Manuscript Title

Authors

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

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

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

Doctorate Project - Report 2 FAPESP process number: #2019/26284-1 Project term: 01/08/2020 to 31/07/2024 Period of Scientific report: 10/07/2021 to 10/07/2022

Scholarship Holder: Tiago Lubiana

Responsible Researcher: Prof. Dr. Helder Takashi Imoto Nakaya

São Paulo - SP July 2022

Project summary

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

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

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

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

Modern advancements of literature-based discovery rely on Natural Language Processing, Machine Learning and Knowledge graphs to make inferences on literature knowledge. Word embeddings, for example, are leading inference of properties of compounds based on their shared neighbourhood of words (the words before and after their mentiongs) with known compounds, thus making use of latent knowledge in the body of knowledge. [8] Other, more explicit approaches, rely on extracted relations embedded in knowledge graphs,fo example, the discovery of new RNA-binding proteins related to Amyotrophic Lateral Sclerosis by analysis of the Watson Drug Discovery gene-disease network. [9] Knowledge graphs have a set of characteristics that make then useful for Literature Based Discovery: the power of representing multiple relations, the power of making inferences on top of those relations, and provide human understandability at every step, allowing for a dialog between

expert humans and computing systems. The field of biomedical ontologies explores that direction in depth, and the community is building many solutions, widely applicable for the biomedical sciences.

An ontology, as used here, is a formal computational representation of reality, which tries to represent each concept (and their relations) as precisely as possible. [10]
Constructing an ontology is a process of selecting and defining terms and 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. [11] [12].

The Gene Ontology is part of a much larger effort to formalize concepts across biology: the Open Biomedical and Biological Ontologies (OBO) Foundry. [13] Created in 2007, the OBO Foundry is a hub of biomedical ontologies that sets guidelines for the design and construction of high-quality ontologies. The initial OBO Foundry united several independent ontologies, like the Cell Ontology (CL), the Disease Ontology (DO) and the Protein Ontology (PRO) under a common framework, a great progress towards interoperability. At the same time, the creation of the Relation Ontology (RO) provided a go-to point for relations in biology that could them be reused by different ontologies.

Ontologies are powerful, but require a high degree of tecnichal expertise to get started. Recently, a new approach for formal knowledge representation arose with the dawn of collaborative knowledge graphs. Wikidata, the collaborative knowledge graph of the Wikimedia foundation, allows users to contribute with classes and statements, in the same spirit of Wikipedia and share its "epistemic virtues, like power, speed and availability. [14] It is powerful because of its large community of contributors. With a community of more than 25,000 active editors (https://www.wikidata.org/wiki/Wikidata:Statistics) and growing, it is able to cover a much wider number of concepts than any user individually. It is fast, because one does not need to install any software or ask for permissions to update it: any user can simply do it via a web interface. That speed makes it easier for newcomers to join and contribute, in contrast to OBO Foundry ontologies, which require extensive training on semantics and knowledge of Git/GitHub for contributions. Finally, the information on Wikidata is available via an user interface, via a SPARQL query service and as large, full-size database dumps, providing full extent reusability. The Wikidata model has been so sucessfull that Google decided to migrate its own knowledge base, Freebase, fully into Wikidata.[15]

Several advances towards biological data integration and biological data analysis in Wikidata have been made before, yielding positive results [16] [17] and showcasing its potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [18]. Wikidata has been proposed as a unified base to gather and distribute biomedical knowledge, with more than 50 000 human gene items indexed and hundreds of biomedical-related properties [19].

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

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

Achievements

In the second year of the project, we addressed 2 different facets of the project.

The first facet was the development of guidelines for describing new cell types in published research in partnership with the Cell Ontology.

The second facet was the formalization of a biocuration routine to curate new cell types on Wikidata, an effort with 2 main products: the consolidation of Wikidata as a database for cell types and Wikidata Bib, a software and workflow for reading large amounts of scientific literature.

Additionally, we collaborated in several different projects related to organizing biomedical knowledge on Wikidata, which are also detailed in this session. ## Biocuration of cell classes with Wikidata Bib {.page_break_before} ### Introduction

Reading scientific articles is an integral part of the routine of modern scientists. Although several literature-management software are available [20], the process of reading is mainly 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 the efficient screening of scientific articles.

Other professional traditions have dealt with similar issues in the past. Notetaking is vital to keep track of financial balances and avoid costly problems in accounting. Double-entry bookkeeping was developed in the 13th century as a professional solution for notetaking in accounting where "every entry to an account requires a corresponding and opposite entry to a different account." [21, =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 [22], 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 standard academic training in the humanities [23][**url?**]. An influential work in presenting methods for academic reading in the humanities is Umberto Eco's book "How to Write a Thesis" [24], 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. Still, its theoretical scope (roughly the humanities) and its date preceding the digital era limits the extent to which it applies to the biomedical sciences.

