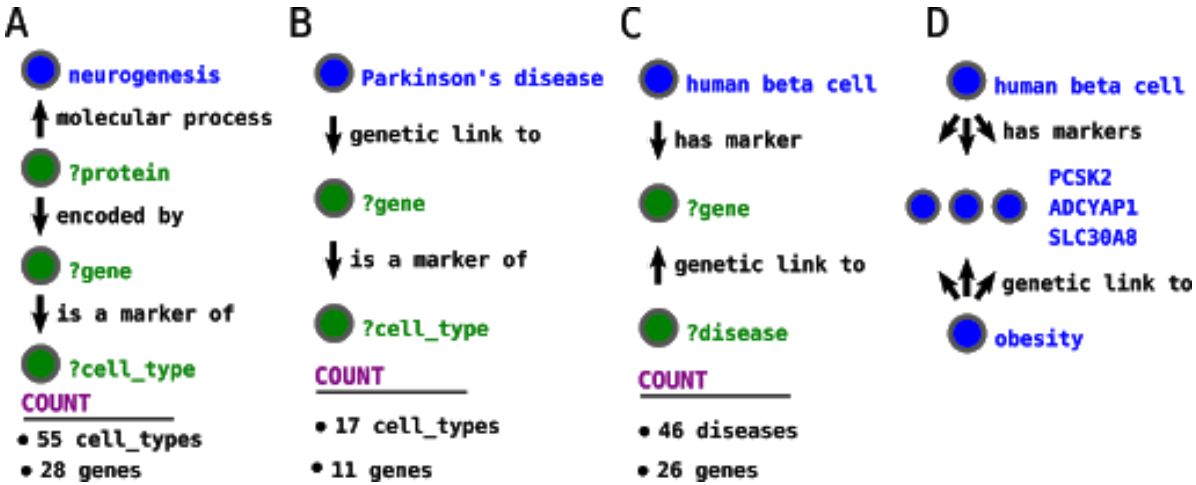


Figure 6: SPARQL queries in Wikidata now harness information from Panglao DB. Queries with the above design were run on Wikidata. Results might change in real time with Wikidata updates by contributors A-C) Graphical representation of feasible SPARQL queries (<https://w.wiki/yQ6>, <https://w.wiki/yQD> 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: Sample of 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

geneLabel	cellTypeLabel
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:** Sample of 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 [103] 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 [43], labome [124], CellFinder [120] and SHOGoin/CELLPEDIA [125/]) (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.

## Wikidata Bib and a professional system for biocuration

### Introduction

Reading scientific articles is an integral part of the routine of modern scientists. Although a number of literature/reference management software are available [[wikidata:https://en.wikipedia.org/wiki/Comparison\\_of\\_reference\\_management\\_software?](https://en.wikipedia.org/wiki/Comparison_of_reference_management_software?)], the process of reading is largely artisanal. There are no standard guidelines on how to probe the literature organize notes for biomedical researchers. Thus, while reading and studying is a core activity, there are few (if any) protocols for efficient screening of scientific articles.

Other professional traditions have dealt with similar issues in the past. In the field of accounting, note-taking is of outstanding importance, to keep track of financial balances and avoid costly problems. Double-entry bookkeeping was developed in the 13th century as a professional solution for note-taking in accounting where “every entry to an account requires a corresponding and opposite entry to a different account.” [20, =Double-entry\_bookkeeping&oldid=1055066428] In software development, Test-Driven Development (TDD) is a popular methodology where tests for code snippets are written before the code itself, therefore ensuring that written software passes minimum quality standards. The similarities of Double-entry bookkeeping and TDD are diverse [[wikidata:https://blog.cleancoder.com/uncle-bob/2017/12/18/Excuses.html?](https://blog.cleancoder.com/uncle-bob/2017/12/18/Excuses.html?)], but for our purpose here suffices to see both as professionalized systems that promote better quality and accountability of works.

In the humanities, there is a well-established practice of annotations of readings. The annotation skills are part of common academic training in the humanities [126/][127\_da26C-QW5qiS7uZ]. An influential work in presenting methods for academic reading in the humanities is Umberto Eco’s book “How to Write a Thesis” [128], which outlines not only *how* to annotate the literature that basis an academic thesis, but also *why* to do so. The book, written originally in 1977, is still influential today, but its theoretical scope (roughly the humanities) and its date, preceding the digital era, limits the extent in which it applies to the biomedical sciences.

Notably, the need of an organized reading system for biocuration studies stems from a difference in methodology. In humanities, the main (if not sole) research material is the written text, the books and articles from which research stems. [127\_da26C-QW5qiS7uZ]. In the biomedical sciences, including a large part of bioinformatics, the object of study is the natural world, observed via experimentation. Thus, naturally, scientific training focuses on the theoretical and practical basis of experimentation and data analysis. With the bloom of scientific articles, however, the scientific literature (and accompanying public datasets) provide already a strong material for the sculpting of scientific projects. Thus, the development of a methodology for academic reading, tailored to the digital environment, presents itself as a need.



This chapter concerns itself with presenting Wikidata Bib, a framework for large scale reading of scientific articles. It is presented as three parts, each of them with a technical overview alongside the theoretical foundations. First, Wikidata Bib is presenting as a reading system, for managing references and notes using a GitHub repository and plain text notes. Then, we present how the system ensures accountability, allowing its user to get personalized analytics on their reading patterns. Finally, we demonstrate how Wikidata Bib fits an active curation environment, connecting the framework with the larger goal of this project of curating information about cell types on Wikidata.

## Wikidata Bib as a reading system

The reading framework of Wikidata bib is built upon a git repository integrated with GitHub, Python3 scripts and SPARQL queries. It has a standard file structure, summarized as the following:

- docs/
  - index.html
- downloads/
  - 10.7554\_ELIFE.52614.pdf
- notes/
  - Q87830400.md
- src/
  - get\_pdf.py
  - helper.py
  - read\_paper.py
  - update\_dashboard.py
- index.md
- toread.md
- config.yaml
- pop
- wadd
- wadd\_all
- wread
- wlog

The docs/ directory contains the live dashboard from the readings, which will be discussed in the following sessions. The downloads/ directory hosts the pdfs of the articles read with the system. These are not committed to the repository, and are only stored locally. The notes/ directory contains markdown files, one for each article read. The src/ directory contains the python code with the mechanics of the system. They contain helper functions for the command line commands discussed below: - wread which receives a Wikidata QID for an article and outputs (1) a notes document, (2) a pdf for the paper obtained from Unpaywall [129/] and (3) an updated version of the dashboard html files in the docs/ directory. - pop, which “pops” an article from toread.md and runs wread for it - wadd, which takes an URL for an Wikidata SPARQL query and adds new QIDs to toread.md - wadd\_all, which parses config.yaml for recurrent SPARQL queries and runs wadd for each - wlog, which adds, commits and pushes recent readings and dashboard updates to GitHub

All the structures described so far are commonly shared by any user of Wikidata Bib. To personalize the use of the system, the user edits three plain text files. toread.md hosts a plain text QIDs of the articles that will be read. These can be added either manually, or via wadd. While the pop command only sees QIDs, articles titles or other identifiers can be added to toread.md temporarily without breaking the system. index.md hosts a numbered list of topics of interest. This file plays the role of Umberto Eco's work plan, with the topics of interest for the academic. [128] These are used to tag articles for retrieval in a later step. config.yaml contains shortcuts for different reading lists. This is better explained by example. In my toread.md file there are two reading lists, one following a # Cell types header, and another following a # Biocuration header. My config.yaml contains the following snippet:

```
lists:
# - shortcut: Title of header in toread.md
ct: Cell types
bioc: Biocuration
```

The shortcuts in config.yaml are used as arguments by the pop command, where \$ ./pop ct retrieves an article from the “Cell types” list, while \$ ./pop bioc retrieves an article from the “Biocuration” list.

