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

This manuscript ([permalink](https://lubianat.github.io/quali_phd/v/bdc5dbf4cb3698c1da65b9939fa935424b927361/)) was automatically generated from [lubianat/quali\_phd@bdc5dbf](https://github.com/lubianat/quali_phd/tree/bdc5dbf4cb3698c1da65b9939fa935424b927361) on December 9, 2021.

## Authors

* **Tiago Lubiana** ORCID icon [0000-0003-2473-2313](https://orcid.org/0000-0003-2473-2313) · GitHub icon [lubianat](https://github.com/lubianat) · Twitter icon [lubianat](https://twitter.com/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 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 significantly impact 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 knowledge management tools. 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 review 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 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.

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 classifying 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 into Wikidata, based on a notion of species-specific cell type clarified in the preceding chapter.

Then, we present Wikidata Bib, a framework for an organized reading of the literature. The framework, although used as a method throughout the PhD project, is presented in the results session. We emphasize 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. 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 is 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 for a deep knowledge of cells, the building blocks of life [[1](#ref-pNGap1Du)]. The Human Cell Atlas (HCA) project has been a major player in the cell knowledge ecosystem, running since 2017 to characterize every cell type in the human body [[2](#ref-1GmbExweg)]. The HCA consortium gathers people from all over the world to tackle different parts of the project to have a diverse and equitable account of the cell type diversity. [[3](#ref-tjdjR2Xf)]

Building a complete 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](#ref-1GmbExweg),[4](#ref-kkwRTArg)]. Every lab inside the project will contribute with its expertise, providing samples representing human diversity.

HCA is set to revolutionize the biomedical sciences by creating tools and standards for basic research, allowing better characterization of disease, and improving diagnostics and therapy. Its products (data, information, knowledge and wisdom) need to be FAIR: findable, accessible, interoperable and reusable. Data stewardship and management are growing as core demands of the scientific community, ranging from data management plans [[5](#ref-1DSEIjFha)] to specialized data personnel [[5](#ref-1DSEIjFha)].

The Human Cell Atlas has a dedicated team for organizing data: the Data Coordination Platform (DCP) [[6](#ref-zDRzmIGu)] [[4](#ref-kkwRTArg)]. The DCP is responsible for tracing the plan for computational interoperability, from the data generators to the consumers.[[4](#ref-kkwRTArg)]. The Human Cell Atlas has its portal for data [[7](#ref-kX6KnbUo)]), which composes the data repository landscape with other resources, like the Broad Institute Single Cell Portal [https://singlecell.broadinstitute.org/single\_cell>) and the Chan-Zuckerberg Biohub Tabula Sapiens (<https://tabula-sapiens-portal.ds.czbiohub.org/>). In addition to its core team, the HCA is poised to grow by community interaction. It 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](#ref-1GmbExweg)]  
Thus, this project inserts itself among the wide-ranging scientific discussions to improve data - and knowledge - interoperability. ]) The work also paves the way for Wikidata reconciling of other databases for cell-type markers, such as CellMarker [[8](#ref-chGii6yw)], labome [[9](#ref-rhRRCtlA)], CellFinder [[10](#ref-4AEy2xhQ)] and SHOGoiN/CELLPEDIA [[11](#ref-6uWWsiSq)/]) (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?**](#X2d0964a6cb7c228bf3d61384638532c561d652c)], 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.” [[12](#ref-uYuz0opI), =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?**](#X8751d1dc64638abd56385e63924893147dfc301)], 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 [[13](#ref-rPKBwmYh)/][[14](#ref-PKhuVRW8)\_da26C-QW5qiS7uZ]. An influential work in presenting methods for academic reading in the humanities is Umberto Eco’s book “How to Write a Thesis” [[15](#ref-1HBVPtZGp)], 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. [[14](#ref-PKhuVRW8)\_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 accompaning 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 commited 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 [[16](#ref-15luL9zZC)/] 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. [[15](#ref-1HBVPtZGp)] 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. [[17](#ref-13HqB23xH)] Every day, I read one article of each list, using the notetaking station displayed in Figure [1](#fig:notetaking). 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 [1](#fig:notetaking) 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 [1](#fig:notetaking) B are shown the links included in the dashboard. A link to a Scholia [[18](#ref-hxzL9pmm)] 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 [[19](#ref-1A9RvszKC)]. By acessing 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 publict 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.