Notably, the need for 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—[**url?**]. 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 experimentation and data analysis's theoretical and practical basis. With the boom of scientific articles, however, the scientific literature (and accompanying public datasets) already provide a strong material for sculpting scientific projects. Thus, developing a methodology for academic reading tailored to the digital environment is a need.

To address this gap, we are developing Wikidata Bib, a framework for large scale reading of scientific articles. It is presented in three parts, each with a technical overview alongside the theoretical foundations. First, Wikidata Bib is presented 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 users 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, Python 3 scripts and SPARQL queries. The code is packaged into a python module to facilitate usage. It also uses the Click library to implement a professional Command Line Interface for end users (https://github.com/pallets/click). It has a standard file structure, summarized as the following:

- `docs/`
  - `index.html`
- `downloads/`
  - `10.7554_ELIFE.52614.pdf`
- `src/`
  - `data/`
    - `config.yaml`
    - `index.yaml`
    - `read.csv`
    - `read.ttl`
    - `toread.yaml`
  - `notes/`
    - `Q87830400.md`
  - `wikidata_bib/`
    - `get_pdf.py`
    - `read_paper.py`
    - `update_dashboard.py`
    - …
- `LICENSE`
- `pyproject.toml`
- `README.md`
- `setup.cfg`

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 `src/` directory contains a `data/` subfolder with the configuration files and the local database of what is read, a `notes` subfolder with the notes on the read articles and a `wikidata_bib` directory containing the actualpython code with the system's mechanics.

After installing the package using the `pip` utility (https://pypi.org/), the user is able to use Wikidata Bib from the terminal as any other command line utility. The basic commands available are described below: - `bib read` receives a Wikidata QID for an article and opens the article web page on a browser and opens a personalized notes document. It also updates the static files for the analytics dashboard, stored in the `docs/` directory. - `bib pop`, which "pops" an article from the reading stash at `toread.yaml` and runs `bib read` for it - `bib log` updates the git repository and adds, commits and pushes notes and the dashboard to to GitHub

To customize the system, a new user would need to update 3 `yaml` files in the `src/data` folder: `toread.yaml`, `index.yaml` and `config.yaml`. The `toread.yaml` stores in a plain text format the QIDs to be read. The file can be edited either manually or via automatic queries (not discussed in this report for conciseness). The `index.yaml` file holds an hierarchy of topics of interest for the researcher, to be used for indexing/tagging the notes in the way required by the topic of research. This file plays the role of Umberto Eco's "work plan", with the topics of interest for the academic. [24] Finally, the `config.yaml` contains options to customize the system for advanced use. For example, the user can configure shortcuts for different reading lists. For example, in my `to read.yaml` file there are two reading lists, one entitled `Cell types` header and another entitled `Biocuration`. My `config.yaml` contains the following snippet:

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

The `config.yaml` shortcuts are used as arguments by the `bib pop` command, where `$ bib pop ct` retrieves an article from the "Cell types" list, while `$ bib pop bioc` retrieves an article from the "Biocuration" list.

The Wikidata bib framework is coupled with a discipline of daily reading. The discipline 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 might be approached. [25] Every day, I read one article of each list, using the notetaking station displayed in Figure [**fig?**]: notetaking. The constancy of reading allows steady coverage of the relevant literature. While the discipline has worked for this research project, it is not required to use the Wikidata Bib system.

The notetaking station of Wikidata Bib, opened in Virtual Studio Code, is depicted on Figure [**fig?**]: notetaking A. The title and publication dates are displayed, and the reading process entails copying snippets from the text to the "Highlights" session. Copying the highlights into plain text makes the sections of interest searchable via command line using `grep` (https://en.wikipedia.org/w/wiki/Grep). Comments can be added either in the comment section or inline, alongside the highlights. Even though there is no enforcement, our recommendation is to use HTML comment syntax `<!--` `Comment goes here -->` to differentiate from highlights. Visual Studio offers a default shortcut (CTRL + /) for adding comments, making the process seamless.