The Wikidata bib framework is coupled with a discipline of daily reading. This is inspired by Robert Cecil Martin's description of Test Driven Development in the book “Clean Code”, which includes not only a technical description, but a *school of thought* of how software development can be approached. [130] Every day, I read one article of each list, using the notetaking station displayed in Figure 7. The constancy of reading allows steady coverage of the relevant literature. While it has worked for this research project, however, it is not required for use of the Wikidata Bib system.

The notetaking station of Wikidata Bib is, by default, opened in Virtual Studio Code, and is depicted on Figure 7 A. The title and publication dates are displayed, and the reading process entails copying snippets from the text to the “Highlights” session. By copying the highlights into plain text, the sections of interest become searchable via command line using grep (https://en.wikipedia.org/w/index.php?title=Grep&oldid=1039541979). Comments can be added either in the comment section or inline, alongside the highlights, using --> Comment goes here to differentiate from highlights. Also searchable by grep are the tags, copied and pasted from index.md in the ## Tags session or alongside the main article.

The discipline also includes, whenever possible, an improvement of the metadata about the article on Wikidata. In 7 B are shown the links included in the dashboard. A link to a Scholia [100] profile allows identification of related articles from a series of pre-made SPARQL queries probing bibliography data on Wikidata. While Scholia provides an overview of a given article, it does not allow direct curation of the metadata. For that, two links are provided, one to Wikidata and one to Author Disambiguator [131]. By accessing the Wikidata page for the entity, one can add new triples, for example curating authors and topics of the article, which are then used by Scholia and by Wikidata Bib's dashboard. Author Disambiguator is a wrapper of an Wikimedia API which facilitates the process of disambiguating author names to unique identifiers on Wikidata, thus feeding the public knowledge graph of publication and authors.

Finally, a link to the article's DOI or full text URL is provided, and serves as a fallback when the automatic download fails. Of note, while the metadata

curation has a technical benefit to Wikidata and the dashboard, it also plays a theoretical role. By curating metadata on authors, the user of Wikidata Bib can better understand the people they read, and expand their metascientific perspective on their domain of interest.

**A**

publication title

citation in Manubot format

tags for document indexing

highlights copied from the main text

links to get extra information and curate metadata (see below)

```

Single-Cell Transcriptome Atlas of Murine Endothelial Cells
=====
[@wikidata:Q89720882]
Publication date : 13 of February, 2020

# Highlights
A comprehensive murine atlas comprising >32,000 single endothelial cell transcriptomes from 11 mouse tissues is reported, and among the subclusters various classical as well as tissue-specialized endothelial cell subtypes are defined.

# Comments

## Tags
--> - 1.4.2. A focus on single-cell RNA sequencing

# Links
* [Scholia Profile](https://scholia.toolforge.org/work/Q89720882)
* [Wikidata](https://www.wikidata.org/wiki/Q89720882)
* [Author Disambiguator](https://author-disambiguator.toolforge.org/work_item_oauth.php?id=Q89720882&batch_id=&match=1&author_list_id=&doit=Get+author+links+for+work)
* [DOI](https://doi.org/10.1016/J.CELL.2020.01.015)
  
```

**B**

Scholia Profile

Related works

Related works from co-citation analysis

Show  entries Search:

Count	Work
3	CD157 Marks Tissue-Resident Endothelial Stem Cells with Homeostatic and Regenerative Properties.
3	A molecular atlas of cell types and zonation in the brain vasculature.

Author Disambiguator

34	[34] Yonglun Luo		
35	[35] Peter Carmeliet	<input checked="" type="checkbox"/> Peter Carmeliet	physician, professor
362	Katholieke Universiteit Leuven		

Match selected authors

## Wikidata entity

main subject

mouse endothelial cell

0 references

mouse endothelial cell  
cell type of Mus musculus

Mouse endothelial cells cross-present lymphocyte-deri...

## Digital Object Identifier (DOI)

Cell Supports open access

RESOURCE | VOLUME 180, ISSUE 4, P764-779 E20, FEBRUARY 20, 2020

Single-Cell Transcriptome Atlas of Murine Endothelial Cells

Joanna Kalucka <sup>1, 10</sup> • Laura P.M.H. de Rooij <sup>10</sup> • Jermaine Goveia <sup>10</sup> • ... Xuri Li <sup>10</sup> • Peter Carmeliet <sup>10, 11</sup> • Show all authors • Show footnotes


Open Archive • Published: February 13, 2020 • DOI: <https://doi.org/10.1016/j.cell.2020.01.015> • Check for updates

**Figure 7:** Wikidata Bib's platform for note taking

The source code for Wikidata Bib is available at [https://github.com/lubianat/wikidata\\_bib](https://github.com/lubianat/wikidata_bib).

## Wikidata Bib as a dashboard

The Wikidata Bib system also enables the reader to get statistics on their readings. Two simple databases are stored on the GitHub repository: \* `read.ttl` - An RDF document recording the dates in which each article was read. \* `read.csv` - An simple, human-readable, index connecting QIDs with article titles. The csv file is only stored for accountability, and as a quick way to glance at the titles read. The .ttl file, in the other hand, is processed by the `update_dashboard.py` script to render 4 different html files under the `docs/` folder: - `index.html` - `last_day.html` - `past_week.html` - `past_month.html`. All files are displayed in a GitHub pages. In the case of this work, they are displayed at [https://lubianat.github.io/wikidata\\_bib/](https://lubianat.github.io/wikidata_bib/).

Figure 8: Wikidata Bib queries for institutions of authors and most read venues

## Wikidata Bib for curation of cells to Wikidata

Articles read with Wikidata Bib were screened for the mention of cell types absent from Wikidata. As discussed on the chapter about the concept of cell type, we considered as a “cell type” as any class of cells described by a domain expert with evidence of reality of its instances. When a mention of such a class appears in an article, I first verify Wikidata for the existence of a related class. If it is absent from the platform, I enter a class name, alongside a superclass, and a QID in a Google Spreadsheet, as shown in Figure 9.

The information from the spreadsheet is pulled by a python script, and processed locally with a series of dictionaries that match common terms to Wikidata IDs. In the example shown in Figure 9, the string “endothelial cell” was matched against a manually curated dictionary to the wikidata entry [Q11394395](https://www.wikidata.org/wiki/Q11394395), the representation of that concept on Wikidata. After reconciling the data, the script uses the Wikidata Integrator python package [87] to insert the new entries on the Wikidata database. The code for integrating a Google Spreadsheet to Wikidata is available at [https://github.com/lubianat/wikidata\\_cell\\_curation](https://github.com/lubianat/wikidata_cell_curation).

## Abstract

The heterogeneity of endothelial cells (ECs) across tissues remains incompletely inventoried. We constructed an atlas of >32,000 single-EC transcriptomes from 11 mouse tissues and identified 78 EC subclusters, including Aqp7<sup>+</sup> intestinal capillaries and angiogenic ECs in healthy tissues. ECs from

The screenshot shows the 'Biocuration of Cell Classes for Wikidata' web application. At the top, there's a menu with 'File', 'Edit', 'View', 'Insert', 'Format', 'Data', 'Tools', 'Add-ons', and 'Help'. Below the menu is a table with four columns: 'label', 'subclass of', 'stated in', and 'aliases'. The 'label' column contains 'angiogenic endothelial cell', 'subclass of' contains 'endothelial cell', 'stated in' contains 'Q89720882', and 'aliases' contains 'angiogenic EC'. A red box highlights 'angiogenic ECs' in the abstract above, with a red arrow pointing to the 'angiogenic EC' alias in the table. A green arrow points from the 'angiogenic endothelial cell' label in the table to the Wikidata page below. The Wikidata page shows the 'angiogenic endothelial cell' (Q109908611) with its description 'cell type' and 'angiogenic EC'. It also lists 'Most relevant properties which are absent: ID' and 'Recoin: Most relevant properties which are absent'. Under the 'Statements' section, it shows 'instance of' as 'cell type' (with 1 reference) and 'subclass of' as 'endothelial cell' (with 0 references).