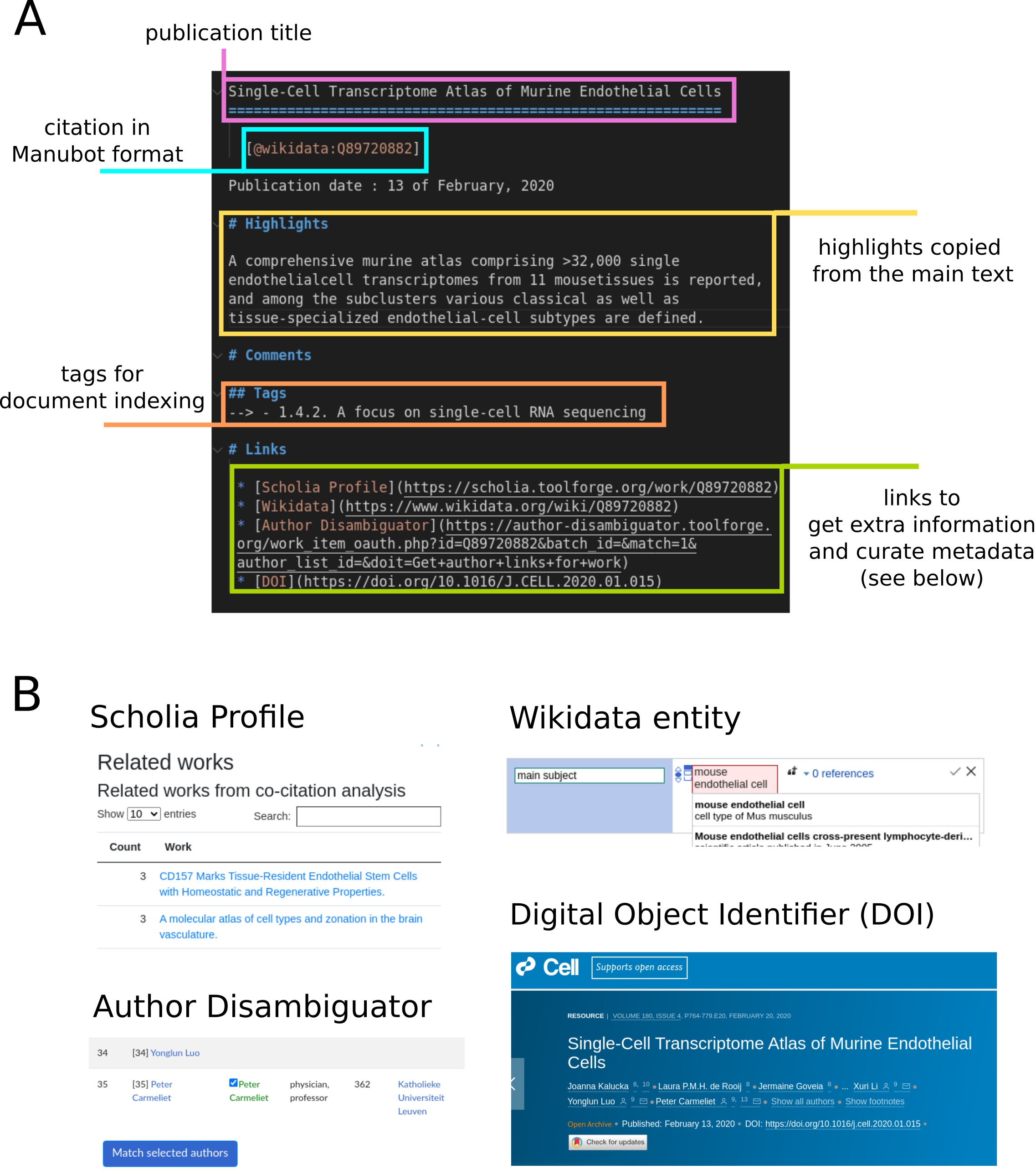


Figure 1: Wikidata Bib’s platform for note taking

The source code for Wikidata Bib is available at 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/.

