Manuscript Title

This manuscript (permalink) was automatically generated from lubianat/phd_thesis@53cfb87 on May 11, 2021.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Index

Master index

Numbers are used as tags for literature and should not be changed!

At best, names can be clarified.

Ex: 1.1.2 refers to “Interoperable data: dataset integration”

Order may be changed at a later step.

-
- 1.3.2. Biological knowledgebases
- 1.3.2.1. Cell-type-oriented knowledgebases
- 1.3.2.2. Other databases relevant for this work
- 1.3.2.3. Bio2RDF and semantic databases
-
- 1.3.3. Wikidata as a platform for representation of biological knowledge
-
- 1.3.4 Semantic Systems Biology
- 1.4. The challenges of the Human Cell Atlas
- 1.4.1. The Human Cell Atlas project and its scope
- 1.4.1.1. Participants
- 1.4.1.2. Overview of main analytical techniques
-
- 1.4.2. A focus on single-cell RNA sequencing
- 1.4.2.1. Wet-lab methods and their differences
- 1.4.2.1.1 Bias introduced by different scRNA-seq methods
- 1.4.2.2. Computational analysis of scRNA-seq data
- 1.4.2.2.1. Clustering algorithms
- 1.4.2.3. Cell label identification
- 1.4.2.3.1. Labelling clusters
- 1.4.2.3.2. Labelling events
-
- 1.4.3. Data availability
- 1.4.3.1. As coordinated by the Human Cell Atlas
- 1.4.3.2. By the community as a whole

- 2.4.2.2. Identification
- 2.4.2.2.1. Transgenic animals and marker-based identification
- 2.4.2.2.2. A posteriori identification in high-throughput metodologies
- 2.4.2.3. Nomenclature
- 2.4.3. States, identities, fates, attractors and phenotypical continuity
- 2.4.4. A pragmatic definition of cell type
- 2.4.4.1. 3+1 rules for defining one type
- 2.4.4.1.1 Cell types must have explicit definitions
- 2.4.4.1.2 Cell types must have a taxonomic scope
- 2.4.4.1.3 Cell types must be useful
- 2.4.4.1.4 Cell types should be inserted in an ontology
- 2.4.4.2. Names for different classes of types
- 2.4.4.2.1 The cell archetype
- 2.4.4.2.2 The sensu stricto cell type
- 2.4.4.2.3 The infratype
- 2.4.4.2.4 The technotype
- 2.4.4.3. Logical implications of the definition
- 2.4.4.4. Practical implications of the definition
- 2.4.4.5. Less-strict, literature based definitions for immediate use
- 2.4.5.1. The use of the concept in the literature
- 2.4.5.2. The use of the concept in metadata and knowledgebases
- 2.4.5.3. The use of the concept in ontologies
- 2.4.5.4. Wikidata and consensus-determined rigorously undefined concepts
- 2.4.6. Claims of novel cell types
- 2.4.7. Levels of classification of cells
- 2.4.7.1 Differentia for cell classes: biological sex

- 2.4.7.2 Differentia for cell classes: strains and subspecific groups
- 2.4.7.3 Differentia for cell classes: age and life stage
- 2.4.7.4 Differentia for cell classes: circadian clock
- 2.4.7.5 Differentia for cell classes: presence in organ
- 2.4.7.5 Differentia for cell classes: sub-organ regions
- 2.4.7.6 Differentia for cell classes: disease condition
- 2.5. The concept of “protein”
- 2.5.1. Proteins, proteins families and protein molecules
- 2.5.2. Protein complexes and their details
- 2.5.3. Protein forms
- 2.6. The concept of “transcript”
- 2.6.1. Transcript expression x gene expression x protein expression
- 2.7. The concept of “cell marker”
- 3. Practical Projects
- 3.1. Cell-type markers in Wikidata
- 3.1.1. PanglaoDB
- 3.1.2. Other bases of markers
- 3.1.3. WikidataMarkers: a website and an R packase for cell type gene sets
- 3.1.4. Cell-Disease networks via Wikidata
- 3.2. Community annotation of texts via Wikidata
- 3.2.1. Pilot: Annotation of the Human Cell Atlas corpus
- 3.2.2. Community curation and gamified science
- 3.2.2.1. ANN
- 3.2.3. Text mining and Wikidata community curation
- 3.3. Cellosaurus cell-lines to Wikidata
- 3.4. Single-cell RNA-seq data reconciliation to Wikidata