	B	C	D
label	subclass of	stated in	aliases
angiogenic endothelial cell	endothelial cell	Q89720882	angiogenic EC

angiogenic endothelial cell (Q109908611)...

cell type  
angiogenic EC

► Most relevant properties which are absent: ID  
► Recoin: Most relevant properties which are absent  
► In more languages

Statements

► instance of  
by TiagoLubiana

► cell type ...  
stated in Single-Cell Transcriptome Atlas of Murine Endothelial Cells  
1 reference

► subclass of  
by TiagoLubiana


► endothelial cell ...  
0 references

**Figure 9:** Wikidata Bib was coupled with a biocuration framework for cell types

Wikidata contains 2940 subclasses of "cell ([Q7868](#))" as of 8 December 2021. From those, 550 cell classes are specific for humans and 318 are specific for mice.

As a comparison, as of 8 of December 2021, Wikidata has more cell classes than the Cell Ontology, which lists 2577 classes. It is worth noticing that classes on the Cell Ontology are added after careful consideration by ontologists and domain experts, and should be considered of higher quality than the ones on Wikidata.

From the 2940 cell classes on Wikidata, 2812 (95.6%) have been edited in some way by User:TiagoLubiana, and 1668 (56.7%) have been created by User:TiagoLubiana. Edits made to the cells were often connecting a dangling term, created automatically from an Wikipedia page to the cell subclass hierarchy, but also included adding of identifiers, images, markers and other pieces of information. From the 1668 entities created, approximately 63 species-neutral cell types, 188 human and 188 mouse cell types were added based on PanglaoDB entries (total of 439). The remaining 1229 entries were created either directly via Wikidata's web interface or using the curation workflow described in this chapter. These statistics are a simple demonstration of how the curation system is efficiently contributing to the status of cell type information on Wikidata.

 Figure 10: Subclasses of "cell" on Wikidata

**Figure 10:** Subclasses of "cell" on Wikidata

## Wikidata and the Cell Ontology interplay

The contributions to cell types on Wikidata will be of most value if they are integrated to the current state-of-art of knowledge representation. Arguably, the Cell Ontology is the current leading source of cell type identifiers in the context of the Human Cell Atlas project.[\[48\]](#) Thus, it is crucial that data about cell types on Wikidata is connected to the Cell Ontology.

To start the improvement in the interplay of both databases, we proposed and got approval of a specific Wikidata identifier for the Cell Ontology, the "Cell Ontology ID" (<https://www.wikidata.org/wiki/Property:P7963>). IDs can be added to Wikidata entities and connect them to external databases enabling integrative SPARQL queries. Besides using the common Wikidata interface, one can crowd-curate identifiers via 3rd-party service, Mix'N'Match, which provides an user-friendly framework for connecting identifier catalogs to Wikidata. [135?p=114], as seen in Figure ?? . Logically, we created a Mix'N'Match catalog for harmonizing Cell Ontology IDs to Wikidata (<https://mix-n-match.toolforge.org/#/catalog/4719>), harnessing the community support for the task.

Enter Cell Ontology ID

granulocyte 🔍 Next entry

Entry	116729110
Catalog ID	CL_0000094
Catalog description	

Enter Q number of matching item

Set Q New item N/A

Search

[Search Wikidata](#) | [Search en.wikipedia](#) | [Google-search Wikipedias](#) | [Google-search Wikisource](#) | [Google-search Wikidata](#)

Wikidata search results

[Q223143](#) [↑] **granulocyte**  
mature white blood cells with granules in the cytoplasm 🔍

**Figure 11:** Mix'N'Match curation system

As of early December 2021, more than 700 Cell Ontology IDs have been manually matched to Wikidata. The integration already enables queries that harness the previously existing information on Wikidata for Cell Ontology - based applications. For example, one can query Wikidata items that have (1) a crossref to a CL ID (2) a picture in Wikimedia Commons (<https://w.wiki/4F6e>, Figure ??). The different possibilities of mutual benefit between the Cell Ontology and Wikidata will continue to be explored in the next years of this PhD project.



{#fig:cl\_images width="85%"}

## Final considerations and next steps



To sum up, this PhD research project aims at improving knowledge representation in the context of the Human Cell Atlas. It is composed by a mixture of theoretical studies on conceptual modelling, practical contributions to knowledge organization projects, (mainly the Cell Ontology and Wikidata), explorations of the data to generate biomedical insights and the development of a technical framework for organized reading. By approaching the object of study from a new perspective, we hope not only to make sizeable contributions, but to promote discussion and fruitful conflation of approaches.

The next years of study will be devoted to improving the projects presented here into mature, useful objects. We hope to improve the interplay of Wikidata and Cell Ontology, developing frameworks to combine community- and expert- based curation of knowledge on cell types. Furthermore, we plan to integrate Wikidata to current single-cell RNA-sequencing pipelines by adapting ontology-based R packages (as OnClass [50] and ontoProc[51]) to use Wikidata. Finally, we aim at moving the Wikidata Bib system to a well documented, user-friendly mature system, testing usability with other academics and distributing it as a durable open-source project.