To organize the code for rendering the dashboard, we created a python package, wbib, and deposited it in PyPi, making it available via pip. [[20](#ref-6chnW6cc)/]. The package implements the logic for rendering complex Wikidata-based academic dashboards and is available in GitHub at https://github.com/lubianat/wbib. It allows the user to build dashboards based on Wikidata records of information such as gender of authors, the region of authors intitutions, topics of articles and similar metascientific information. The dashboard is composed of SPARQL queries written for the Wikidata Query Service [@url:https://query.wikidata.org] It also allows users to feed an arbitrary list of articles and obtain a custom dashboard. Wikidata Bib obtains the html dashboards after feeding wbib the lists of articles read in total (index.html) or in pre-determined time spans (last\_day.html,past\_week.htmlandpast\_month.html` )

Figure 2: Wikidata Bib queries for institutions of authors and most read venues

Figure 2: Wikidata Bib queries for institutions of authors and most read venues

The dashboard includes not only a basic list of read articles, but also statistics on most read authors and most read venues. It also displays an interactive map of the instituions of articles read, permitting a glance on geographic biases in activities. An example of queries is show in [2](#fig:dashboard). As the queries are rendered live, they evolve in quality with the growth of Wikidata. Finally, the clean 5-star-open data format enables users to adapt the queries to include different aspects of Wikidata. For example, table ?? showcases 10 articles that (1) I have read in the past year and (2) were authored by a speaker of the 1st Human Cell Atlas Latin America Single Cell RNA-seqData Analysis Workshop [[21](#ref-1hag8XE6)/]. One practical application that the dashboard enables, thus, is to identify people in an event, institution or location that the user has read before, therefore catalysing the possibility of colaborations. Anedoctally, this strategy was tested successfully at Biohackathon Europe 2021 [[22](#ref-kHL3NVxk)/], where I used the system both to identify possible collaborators and as a conversation starter. |workLabel |authors | |————————————————————————————————-|————————————————————————–| |A promoter-level mammalian expression atlas |Jay W Shin | |Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.|Muzlifah Haniffa | |The Human Cell Atlas. |Musa Mhlanga, Jay W Shin, Muzlifah Haniffa, Menna R Clatworthy, Dana Pe’er| |The Human Cell Atlas: Technical approaches and challenges. |Jay W Shin | |Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. |Dana Pe’er | |Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations |Sonya A MacParland | |Single-cell reconstruction of the early maternal–fetal interface in humans |Muzlifah Haniffa | |Distinct microbial and immune niches of the human colon |Rasa Elmentaite, Menna R Clatworthy | |A cell atlas of human thymic development defines T cell repertoire formation |Muzlifah Haniffa, Menna R Clatworthy | |Decoding human fetal liver haematopoiesis |Muzlifah Haniffa | Table: Articles read by Tiago Lubiana before 8 December 2021 in which an author was a speaker at HCA Latin America {#tbl:articles\_read\_hca} |

## Wikidata Bib for curation of cells to Wikidata

The Wikidata Bib system was devised originally to allow an overview of the fields of cell classification and biocuration. However, during the process, it was also repurposed for biocuration of new cell classes in Wikidata. By fast-tracking the reading of new articles, Wikidata Bib enables an efficient parsing of the literature, and, thus, the identification of previously uncatalogued cell types.

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 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 [3](#fig:biocuration_of_cells).

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 [3](#fig:biocuration_of_cells), 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 [[23](#ref-qDI8I4IJ)] 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.