- 3.4.1. Integration of single-cell RNA-seq bibliometric data to Scholia
- 3.4.2. scRNA-seq metadata reconciliation
- 3.4.2.1. COVID-19 cell types collection
- 3.4.2.2. Other single-cell RNA-seq datasets
- 3.4.2.3. Cell-disease networks inferred from Wikidata
- 3.5. Practical applications of a pragmatic cell type definition in 2.4.4
- 3.5.1. Basic dictionary for cell type identification based on regular expression of cell type markers
- 3.5.2. R package for suggestion of cell-type rules based on Seurat clusters
- 3.5.3. R package for rule-based identification of cell types
- 3.5.3.1. Data model and development
- 3.5.3.2. Large scale application to many datasets
- 3.6. Wikidata Bib and scientific practice as a first-class citizen in the knowledge world
- 3.6.1. The Wikidata Bib platform and the formalization of a reading discipline
- 3.6.2. Linked Open Data for personalized analysis of bibliography
- 3.6.3 Sowing the seeds of knowledge: Meta contributions to Wikidata during the PhD
- 3.7 Scholarly and metascientific information on Scholia/Wikidata
- 3.8 The long tail of biocuration
- 3.9 Semantic Systems and Synthetic Biology
- 3.10 Minimal Information About New Cell Classes
- 3.11 The Cell Wiki Project
- 3.12 Linked GEO and curation of transcriptomics datasets
- 0 Master's projects
- 0.1 fcoex & FCBF
- 0.2 Alexa
- 0.3 PubScore

Other tags

- X.1 - Phrase to quote

Abstract

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

Introduction

The Human Cell Atlas (HCA) Project

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

Building a full atlas of human cells comes with multiple challenges. The project includes detection, in single cells, of RNA content (scRNA-Seq), chromatin accessibility (scATAC-Seq), and protein markers (primarily by CYTOF), as well as spatial information on cells with multiplexed *in situ* hybridization (such as MERFISH) and imaging mass cytometry [2,3]. Every lab will contribute with its expertise, providing samples that are representative of human diversity.

HCA is set to revolutionize the biomedical sciences, by creating tools and standards for basic research, as well as allowing better characterization of disease, and thus, ultimately, improving diagnostics and therapy. 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 [4] to specialized personnel [4].

The Human Cell Atlas has a dedicated team for organizing data: the Data Coordination Platform (DCP) [5] [3]. The DCP is responsible for tracing the plan for computational interoperability, from the data generators to the consumers.[3]. 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 highlight of “knowledge” in the last paragraph is meant to stress that data *per se* is not enough. 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 Commitee 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. ## Literature Based Discovery, hidden knowledge and text-mining It is not recent news that the amount of scholar information vastly outnumbers what single researchers can fathom. Nevertheless, the gap between single individuals and the collectively body of knowledge has been widening in an accelerated fashion. The explosion in the number of published articles is leading to a “tsunami of knowlegde”, flooding the scientific literature with rich information. Moreover, articles themselves are becoming denser, as high-throughput (and high-information) technologies like single-cell RNA-sequencing get cheaper and widely used.

The technological advances, however, are no yet met by equivalent knowledge-handling systems. Mainstream scientific publication is, nowadays, barely readable by machines. Articles are written for human consumption, using ambiguous natural language and relying on implicit conventions. Tables and data rarely make use of URI (Uniform Resource Identifiers), RDF (Resource Description Framework) formating and other W3C (World Wide Web Standards). In fact, those standards and their acronyms are completely foreign for most life scientists (personal observations), despite being the *de jure* gold standard for data quality. [6] Thus, interconnecting biomedical knowledge is an open challenge of our century, and there is a large way to go before society can fully benefit from the sum of all knowledge we generate.

The scientific community has pursue solutions for this tsunami of information from many different angles. Narrative reviews, systematic reviews and textbooks compile and synthetize information, providing a layer of processing. Biocuration efforts go a step further and transform unstructured information into structered information in knowledgebases, such as UniProt and PDB. Text-mining apply a range of Natural Language Processing tools to try and extract biological relations, or provide guidance for biocurators. Elaborate knowlegde networks, like the STRING database [7] and Wikidata[8], combine information from different sources. Overall, approaches mix and intermingle to mature hidden knowledge into solid theories and practical applications.