## Additional Work

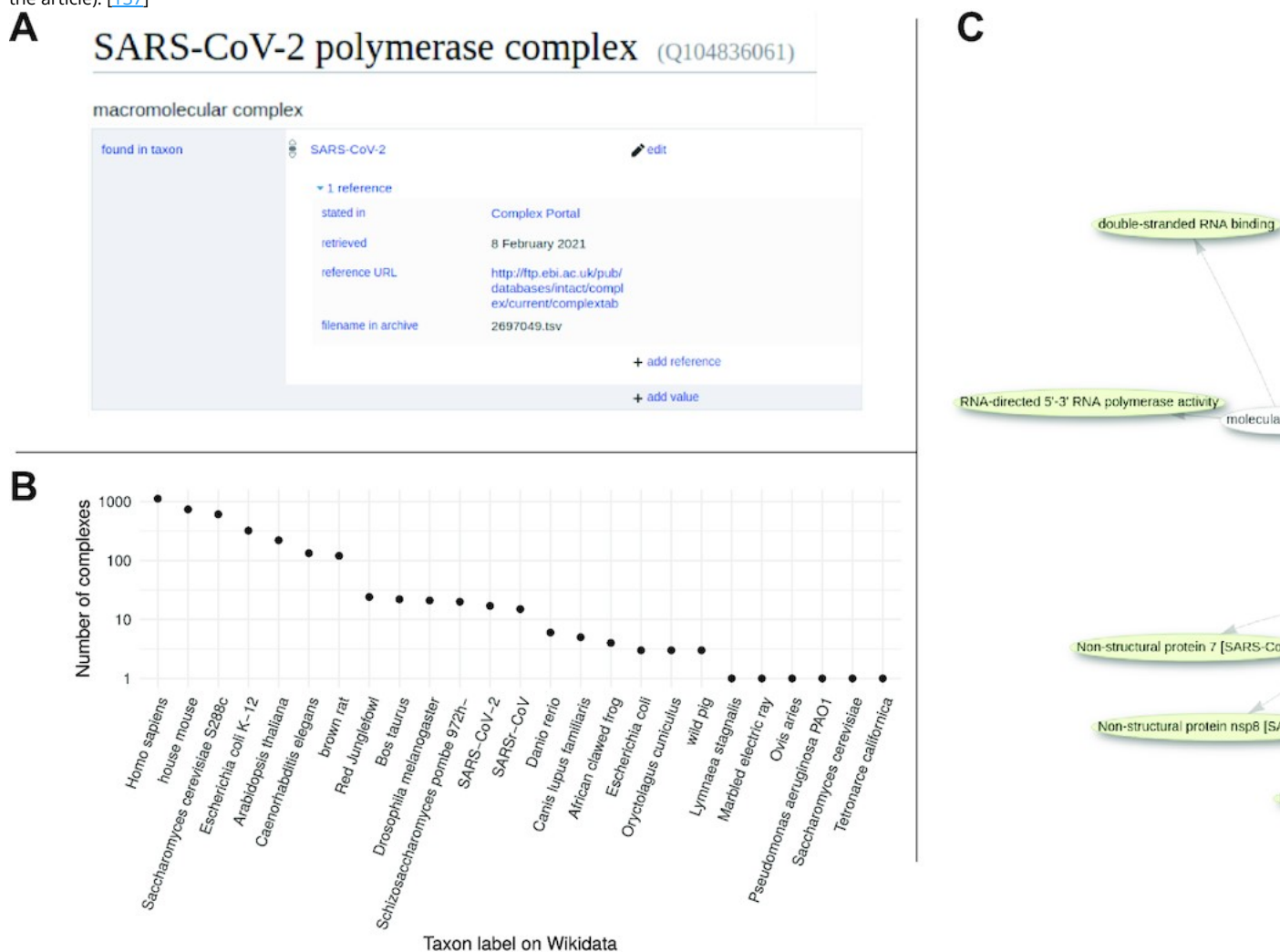
### Collaborations and manuscripts

#### fcoex

During the initial course of this PhD work, we also completed the development and reportin of *fcoex*, an R package for investigating cellular phenotypes using co-expression networks. [136] The software was maintained to withstand new releases of dependencies and new R version, and *WAS PUBLISHED AS A PRE\_PRINT ADD HERE THE LINK*.

#### Wikidata Bots

Alongside the editing of cell-type information on Wikidata, I have joined different efforts to improve biological information on Wikidata. I have collaborated with the ComplexPortal curators, as part of the Virtual Elixir BioHackathon 2020 (<https://github.com/virtual-biohackathons/covid-19-bh20/wiki>) and for the following year, to build an Wikidata Bot to integrate information on protein complexes to Wikidata. An overview of the Wikidata integration is in Figure ??, presented in an article published in Nucleic Acid Research (re-use of the image and legend possible under the CC-BY license of the article). [137]



I have also collaborated with the Cellosaurus database [138] to revive the CellosaurusBot [139], responsible for updating the metadata on more than 100,000 cell lines on Wikidata. The bot code, written in Python, was completely refactored, and is run by me semi-automatically after the Cellosaurus database releases. A write-up of the integration is in progress, and is planned for release/submission in the first semester of 2022.

### Systematic Reviews and publishing of intermediary tables

Finally, in a collaboration with Olavo Amaral and Kleber, from the Brazilian Reproducibility Initiative [140] I wrote a commentary on the value of publishing intermediate datasets as citable products. [141/] The pieces discuss the value of small curations done both in systematic reviews and by

experimentalists in the course of their research projects. Published curation tables can serve as a source for improving the ecosystem of open knowledge, not less by reconciliation to Wikidata (thereby bridging the commentary with this project)

## WiseCube - enterprise biomedical question and answering

During a part of this project, I have worked part-time as a consultant for the Wisecube company, based in Seattle, United States. [142] The job was approved by FAPESP, and consisted mainly in writing SPARQL queries that probe Wikidata for answers to the questions posed by the BioASQ competition. [143] It also entails on-demand curation of biomedical topics on Wikidata based on requests by pharmaceutical companies as well as the development of dashboards targeted at providing insights to customers.

## Awards and participation in events

During the initial course of this PhD project, I have participated in several events:

- (Feb-2021) Presented an open talk at the “Semana da Bioinformática” event about modelling of biological systems (1020 views as of December 2021) [144, =VDvCxskIGE]
- (Jun-Aug 2021) Helped to organize the No-Budget-Science HackWeek virtual hackathon [145]
- (Jul - 2021) Presented the work “Wikidata for 5-star Linked Open Databases: A case study of PanglaoDB” at the Bio-Ontologies section of the Annual International Conference on Intelligent Systems for Molecular Biology. [146]. The presentation was awarded the best
- (Jul - 2021) Awarded the 2nd place in the International Society for Computational Biology (ISCB) Wikipedia Competition for the contributions to the Wikipedia page on Biocuration (<https://en.wikipedia.org/wiki/Biocuration>) [147]
- (Nov - 2021) Managed a project during BioHackathon Europe 2021, in Barcelona, Spain, on the representation of ELIXIR information on Wikidata. [148]

## Course work

During the first year of the PhD program, I took 4 different classes, acquiring a total of 36 academic credits. Figure 12 displays the disciplines taken, available only in portuguese.

### 95131 - 8945857/2 - Tiago Lubiana Alves

Sigla	Nome da Disciplina	Início	Término	Carga Horária	Cred.	Freq.	Conc.	Exc.	Situação
SCC5929-2/4	Introdução à Web Semântica (Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo)	24/08/2020	15/12/2020	180	12	75	A	N	Concluída
SCC5908-3/5	Introdução ao Processamento de Língua Natural (Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo)	26/08/2020	18/12/2020	180	12	100	A	N	Concluída
MAC6967-1/1	Laboratório Avançado de Ciência de Dados (Instituto de Matemática e Estatística - Universidade de São Paulo)	31/08/2020	18/12/2020	120	8	93	A	N	Concluída
ICB5774-1/2	O Significado de Modelos e Teorias em Ciências Biológicas (Instituto de Ciências Biomédicas - Universidade de São Paulo)	10/05/2021	18/07/2021	60	4	100	A	N	Concluída

	Créditos mínimos exigidos		Créditos obtidos
	Para exame de qualificação	Para depósito de tese	
Disciplinas:	0	32	36
Estágios:			
Total:	0	32	36

Créditos Atribuídos à Tese: 140

Figure 12: Courses taken

## 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](https://doi.org/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](https://doi.org/10.7554/elife.27041)
3. **The Human Cell Atlas and equity: lessons learned**  
Partha P Majumder, Musa M Mhlanga, Alex K Shalek  
*Nature Medicine* (2020-10-01) <https://www.wikidata.org/wiki/Q100491106>  
DOI: [10.1038/s41591-020-1100-4](https://doi.org/10.1038/s41591-020-1100-4)
4. **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>
5. **Everyone needs a data-management plan**  
Nature  
(2018-03-15) <https://www.wikidata.org/wiki/Q56524391>  
DOI: [10.1038/d41586-018-03065-z](https://doi.org/10.1038/d41586-018-03065-z)
6. **About the Data Coordination Platform**  
HCA Data Portal  
<https://data.humancellatlas.org/about/>
7. **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](https://doi.org/10.1016/j.cels.2017.03.006)
8. **A periodic table of cell types**  
Bo Xia, Itai Yanai  
*Development* (2019-06-15) <https://doi.org/ggctwf>  
DOI: [10.1242/dev.169854](https://doi.org/10.1242/dev.169854) · PMID: [31249003](https://pubmed.ncbi.nlm.nih.gov/31249003/) · PMCID: [PMC6602355](https://pubmed.ncbi.nlm.nih.gov/PMC6602355/)
9. **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](https://doi.org/10.1242/dev.178996) · PMID: [31537583](https://pubmed.ncbi.nlm.nih.gov/31537583/)
10. **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](https://doi.org/10.1038/550451a)
11. **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)
12. **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)
13. **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)
14. **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)
15. **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)
16. **Evolution of Cellular Differentiation: From Hypotheses to Models**  
Pedro Márquez-Zacarias, 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)