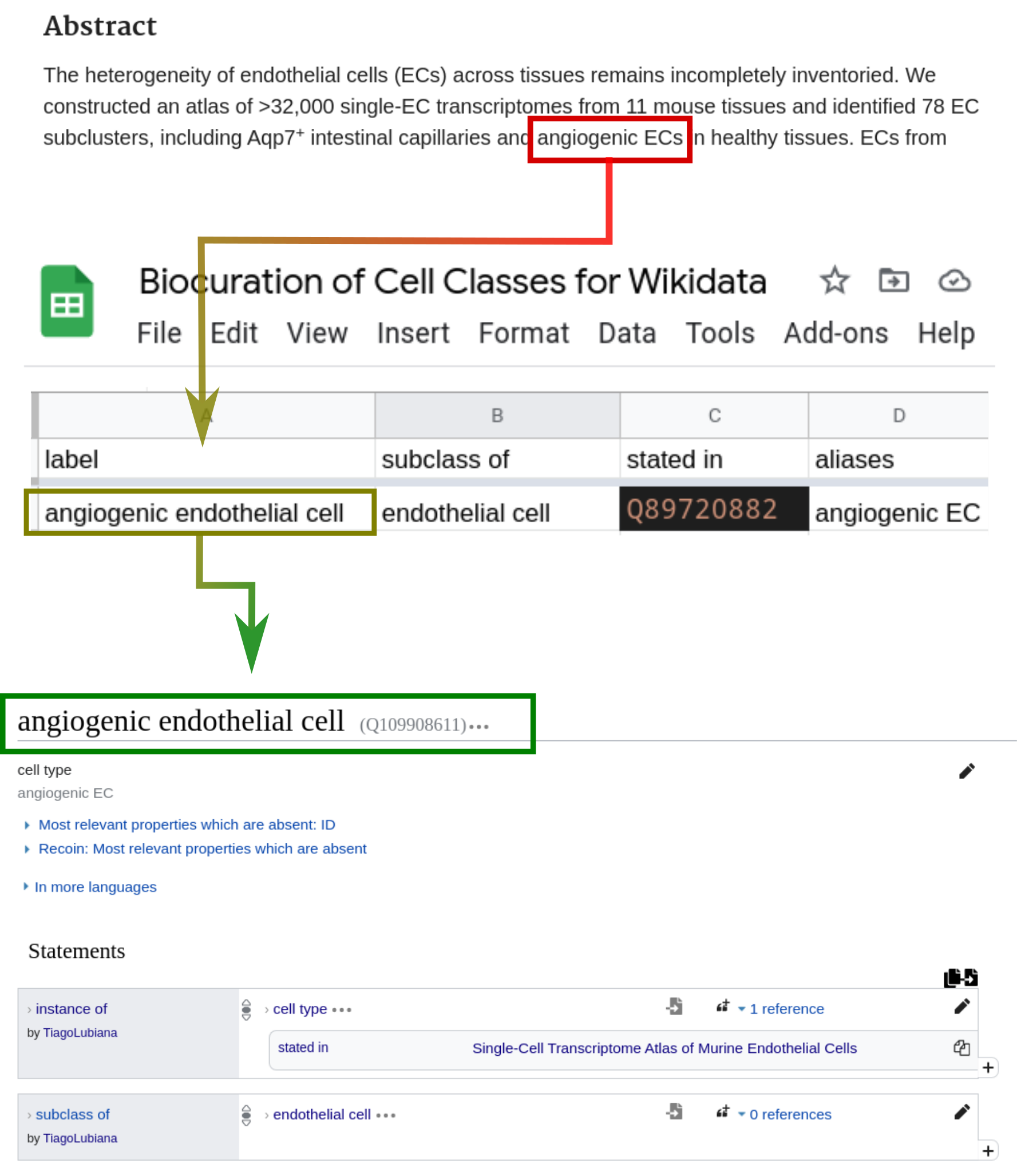


Figure 3: Wikidata Bib was coupled with a biocuration framework for cell types

Wikidata contains 2940 subclasses of “cell ([Q7868](https://www.wikidata.org/wiki/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. Theses statistics are a simple demonstration of how the curation system is efficiently contributing to the status of cell type information on Wikidata.

![Figure 4: Subclasses of “cell” on Wikidata](data:text/html; charset=UTF-8;base64,)

Figure 4: 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.[[24](#ref-qT8WxqjA)] 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 idenfier catalogs to Wikidata. [[25](#ref-JgiKEEdq)/?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.