The Wikidata Bib framework also includes, whenever possible, an improvement of the metadata about the article on Wikidata. In [**fig?**]: notetaking B are shown the links included in the dashboard. A link to a Scholia [26] 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 [27]. 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 that facilitates 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. Furthermore, by embedding the metascientific curation in the reading process WikidataBib directs the user to get acquainted with the people producing the science they read, thus contributing to a more complete perspective of the research process.
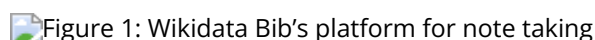
Figure 1: Wikidata Bib's platform for note taking

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, on 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 static website. 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` . [28]. 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 author's institutions, topics of articles and similar metascientific information. The dashboard is composed of SPARQL queries written for the Wikidata Query Service [29] 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.html` and `past_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 institutions of articles read, permitting a glance at geographic biases in activities. An example of queries is shown in 2. 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 1 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 [30]. 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 catalyzing the possibility of collaborations. Anecdotally, this strategy was tested successfully at Biohackathon Europe 2021 [31], where we used the system both to identify possible collaborators and as a conversation starter.

Table 1: Articles read by Tiago Lubiana before 8 December 2021 in which an author was a speaker at HCA Latin America |

workLabel	authors
A promoter-level mammalian expression atlas	Jay W Shin

workLabel	authors
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

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 to mention cell types absent from Wikidata. As discussed in the chapter about the concept of cell type, we considered a "cell type" as any class of cells described by a domain expert with evidence of the 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.

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, the string "endothelial cell" was matched against a manually curated dictionary to the Wikidata entry Q11394395, the representation of that concept on Wikidata. After reconciling the data, the script uses the Wikidata Integrator python package [32] 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

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

Wikidata contains 4075 subclasses of "cell (Q7868)" as of 16 of June of 2022. From those, 550 cell classes are specific for humans, and 318 are specific for mice.
Currently 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 4075 cell classes on Wikidata, 3946 (96.8%) have been edited somehow by User:TiagoLubiana, and 2737 (67.2%) have been created by User:TiagoLubiana. Edits included adding multilanguage labels, connecting a dangling Wikipedia page to the cell subclass hierarchy, adding identifiers, images, markers, and other pieces of information. Approximatedly 430 hundred terms were added via manual curation based on PanglaoDB entries, while the remaining 2307 entries were created either via Wikidata's web interface or via the curation workflow described in this chapter. These statistics are demonstration of how the curation system efficiently contributes to the status of cell type information on Wikidata.

As mentioned by Aviv Regev in the Human Cell Atlas General Meeting 2021, "it's everyone's collective responsibility to participate in the annotation efforts, because that relies on domain expertise. To really tease apart things and give them names. Until we have names, people will have really a hard time working with things in biology."[**url?**]" We hope that by developing simplified curation tools we will engage more domain experts into the curation efforts.

Guidelines for reporting new cell types

Additional Work

Collaborations and manuscripts

fcoex

During the initial course of this PhD work, we also completed the development and reporting of *fcoex*, an R package for investigating cellular phenotypes using co-expression networks. [33] The software was maintained to withstand new releases of dependencies and new R version and was published as a preprint on biorxiv. [**doi:10.1101/2021.12.07.471603v1?**]

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 a Wikidata Bot to integrate information on protein complexes to Wikidata. An overview of the Wikidata integration is in Figure 4, 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). [34]
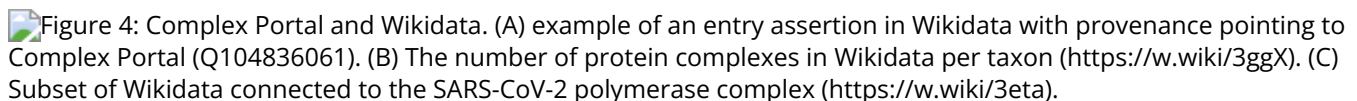
![Figure 4: Complex Portal and Wikidata. (A) example of an entry assertion in Wikidata with provenance pointing to Complex Portal (Q104836061). (B) The number of protein complexes in Wikidata per taxon (https://w.wiki/3ggX). (C) Subset of Wikidata connected to the SARS-CoV-2 polymerase complex (https://w.wiki/3eta).

Figure 4: Complex Portal and Wikidata. (A) example of an entry assertion in Wikidata with provenance pointing to Complex Portal (Q104836061). (B) The number of protein complexes in Wikidata per taxon (https://w.wiki/3ggX). (C) Subset of Wikidata connected to the SARS-CoV-2 polymerase complex (https://w.wiki/3eta).

I have also collaborated with the Cellosaurus database [35] to revive the CellosaurusBot [36], responsible for updating the metadata on more than 100,000 cell lines on Wikidata. The bot code, written in Python, was refactored entirely and runs semi-automatically after the Cellosaurus database was released. A write-up of the integration is in progress and is planned for release/submission in the second semester of 2022.