17. **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)
18. **Towards a pragmatic definition of cell type**  
Tiago Lubiana, Helder Nakaya  
(2021-01-04) <https://www.wikidata.org/wiki/Q108723646>  
DOI: [10.22541/au.160979530.02627436/v1](https://doi.org/10.22541/au.160979530.02627436/v1)
19. **BioNumbers--the database of key numbers in molecular and cell biology**  
Ron Milo, Paul Jorgensen, Uri Moran, Griffin M Weber, Michael Springer  
*Nucleic Acids Research* (2010-01-01) <https://www.wikidata.org/wiki/Q24643881>  
DOI: [10.1093/nar/gkp889](https://doi.org/10.1093/nar/gkp889)
20. **Wikipedia, the free encyclopedia** [https://en.wikipedia.org/wiki/Main\\_Page](https://en.wikipedia.org/wiki/Main_Page)
21. **Cell types and ontologies of the Human Cell Atlas**  
David Osumi-Sutherland, Chuan Xu, Maria C Keays, Peter V Kharchenko, Aviv Regev, Ed S Lein, Sarah Teichmann  
(2021-06-28) <https://www.wikidata.org/wiki/Q107373831>
22. **An estimation of the number of cells in the human body**  
Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, ... Silvia Canaider  
*Annals of Human Biology* (2013-07-05) <https://www.wikidata.org/wiki/Q34037445>  
DOI: [10.3109/03014460.2013.807878](https://doi.org/10.3109/03014460.2013.807878)
23. **A curated database reveals trends in single-cell transcriptomics**  
Valentine Svensson, Eduardo da Veiga Beltrame, Lior Pachter  
*Database* (2020-11-01) <https://www.wikidata.org/wiki/Q103034964>  
DOI: [10.1093/database/baaa073](https://doi.org/10.1093/database/baaa073)
24. **The evolution of cell types in animals: emerging principles from molecular studies.**  
Detlev Arendt  
*Nature reviews. Genetics* (2008-11) <https://www.ncbi.nlm.nih.gov/pubmed/18927580>  
DOI: [10.1038/nrg2416](https://doi.org/10.1038/nrg2416) · PMID: [18927580](https://pubmed.ncbi.nlm.nih.gov/18927580/)
25. **The origin and evolution of cell types**  
Detlev Arendt, Jacob M Musser, Clare VH Baker, Aviv Bergman, Connie Cepko, Douglas H Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D Laubichler, Günter P Wagner  
*Nature Reviews Genetics* (2016-11-07) <https://doi.org/f9b62x>  
DOI: [10.1038/nrg.2016.127](https://doi.org/10.1038/nrg.2016.127) · PMID: [27818507](https://pubmed.ncbi.nlm.nih.gov/27818507/)
26. **Stem cell states, fates, and the rules of attraction.**  
Tariq Enver, Martin Pera, Carsten Peterson, Peter W Andrews  
*Cell Stem Cell* (2009-05-01) <https://www.wikidata.org/wiki/Q37475461>  
DOI: [10.1016/j.stem.2009.04.011](https://doi.org/10.1016/j.stem.2009.04.011)
27. **Theory of cell fate**  
Michael J Casey, Patrick S Stumpf, Ben D MacArthur  
*Wiley interdisciplinary reviews. Systems biology and medicine* (2019-12-12) <https://www.wikidata.org/wiki/Q91908361>  
DOI: [10.1002/wsbm.1471](https://doi.org/10.1002/wsbm.1471)
28. **Perspectives on defining cell types in the brain**  
Eran A Mukamel, John Ngai  
*Current Opinion in Neurobiology* (2018-12-06) <https://www.wikidata.org/wiki/Q90361677>  
DOI: [10.1016/j.conb.2018.11.007](https://doi.org/10.1016/j.conb.2018.11.007)
29. **Ensembles, dynamics, and cell types: Revisiting the statistical mechanics perspective on cellular regulation**  
Stefan Bornholdt, Stuart Kauffman  
*Journal of Theoretical Biology* (2019-01-31) <https://www.wikidata.org/wiki/Q91316993>  
DOI: [10.1016/j.jtbi.2019.01.036](https://doi.org/10.1016/j.jtbi.2019.01.036)
30. **Species Concepts and Species Delimitation**  
Kevin De Queiroz  
*Systematic Biology* (2007-12) <https://doi.org/c34kzf>  
DOI: [10.1080/10635150701701083](https://doi.org/10.1080/10635150701701083) · PMID: [18027281](https://pubmed.ncbi.nlm.nih.gov/18027281/)
31. **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)
32. **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://pubmed.ncbi.nlm.nih.gov/PMC3024222/)
33. **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 Saritvijai, ... Christopher J Mungall  
*Journal of Biomedical Semantics* (2016-07-04) <https://doi.org/gg99b9>

DOI: [10.1186/s13326-016-0088-7](https://doi.org/10.1186/s13326-016-0088-7) · PMID: [27377652](https://pubmed.ncbi.nlm.nih.gov/27377652/) · PMCID: [PMC4932724](https://pubmed.ncbi.nlm.nih.gov/PMC4932724)

34. **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](https://doi.org/10.1186/1471-2105-14-228)
35. **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](https://doi.org/10.1186/s12859-017-1976-2) · PMID: [29322916](https://pubmed.ncbi.nlm.nih.gov/29322916/) · PMCID: [PMC5763506](https://pubmed.ncbi.nlm.nih.gov/PMC5763506)
36. **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](https://doi.org/10.1186/s12859-019-2721-9) · PMID: [31272374](https://pubmed.ncbi.nlm.nih.gov/31272374/) · PMCID: [PMC6509796](https://pubmed.ncbi.nlm.nih.gov/PMC6509796)
37. **Scaled, high fidelity electrophysiological, morphological, and transcriptomic cell characterization**  
Brian R Lee, Agata Budzillo, Kristen Hadley, Jeremy A Miller, Tim Jarsky, Katherine Baker, Dijon Hill, Lisa Kim, Rusty Mann, Lindsay Ng, ... Jim Berg  
*eLife* (2021-08-13) <https://www.wikidata.org/wiki/Q109717199>  
DOI: [10.7554/elife.65482](https://doi.org/10.7554/elife.65482)
38. **Current best practices in single-cell RNA-seq analysis: a tutorial**  
Malte D Luecken, Fabian J Theis  
*Molecular Systems Biology* (2019-06-19) <https://www.wikidata.org/wiki/Q64974172>  
DOI: [10.15252/msb.20188746](https://doi.org/10.15252/msb.20188746)
39. **Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods**  
Zoe A Clarke, Tallulah Andrews, Jawairia Atif, Delaram Pouyababar, Brendan T Innes, Sonya A MacParland, Gary D Bader  
*Nature Protocols* (2021-05-24) <https://www.wikidata.org/wiki/Q107158224>  
DOI: [10.1038/s41596-021-00534-0](https://doi.org/10.1038/s41596-021-00534-0)
40. **Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis**  
Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, El-ad D Amir, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, ... Garry P Nolan  
*Cell* (2015-06-18) <https://www.wikidata.org/wiki/Q30975629>  
DOI: [10.1016/j.cell.2015.05.047](https://doi.org/10.1016/j.cell.2015.05.047)
41. **Fast unfolding of communities in large networks**  
Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre  
*Journal of Statistical Mechanics: Theory and Experiment* (2008-10-09) <https://www.wikidata.org/wiki/Q29305711>  
DOI: [10.1088/1742-5468/2008/10/p10008](https://doi.org/10.1088/1742-5468/2008/10/p10008)
42. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**  
Oscar Franzén, Li-Ming Gan, Johan LM Björkegren  
*Database* (2019-01-01) <https://www.wikidata.org/wiki/Q63664483>  
DOI: [10.1093/database/baz046](https://doi.org/10.1093/database/baz046)
43. **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](https://doi.org/10.1093/nar/gky900)
44. **Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data**  
Julie A McMurphy, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Melanie Courtot, John Deck, Michel Dumontier, Donal K Fellows, ... Helen Parkinson  
*PLOS Biology* (2017-06-29) <https://www.wikidata.org/wiki/Q33037209>  
DOI: [10.1371/journal.pbio.2001414](https://doi.org/10.1371/journal.pbio.2001414)
45. **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](https://doi.org/10.1186/s13326-016-0088-7)
46. **Cell type discovery using single-cell transcriptomics: implications for ontological representation**  
Brian D Aevermann, Mark Novotny, Trygve E Bakken, Jeremy A Miller, Alexander D Diehl, David Osumi-Sutherland, Roger S Lasken, Ed S Lein, Richard H Scheuermann  
*Human Molecular Genetics* (2018-05-01) <https://www.wikidata.org/wiki/Q52625486>  
DOI: [10.1093/hmg/ddy100](https://doi.org/10.1093/hmg/ddy100)
47. **Cell ontology in an age of data-driven cell classification.**  
David Osumi-Sutherland, David Osumi-Sutherland  
*BMC Bioinformatics* (2017-12-21) <https://www.wikidata.org/wiki/Q49192555>  
DOI: [10.1186/s12859-017-1980-6](https://doi.org/10.1186/s12859-017-1980-6)
48. **Cell type ontologies of the Human Cell Atlas**  
David Osumi-Sutherland, Chuan Xu, Maria Keays, Adam P Levine, Peter V Kharchenko, Aviv Regev, Ed Lein, Sarah Teichmann  
*Nature Cell Biology* (2021-11-01) <https://www.wikidata.org/wiki/Q109755180>  
DOI: [10.1038/s41556-021-00787-7](https://doi.org/10.1038/s41556-021-00787-7)
49. **Besca, a single-cell transcriptomics analysis toolkit to accelerate translational research**