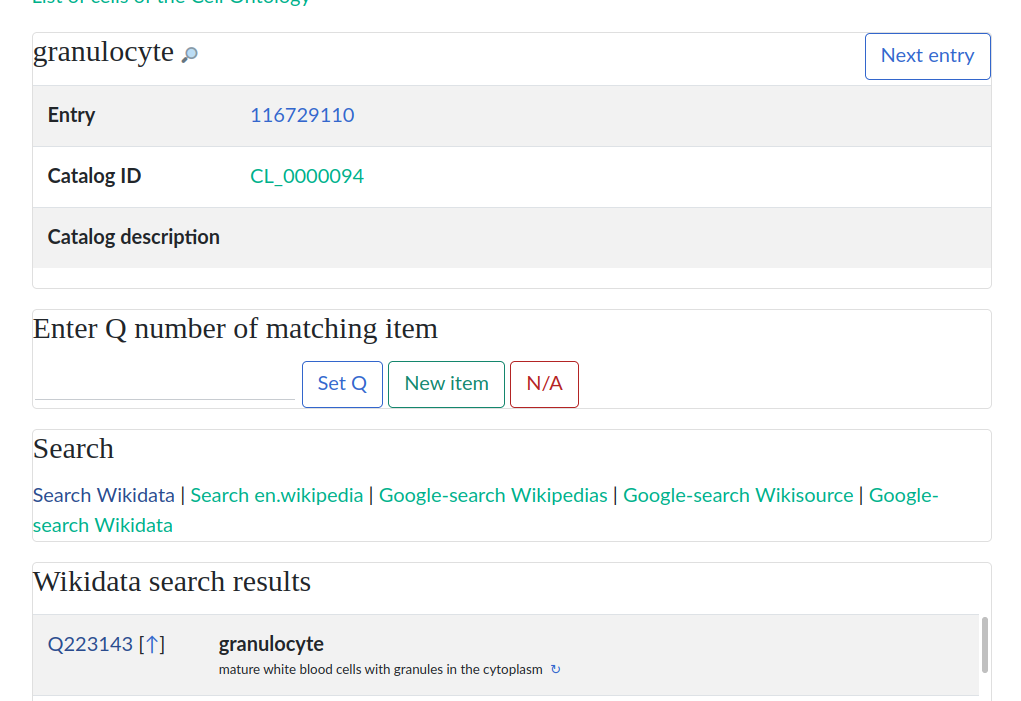


Figure 5: Mix’N’Match curation system

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 [[26](#ref-sW6aNZJB)] and ontoProc[[27](#ref-15YmDXALp)])) 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.