The synthesis effort of literature mining goes beyond purely detecting what science claims to be true. Interconnected knowledge provides a way to discover new, implicit knowledge, by applying logical reasoning to a dataset. A field denominated Literature Based Discovery [9] dedicates itself to this challenge: make actual discoveries (or at least very strong hypothesis) using as material plainly the existing literature. [10] 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 [11] 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 [11]

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. [12] 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. [13]

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.

For the Human Cell Atlas Project (as presented in the chapter) to maximize its benefit for society, its knowledge products will need to be inserted into the main route of automated knowledge discovery . That implies a daunting task of building knowledge graphs able to deal with it at all layers, including the generated data and metadata, its range of different protocols, and the purified knowledge projects that are enshrined in publications. Thus, the chapter will present challenges and paths for applying literature based discovery on an enormous scale and with sufficient flexibility to deal with the Human Cell Atlas.

Knowledge graphs as tools for interoperability

- The OBO Foundry and biomedical ontologies
- Knowledge graphs and a different approach to biomedical semantics
- Wikidata as a knowledge graph for the life sciences

Objectives

- Set up the semantic infrastructure on Wikidata for handling knowledge about cell types
    - Refine the theories of types/states/classes of cells within the constraints of ontologies and knowledge bases
    - Investigate the types of statements done about cell types
        - On Wikidata
        - On OBO Foundry ontologies
        - Freely on the biomedical literature
    - Craft wikidata relations ("properties") for making cell-type-related assertions (like "has marker" or "is the progenitor of")
- Devise ways to connect the Human Cell Atlas products to Wikidata and the Linked Open Data cloud

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

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

Methodology

Organized reading

Given the breadth of the task envisioned for this project, a standard methodology of reading was followed.

- Describe Wikidata bib
- Integrate with ECO's views

Wikidata updates

- Property proposals
- Wikidata bots
- PanglaoDB integration
- Semi-manual integration: Google Sheets and Quickstatements

Data retrieval

- SPARQL queries

Data analysis

- Packages used in R and Python
- For interacting with Wikidata

Annotation of Human Cell Atlas articles

Status of cell type info on Wikidata

Cell-disease network analysis

Preliminary Results

The concept of cell type

- Describe background
- Cell types, cell states and cell classes
- Levels of cell type information: archetype, senso stritu cell type, infratype and technotype.
- Infratypes and technotypes as theoretical innovations
- Current usage mixes archetypes and species-specific cell types

- Annotation of HCA articles for grasping the use of different concepts in the context of HCA

Next steps

- Improve formalization of cell types in connection with the biomedical semantics community

HCA

- “Sky dive” approach: hand annotation of all abstracts and the core Human Cell Atlas paper
- Benefits of using a single ontology that anyone can edit (new terms and speed of science)
- Figure: The different concepts in use by the HCA paper
- Figure: The different concepts in use by the different HCA papers
- Discussion
- Information by HCA and related efforts is already targeted by biocurators. PanglaoDB is one of these resources etc etc

Next steps

- Mature the annotation system into a curation tool (based on ANN, perhaps reuse figure)
- Explore the use of SciSpacy and natural language processing for making it easier

PanglaoDB integration to Wikidata

- The architecture of marker information on Wikidata
- Integration of information to the larger scope –> live updates by everyone
- Overview of the stats

Cell-disease networks

- Systems-biology explorations: what can we discover based on the literature distilled on wikidata?
- Cell-disease networks based on shared genes
- Hub diseases and cell types
- ShinyApp to explore the data in real time

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. **The Human Cell Atlas White Paper**
   Aviv Regev, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael J. T. 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

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

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

6. **5-star Open Data** http://5stardata.info/en/

7. **The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets**
   Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda Tsankova Doncheva, Marc Legeay, Tao Fang, Peer Bork, … Christian von Mering
   Nucleic Acids Research (2020-11-25) https://www.wikidata.org/wiki/Q102383784
   DOI: 10.1093/nar/gkaa1074

8. **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

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

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

11. **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

12. **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

13. **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