Sophia Clara Mädler, Alice Julien-Laferrriere, Luis Wyss, Miroslav Phan, Albert SW Kang, Eric Ulrich, Roland Schmucki, Jitao David Zhang, Martin Ebeling, Laura Badi, ... Klas Hatje  
*bioRxiv* (2020-08-12) <https://www.wikidata.org/wiki/Q104450593>  
DOI: [10.1101/2020.08.11.245795](https://doi.org/10.1101/2020.08.11.245795)

50. **Leveraging the Cell Ontology to classify unseen cell types**  
Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbić, Marinka Žitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanas, Russ Altman  
*Nature Communications* (2021-09-21) <https://www.wikidata.org/wiki/Q108929315>  
DOI: [10.1038/s41467-021-25725-x](https://doi.org/10.1038/s41467-021-25725-x)
51. **ontoProc: processing of ontologies of anatomy, cell lines, and so on** <https://www.wikidata.org/wiki/Q101074371>
52. **Tabula Muris** <https://tabula-muris.ds.czbiohub.org/>
53. **Tabula Sapiens** <https://tabula-sapiens-portal.ds.czbiohub.org/celltypes>
54. **Azimuth** <https://azimuth.hubmapconsortium.org/>
55. **Construction and Usage of a Human Body Common Coordinate Framework Comprising Clinical, Semantic, and Spatial Ontologies**  
Katy Börner, Ellen Quardokus, Bruce WHerr II, Leonard E Cross, Elizabeth G Record, Yingnan Ju, Andreas D Bueckle, James P Sluka, Jonathan C Silverstein, Kristen M Browne, ... Griffin M Weber  
(2020-07-28) <https://www.wikidata.org/wiki/Q109755184>
56. **Cell Annotation Platform | Coming Soon** <http://celltype.info/>
57. **Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture across the human body**  
Conde C Domínguez, Tomás Gomes, Lorna B Jarvis, C Xu, SK Howlett, DB Rainbow, Ondrej Suchanek, Hamish W King, Lira Mamanova, Krzysztof Polański, ... Sarah Teichmann  
(2021-04-28) <https://www.wikidata.org/wiki/Q107363182>  
DOI: [10.1101/2021.04.28.441762](https://doi.org/10.1101/2021.04.28.441762)
58. **Ontology based molecular signatures for immune cell types via gene expression analysis**  
Terrence F Meehan, Nicole Vasilevsky, Christopher J Mungall, David S Dougall, Melissa Haendel, Judith A Blake, Alexander D Diehl  
*BMC Bioinformatics* (2013-08-30) <https://www.wikidata.org/wiki/Q34978215>  
DOI: [10.1186/1471-2105-14-263](https://doi.org/10.1186/1471-2105-14-263)
59. **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://www.wikidata.org/wiki/Q33786317>  
DOI: [10.1186/1471-2105-12-6](https://doi.org/10.1186/1471-2105-12-6)
60. **Ontologies for the life sciences**  
Steffen Schulze-Kremer, Barry Smith  
(2005-11-15) <https://www.wikidata.org/wiki/Q105870680>  
DOI: [10.1002/047001153x.g408213](https://doi.org/10.1002/047001153x.g408213)
61. **The Philosophy of Logical Atomism, Lecture 1: Facts and Propositions** <https://www.wikidata.org/wiki/Q105105637>
62. **Logik der Forschung**  
Karl Popper  
(1934-01-01) <https://www.wikidata.org/wiki/Q1868040>
63. **The semantic conception of truth: and the foundations of semantics**  
Alfred Tarski  
*Philosophy and Phenomenological Research* (1944-03-01) <https://www.wikidata.org/wiki/Q106090790>  
DOI: [10.2307/2102968](https://doi.org/10.2307/2102968)
64. **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](https://doi.org/10.1093/nar/gkaa1113)
65. **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](https://doi.org/10.1038/75556)
66. **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](https://doi.org/10.1038/nbt1346)
67. **Introducing the Knowledge Graph: things, not strings**  
Google  
(2012-05-16) <https://blog.google/products/search/introducing-knowledge-graph-things-not/>
68. **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](https://doi.org/10.1002/asi.20870)
69. **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](https://doi.org/10.1145/2872427.2874809)