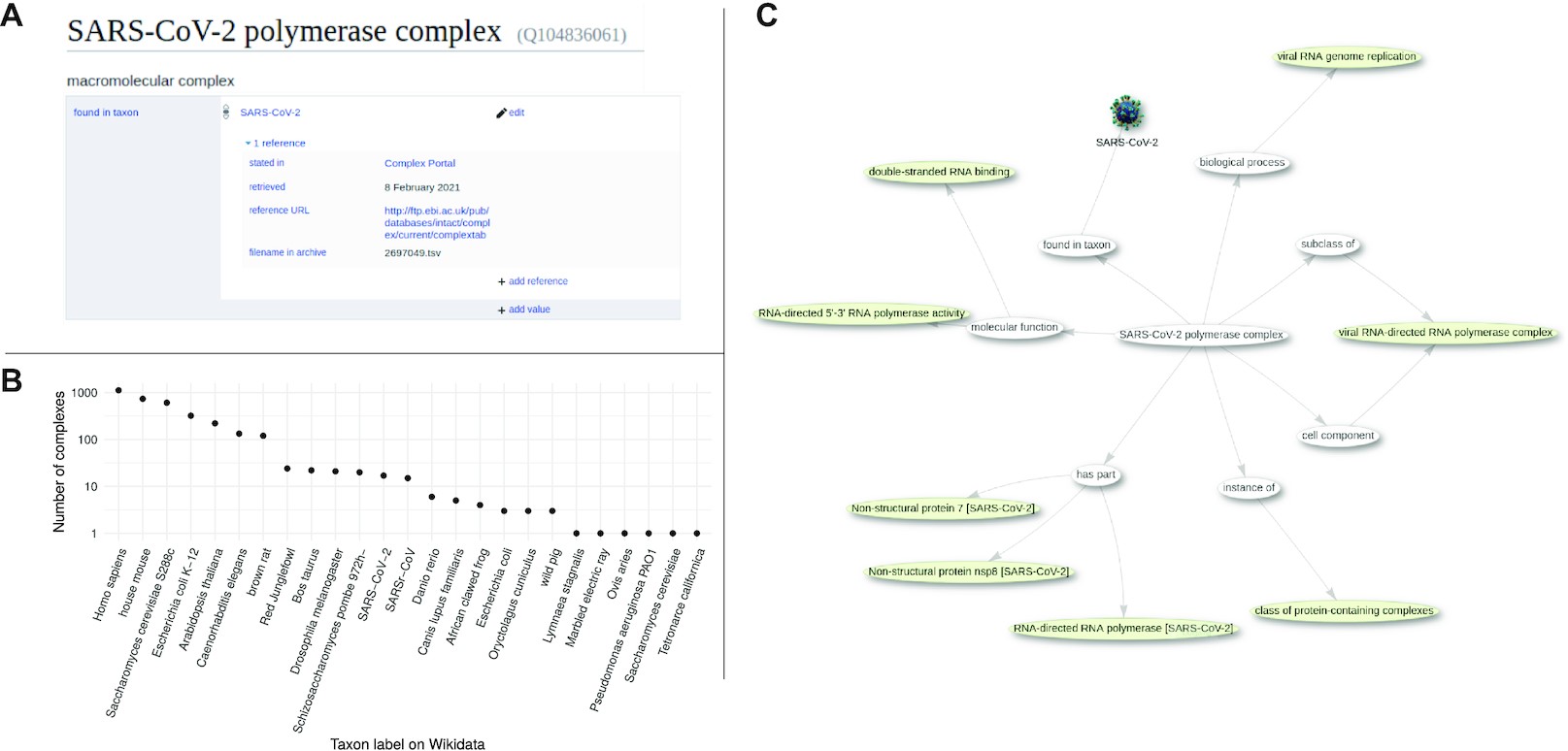
# Additional Work

## 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. [[28](#ref-MxIeSJYt)] The software was mantained 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). [[29](#ref-CQRJ53gu)] 

I have also collaborated with the Cellosaurus database [[30](#ref-1DguATd9G)] to revive the CellosaurusBot [[31](#ref-lMQxhx3q)], 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 he integration is in progress, and is planned for release/submisson 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 [[32](#ref-F2mYjDJ0)] I wrote a commentary on the value of publishing intermediate datasets as citable products. [[33](#ref-HbZ13t8l)/] 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. [[34](#ref-gdYsBE7d)/] 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. [[35](#ref-rkXotO9x)] 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) [[36](#ref-1C9uHr1Zk), =VDvCxskIGEI]
* (Jun-Aug 2021) Helped to organize the No-Budget-Science HackWeek virtual hackathon [[37](#ref-12LOzmXRs)]
* (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. [[38](#ref-SALI6Ywb)]. 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>) [[39](#ref-IJG65hFm)]
* (Nov - 2021) Managed a project during BioHackathon Europe 2021, in Barcelona, Spain, on the representation of ELIXIR information on Wikidata. [[40](#ref-14Wi842eZ)]

## Course work

During the first year of the PhD program, I took 4 different classes, acquiring a total of 36 academic credits. Figure [6](#fig:courses_taken) displays the disciplines taken, available only in portuguese.

