Manuscript Title

This manuscript (permalink) was automatically generated from lubianat/phd_thesis@e71b880 on April 7, 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. Introduction
- 1.1. The quest for interoperable knowledge
- 1.1.1. Literature Based Discovery, hidden knowledge and text-mining
- 1.1.2. Interoperable data: dataset integration
- 1.1.3. Interoperable publication processes: nanopublications
- 1.2. Formal representation of knowledge
- 1.2.1. Descriptional logic and its historical context
- 1.2.2. Computational ontologies and their methods
- 1.2.3. Web of Data and Linked Open Data
- 1.2.4. Wikidata and Knowledge Graphs
- 1.2.5. The role of definitions in formal knowledge
- 1.3. Knowledge Representation in biology
- 1.3.1. OBO Foundry and biomedical ontologies
- 1.3.1.1 Gene Ontology
- 1.3.1.2. UBERON
- 1.3.1.3. MONDO and DO
- 1.3.1.4. Cell Ontology (and CELDA)
- 1.3.1.5. UMLS, NCiT and non-OBO resources
-

- 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
-

- 2.4.2.2.1. 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
- 2.5. 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
- 3.2. 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
- 3.5. 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 Abount New Cell Classes
- 0 Master's projects
- 0.1 fcoex & FCBF
- 0.2 Alexa
- 0.3 PubScore

Introduction

The Human Cell Atlas

- Data interoperability challenges
- Data interoperability challenges
- knowledge interoperability challenges

Some of the interoperability challenges fit within the larger quest to extract and integrate literature knowledge. Biocuration and literature based discovery.

Literature Based Discovery, hidden knowledge and text-mining

- Tsunami of knowledge - o parse, connect and benefit society
- Literature based discovery is a way to connect knowledge
  - Biocuration and organization
  - Actual processing of the curated information

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