70. **Wikibase/DataModel - MediaWiki** <https://www.mediawiki.org/wiki/Wikibase/DataModel>
71. **Help:Data type - Wikidata** [https://www.wikidata.org/wiki/Help:Data\\_type](https://www.wikidata.org/wiki/Help:Data_type)
72. **Wikidata:Statistics - Wikidata** <https://www.wikidata.org/wiki/Wikidata:Statistics>
73. **Help:Multilingual - Wikidata** <https://www.wikidata.org/wiki/Help:Multilingual>
74. **RDF 1.1 Semantics** <https://www.w3.org/TR/rdf11-nt/>
75. **Wikidata:Data access - Wikidata** [https://www.wikidata.org/wiki/Wikidata:Data\\_access](https://www.wikidata.org/wiki/Wikidata:Data_access)
76. **WikidataR package - RDocumentation** <https://www.rdocumentation.org/packages/WikidataR/versions/2.2.0>
77. **wikidata2df: Utility package for easily turning a SPARQL query into a dataframe**  
João Vitor F Cavalcante  
<https://github.com/jvfe/wikidata2df>
78. **Wikidata:Licensing - Wikidata** <https://www.wikidata.org/wiki/Wikidata:Licensing>
79. <https://query.wikidata.org/>
80. [Q56010228](https://www.wikidata.org/wiki/Q56010228)
81. **Scholia**  
Scholia  
<https://scholia.toolforge.org/>
82. **SARS-CoV-2-Queries**  
SARS-CoV-2-Queries  
<https://egonw.github.io/SARS-CoV-2-Queries/>
83. **Wikidata:Tools/OpenRefine - Wikidata** <https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine>
84. **Help:QuickStatements - Wikidata** <https://www.wikidata.org/wiki/Help:QuickStatements>
85. **Wikidata:Bots - Wikidata** <https://www.wikidata.org/wiki/Wikidata:Bots>
86. **Wikidata:Pywikibot - Python 3 Tutorial - Wikidata** [https://www.wikidata.org/wiki/Wikidata:Pywikibot - Python 3 Tutorial](https://www.wikidata.org/wiki/Wikidata:Pywikibot_-_Python_3_Tutorial)
87. **GitHub - SuLab/WikidataIntegrator: A Wikidata Python module integrating the MediaWiki API and the Wikidata SPARQL endpoint**  
GitHub  
<https://github.com/SuLab/WikidataIntegrator>
88. **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](https://doi.org/10.7554/elife.52614)
89. **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* (2019-09-23) <https://www.wikidata.org/wiki/Q68471881>  
DOI: [10.1016/j.jbi.2019.103292](https://doi.org/10.1016/j.jbi.2019.103292)
90. **Big data: Wikiomics**  
Mitch Waldrop  
*Nature* (2008-09-04) <https://www.wikidata.org/wiki/Q28292893>  
DOI: [10.1038/455022a](https://doi.org/10.1038/455022a)
91. **Calling on a million minds for community annotation in WikiProteins**  
Barend Mons, Michael Ashburner, Christine Chichester, Erik M van Mulligen, Marc Weeber, Johan den Dunnen, Gert-Jan van Ommen, Mark A Musen, Matt Cockerill, Henning Hermjakob, ... Amos Bairoch  
*Genome Biology* (2008-01-01) <https://www.wikidata.org/wiki/Q21183907>  
DOI: [10.1186/gb-2008-9-5-r89](https://doi.org/10.1186/gb-2008-9-5-r89)
92. **Ten Simple Rules for Developing Public Biological Databases**  
Mohamed Helmy, Alexander Crits-Christoph, Gary D Bader  
*PLOS Computational Biology* (2016-11-01) <https://www.wikidata.org/wiki/Q28595967>  
DOI: [10.1371/journal.pcbi.1005128](https://doi.org/10.1371/journal.pcbi.1005128)
93. **Inside the Alexa-Friendly World of Wikidata**  
Tom Simonite  
*Wired* <https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/>
94. **A gene wiki for community annotation of gene function**  
Jon W Huss, Camilo Orozco, James Goodale, Chunlei Wu, Serge Batalov, Tim J Vickers, Faramarz Valafar, Andrew I Su  
*PLOS Biology* (2008-07-08) <https://www.wikidata.org/wiki/Q21092744>  
DOI: [10.1371/journal.pbio.0060175](https://doi.org/10.1371/journal.pbio.0060175)
95. **Making your database available through Wikipedia: the pros and cons**  
Robert D Finn, Paul P Gardner, Alex Bateman  
*Nucleic Acids Research* (2012-01-01) <https://www.wikidata.org/wiki/Q28254676>  
DOI: [10.1093/nar/gkr1195](https://doi.org/10.1093/nar/gkr1195)

96. **Wikidata as a semantic framework for the Gene Wiki initiative**  
Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mittra, Julia Turner, Timothy Elliott Putman, Justin Leong, Chinmay Naik, Paul Pavlidis, Lynn Schriml, Benjamin M Good, Andrew I Su  
*Database* (2016-01-01) <https://www.wikidata.org/wiki/Q23712646>  
DOI: [10.1093/database/baw015](https://doi.org/10.1093/database/baw015)
97. **WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata**  
Timothy Elliott Putman, Sebastien Lelong, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Colin Diesh, Nathan Dunn, Monica Munoz-Torres, Gregory Stupp, Chunlei Wu, Andrew I Su, Benjamin M Good  
*Database* (2017-03-08) <https://www.wikidata.org/wiki/Q28529449>
98. **ChlamBase: a curated model organism database for the Chlamydia research community**  
Timothy Elliott Putman, Kevin Hybiske, Derek Jow, Cyrus Afrasiabi, Sebastien Lelong, Marco Alvarado Cano, Chunlei Wu, Andrew I Su  
*Database* (2019-01-01) <https://www.wikidata.org/wiki/Q63286185>  
DOI: [10.1093/database/baz041](https://doi.org/10.1093/database/baz041)
99. **Submit a Topic Page to PLOS Computational Biology and Wikipedia**  
Daniel Mietchen, Shoshana Wodak, Szymon Wasik, Natalia Szostak, Christophe Dessimoz  
*PLOS Computational Biology* (2018-05-31) <https://www.wikidata.org/wiki/Q54655231>  
DOI: [10.1371/journal.pcbi.1006137](https://doi.org/10.1371/journal.pcbi.1006137)
100. **Scholia, Scientometrics and Wikidata**  
Finn Årup Nielsen, Daniel Mietchen, Egon Willighagen  
*The Semantic Web: ESWC 2017 Satellite Events* (2017-10-01) <https://www.wikidata.org/wiki/Q41799194>  
DOI: [10.1007/978-3-319-70407-4\\_36](https://doi.org/10.1007/978-3-319-70407-4_36)
101. **Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata**  
Lane Rasberry, Egon Willighagen, Finn Årup Nielsen, Daniel Mietchen  
*Research Ideas and Outcomes* (2019-05-02) <https://www.wikidata.org/wiki/Q63433973>  
DOI: [10.3897/rio.5.e35820](https://doi.org/10.3897/rio.5.e35820)
102. **Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata**  
Houcemeddine Turki, Mohamed Ali Hadj Taieb, Thomas Shafee, Tiago Lubiana, Dariusz Jemielniak, Mohamed Ben Aouicha, José Emilio Labra Gayo, Eric Youngstrom, Mossab Banat, Diptanshu Das, ... WikiProject COVID-19  
*Semantic Web: Interoperability, Usability, Applicability* (2021-09-28) <https://www.wikidata.org/wiki/Q108766311>  
DOI: [10.3233/sw-210444](https://doi.org/10.3233/sw-210444)
103. **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 J 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](https://doi.org/10.1186/s12915-020-00940-y)
104. **Wikidata Queries around the SARS-CoV-2 virus and pandemic** <https://www.wikidata.org/wiki/Q88647643>
105. **COVIWD: COVID-19 Wikidata Dashboard**  
Fariz Darari  
*Jurnal Ilmu Komputer dan Informasi* (2021-03-01) <https://www.wikidata.org/wiki/Q105833381>  
DOI: [10.21609/jiki.v14i1.941](https://doi.org/10.21609/jiki.v14i1.941)
106. **Painel de informação sobre a COVID-19: consultas SPARQL na Wikidata**  
Ana Carolina Simionato Arakaki, Fabiano Ferreira de Castro, Felipe Augusto Arakaki  
*Atoz: Novas Práticas em Informação e Conhecimento* (2020-12-03) <https://www.wikidata.org/wiki/Q106249454>  
DOI: [10.5380/atoz.v9i2.76684](https://doi.org/10.5380/atoz.v9i2.76684)
107. **Uso de Wikidata y Wikipedia para la generación asistida de un vocabulario estructurado multilingüe sobre la pandemia de Covid-19**  
Tomás Saorín, Juan-Antonio Pastor-Sánchez, María-José Baños-Moreno  
*Profesional de la Información* (2020-09-13) <https://www.wikidata.org/wiki/Q107377131>  
DOI: [10.3145/epi.2020.sep.09](https://doi.org/10.3145/epi.2020.sep.09)
108. **The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals**  
Frederic B Bastian, Julien Roux, Anne Niknejad, Aurélie Comte, Sara SFonseca Costa, Tarcisio M Farias, Sébastien Moretti, Gilles Parmentier, Valentine Rech de Laval, Marta Rosikiewicz, ... Marc Robinson-Rechavi  
*Nucleic Acids Research* (2020-10-10) <https://www.wikidata.org/wiki/Q100513179>  
DOI: [10.1093/nar/gkaa793](https://doi.org/10.1093/nar/gkaa793)
109. **Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages: a pilot study**  
Alexander Pfundner, Tobias Schönberg, John Horn, Richard David Boyce, Matthias Samwald  
*Journal of Medical Internet Research* (2015-05-05) <https://www.wikidata.org/wiki/Q21503276>  
DOI: [10.2196/jmir.4163](https://doi.org/10.2196/jmir.4163)
110. **Pesquisa-ação: uma introdução metodológica**  
David Tripp  
*Educação e Pesquisa* (2005-12-01) <https://www.wikidata.org/wiki/Q108479295>  
DOI: [10.1590/s1517-97022005000300009](https://doi.org/10.1590/s1517-97022005000300009)
111. **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/bfpg/elx029](https://doi.org/10.1093/bfpg/elx029)
112. **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](https://doi.org/10.22541/au.160979530.02627436/v1)