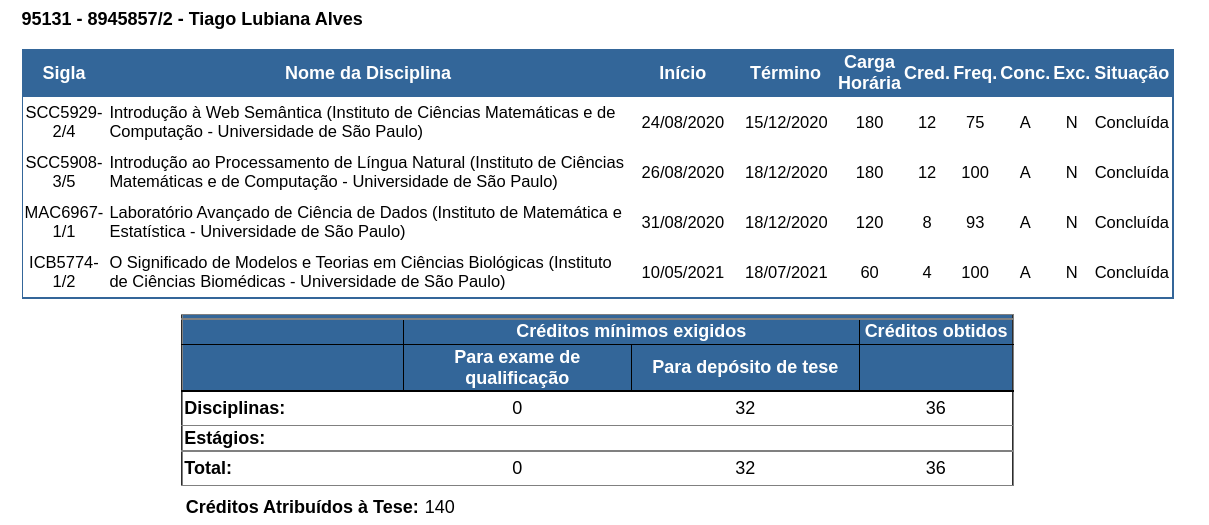


Figure 6: Courses taken

## 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. **Mapping the Human Body at the Cellular Level** HCA Data Portal <https://data.humancellatlas.org/>

8. **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)

9. **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)

10. **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)

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

12. **Wikipedia, the free encyclopedia** <https://en.wikipedia.org/wiki/Main_Page>

13. **Como fazer um fichamento** Priscilla de Carvalho Nunes disse *Blog da Biblioteca da ECA-USP* (2019-09-30) <https://bibliotecadaeca.wordpress.com/2019/09/30/como-fazer-um-fichamento/>

14. <https://www.youtube.com/playlist?list>

15. **Come si fa una tesi di laurea** <https://www.wikidata.org/wiki/Q3684178>

16. **Unpaywall** <https://unpaywall.org/>

17. **Clean Code: A Handbook of Agile Software Craftsmanship** <https://www.wikidata.org/wiki/Q109996684>

18. **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)

19. **Wikidata:Tools/Author Disambiguator - Wikidata** <https://www.wikidata.org/wiki/Wikidata:Tools/Author_Disambiguator>

20. **wbib: A helper for building Wikidata-based literature dashboards via SPARQL queries.** Tiago Lubiana <https://github.com/lubianat/wbib>

21. **HCA Latin America - 2021 Workshop** <https://www.humancellatlas.org/hca-latin-america-2021-workshop/>

22. **BioHackathon Europe** <https://biohackathon-europe.org/>

23. **GitHub - SuLab/WikidataIntegrator: A Wikidata Python module integrating the MediaWiki API and the Wikidata SPARQL endpoint** GitHub <https://github.com/SuLab/WikidataIntegrator>

24. **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)

25. **The Whelming › Tech, tools, and tribulations** Scott Allan Wallick <http://magnusmanske.de/wordpress/>

26. **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 Karkanias, 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)

27. **ontoProc: processing of ontologies of anatomy, cell lines, and so on** <https://www.wikidata.org/wiki/Q101074371>

28. **fcoex: FCBF-based Co-Expression Networks for Single Cells** Tiago Lubiana, Helder Nakaya *Bioconductor version: Release (3.14)* (2021) <https://bioconductor.org/packages/fcoex/>

29. **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)

30. **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)

31. **User:CellosaurusBot - Wikidata** <https://www.wikidata.org/wiki/User:CellosaurusBot>

32. **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)

33. <https://osf.io/preprints/metaarxiv/vbwa9>

34. **Wisecube AI | Knowledge Graph Engine** <https://www.wisecube.ai/>

35. **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)

36. **YouTube** <https://www.youtube.com/>

37. **No Budget Science Hack Week** reprodutibilidade <https://www.reprodutibilidade.bio.br/hack-week-2021>

38. **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)

39. **Biocuration - Wikipedia** <https://en.wikipedia.org/wiki/Biocuration>

40. **biohackathon-projects-2021/projects/32 at main · elixir-europe/biohackathon-projects-2021** GitHub <https://github.com/elixir-europe/biohackathon-projects-2021>