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. [37] 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. [38] It also entailed 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

- (Nov - 2021) Managed a project during BioHackathon Europe 2021, in Barcelona, Spain, on the representation of ELIXIR information on Wikidata. [39]
- (May - 2022) Presented the WikidataBib in a talk the 1st UK Local Biocuration conference, which got awarded as the Runner-Up Best Talk

Next steps

After this second year of work, the next steps of this PhD project will be geared towards robustifying the systems for biocuration and improving the quality of cell type information on Wikidata. This robustification will pave the way for the final phase of the project, where we plan to incorporate the organized information on Wikidata to standar workflows of single-cell RNA-seq data analysis.

In particular, the next part of the project includes the following steps:

- Develop Wikidata Bib into a mature system and deploy it as Open Source Software for the research community

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

  - Establish a comprehensive catalog of cell types on Wikidata
  - Develop a portal for access and re-use of the Wikidata database, e.g. providing datasets for enrichment analysis of differentially expressed genes
  - Complete the mapping between Wikidata and Cell Ontology, as a stepping stone for integrating Wikidata in existing frameworks

- Improve the quality of bibliometric data on Wikidata, laying the foundations to ensure complete coverage of the Brazilian contributions to the Human Cell Atlas

Chronogram

Awards and participation in events

- (Nov - 2021) Managed a project during BioHackathon Europe 2021, in Barcelona, Spain, on the representation of ELIXIR information on Wikidata. [39]
- (May - 2022) Presented the WikidataBib in a talk the 1st UK Local Biocuration conference, which got awarded as the Runner-Up Best Talk

Preprints and articles

Pre-prints

- Using coexpression to explore cell-type diversity with the fcoex package (https://www.biorxiv.org/content/10.1101/2021.12.07.471603v1)

- Guidelines for reporting cell types: the MIRACL standard https://arxiv.org/abs/2204.09673 ## Peer-reviewed articles

- Complex Portal 2022: new curation frontiers (https://doi.org:: 10.1093/nar/gkab991)

-

References

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

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

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

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

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

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

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

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

9. **Artificial intelligence in neurodegenerative disease research: use of IBM Watson to
    identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis.**
    Nadine Bakkar, Tina Kovalik, Ileana Lorenzini, Scott Spangler, Alix Lacoste, Kyle Sponaugle,
    Philip Ferrante, Elenee Argentinis, Rita Sattler, Robert Bowser
    Acta Neuropathologica (2017-11-13) https://www.wikidata.org/wiki/Q47406275
    DOI: 10.1007/s00401-017-1785-8

10. **Ontologies for the life sciences**

Steffen Schulze-Kremer, Barry Smith

(2005-11-15) https://www.wikidata.org/wiki/Q105870680

DOI: 10.1002/047001153x.g408213

11. **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

12. **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

13. **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

14. **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

15. **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

16. **Wikidata: A platform for data integration and dissemination for the life sciences and beyond**
Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M Schriml, Andrew I Su, Benjamin M Good
Bioinformatics (2015-11-16) https://doi.org/gg9dk4
DOI: 10.1101/031971

17. **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

18. **Wikidata as a knowledge graph for the life sciences**
Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, … Andrew I Su
eLife (2020-03-17) https://doi.org/ggqqc6
DOI: 10.7554/elife.52614 · PMID: 32180547 · PMCID: PMC7077981

19. **Wikidata: A large-scale collaborative ontological medical database**
Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, Helmi Hamdi
Journal of Biomedical Informatics (2019-11) https://doi.org/gg9dnt
DOI: 10.1016/j.jbi.2019.103292 · PMID: 31557529

20. **Comparison of reference management software - Wikipedia**
https://en.wikipedia.org/wiki/Comparison_of_reference_management_software

21. **Wikipedia, the free encyclopedia** https://en.wikipedia.org/wiki/Main_Page

22. **Clean Coder Blog** https://blog.cleancoder.com/uncle-bob/2017/12/18/Excuses.html

23. **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/

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

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

26. **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

27. **Wikidata:Tools/Author Disambiguator - Wikidata**
https://www.wikidata.org/wiki/Wikidata:Tools/Author_Disambiguator

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

29. https://query.wikidata.org/

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

31. **BioHackathon Europe** https://biohackathon-europe.org/

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

33. **fcoex: FCBF-based Co-Expression Networks for Single Cells**
Tiago Lubiana, Helder Nakaya
Bioconductor version: Release (3.15) (2022) https://bioconductor.org/packages/fcoex/

34. **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

35. **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](10.7171/jbt.18-2902-002)

36. **User:CellosaurusBot - Wikidata** https://www.wikidata.org/wiki/User:CellosaurusBot

37. **Wisecube AI | Knowledge Graph Engine** https://www.wisecube.ai/

38. **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](10.1186/s12859-015-0564-6)

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