113. **PhyloCode** <https://www.wikidata.org/wiki/Q1189395>
114. **Ontological realism: A methodology for coordinated evolution of scientific ontologies**  
Barry Smith, Werner Ceusters  
*Applied Ontology* (2010-11-15) <https://www.wikidata.org/wiki/Q28239464>  
DOI: [10.3233/ao-2010-0079](https://doi.org/10.3233/ao-2010-0079)
115. **Multi-level ontology-based conceptual modeling**  
Victorio A Carvalho, João Paulo A Almeida, Claudenir M Fonseca, Giancarlo Guizzardi  
*Data and Knowledge Engineering* (2017-05-01) <https://www.wikidata.org/wiki/Q108926456>  
DOI: [10.1016/j.datak.2017.03.002](https://doi.org/10.1016/j.datak.2017.03.002)
116. **Popper on Definitions**  
Wilhelm Büttemeyer  
*Zeitschrift für allgemeine Wissenschaftstheorie. Journal for general philosophy of science* (2005-01-01) <https://www.wikidata.org/wiki/Q108925548>  
DOI: [10.1007/s10838-005-6037-2](https://doi.org/10.1007/s10838-005-6037-2)
117. **PanglaoDB - A Single Cell Sequencing Resource For Gene Expression Data** <https://panglaoDB.se/index.html>
118. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**  
Oscar Franzén, Li-Ming Gan, Johan LM Björkegren  
*Database* (2019) <https://doi.org/ggkzxr>  
DOI: [10.1093/database/baz046](https://doi.org/10.1093/database/baz046) · PMID: [30951143](https://pubmed.ncbi.nlm.nih.gov/30951143/) · PMCID: [PMC6450036](https://pubmed.ncbi.nlm.nih.gov/PMC6450036/)
119. **Linked Data - Design Issues** <https://www.w3.org/DesignIssues/LinkedData.html>
120. **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](https://doi.org/10.1093/nar/gkt1264)
121. **Type or Individual? Evidence of Large-Scale Conceptual Disarray in Wikidata**  
Atilio A Dadalto, João Paulo A Almeida, Claudenir M Fonseca, Giancarlo Guizzardi  
*Lecture Notes in Computer Science* (2021-01-01) <https://www.wikidata.org/wiki/Q109990743>
122. **UniProt** <https://sparql.uniprot.org/sparql>
123. **Portal:Semantic Web - WikiPathways** [https://www.wikipathways.org/index.php/Portal:Semantic\\_Web](https://www.wikipathways.org/index.php/Portal:Semantic_Web)
124. **Cell Markers**  
Konstantin Yakimchuk  
*Materials and Methods* (2013-05-02) <https://doi.org/ghq494>  
DOI: [10.13070/mm.en.3.183](https://doi.org/10.13070/mm.en.3.183)
125. **SHOGoin: Shogoin Human Omics database for the Generation of iPS and Normal cells** <https://stemcellinformatics.org/>
126. **Como fazer um fichamento**  
Priscilla de Carvalho Nunes disse  
*Blog da Biblioteca da ECA-USP* (2019-09-30) <https://bibliotecaeca.wordpress.com/2019/09/30/como-fazer-um-fichamento/>
127. <https://www.youtube.com/playlist?list>
128. **Come si fa una tesi di laurea** <https://www.wikidata.org/wiki/Q3684178>
129. **Unpaywall** <https://unpaywall.org/>
130. **Clean Code: A Handbook of Agile Software Craftsmanship** <https://www.wikidata.org/wiki/Q109996684>
131. **Wikidata:Tools/Author Disambiguator - Wikidata** [https://www.wikidata.org/wiki/Wikidata:Tools/Author\\_Disambiguator](https://www.wikidata.org/wiki/Wikidata:Tools/Author_Disambiguator)
132. **wbib: A helper for building Wikidata-based literature dashboards via SPARQL queries.**  
Tiago Lubiana  
<https://github.com/lubianat/wbib>
133. **HCA Latin America - 2021 Workshop** <https://www.humancellatlas.org/hca-latin-america-2021-workshop/>
134. **BioHackathon Europe** <https://biohackathon-europe.org/>
135. **The Whelming › Tech, tools, and tribulations**  
Scott Allan Wallick  
<http://magnusmanske.de/wordpress/>
136. **fcoex: FCBF-based Co-Expression Networks for Single Cells**  
Tiago Lubiana, Helder Nakaya  
*Bioconductor version: Release (3.14)* (2021) <https://bioconductor.org/packages/fcoex/>
137. **Complex Portal 2022: new curation frontiers**  
Birgit HM Meldal, Livia Perfetto, Colin Combe, Tiago Lubiana, João Vitor Ferreira Cavalcante, Hema Bye-A-Jee, Andra Waagmeester, Noemi del-Toro, Anjali Shrivastava, Elisabeth Barrera, ... Sandra Orchard  
*Nucleic Acids Research* (2021-10-29) <https://www.wikidata.org/wiki/Q109348309>  
DOI: [10.1093/nar/gkab991](https://doi.org/10.1093/nar/gkab991)
138. **The Cellosaurus, a cell-line knowledge resource.**

Amos Bairoch

*Journal of Biomolecular Techniques* (2018-05-01) <https://www.wikidata.org/wiki/Q54370168>

DOI: [10.7171/jbt.18-2902-002](https://doi.org/10.7171/jbt.18-2902-002)

139. **User:CellosaurusBot - Wikidata** <https://www.wikidata.org/wiki/User:CellosaurusBot>
140. **The Brazilian Reproducibility Initiative**  
Ana P Wasilewska-Sampaio, Olavo Bohrer Amaral, Kleber Neves, Ana P Wasilewska-Sampaio, Clarissa FD Carneiro, Olavo Bohrer Amaral, Clarissa FD Carneiro  
*eLife* (2019-02-05) <https://www.wikidata.org/wiki/Q61799268>  
DOI: [10.7554/elife.41602](https://doi.org/10.7554/elife.41602)
141. <https://osf.io/preprints/metaarxiv/vbwa9>
142. **Wisecube AI | Knowledge Graph Engine** <https://www.wisecube.ai/>
143. **An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition**  
George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, ... Georgios Paliouras  
*BMC Bioinformatics* (2015-04-30) <https://www.wikidata.org/wiki/Q28646342>  
DOI: [10.1186/s12859-015-0564-6](https://doi.org/10.1186/s12859-015-0564-6)
144. **YouTube** <https://www.youtube.com/>
145. **No Budget Science Hack Week**  
reprodutibilidade  
<https://www.reprodutibilidade.bio.br/hack-week-2021>
146. **Wikidata for 5-star Linked Open Databases: A case study of PanglaoDB**  
Tiago Lubiana, João Vitor Ferreira Cavalcante  
*Zenodo* (2021-12-01) <https://doi.org/gnpzvr>  
DOI: [10.5281/zenodo.5747849](https://doi.org/10.5281/zenodo.5747849)
147. **Biocuration - Wikipedia** <https://en.wikipedia.org/wiki/Biocuration>
148. **biohackathon-projects-2021/projects/32 at main · elixir-europe/biohackathon-projects-2021**  
GitHub  
<https://github.com/elixir-europe/biohackathon-projects-2021>