Manuscript Title

This manuscript (<u>permalink</u>) was automatically generated from <u>lubianat/phd_thesis@cca7252</u> on March 16, 2021.

Authors

- John Doe

Department of Something, University of Whatever \cdot Funded by Grant XXXXXXXX

- Jane Roe

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
- 1.4.2.2. 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

•

1.4.4. Types of knowledge about cell types

- •
- 1.4.5. Goals of this project
- 2. Definitions
- 2.1. The concept of "gene"
- 2.1.1. Gene as phenotype x gene as DNA
- 2.1.2. Species-specific genes and multispecies genes
- 2.1.3. Representations of genes in knowledge bases
- 2.1.4. A simple theory of the molecular gene
- 2.1.4.1. Allele x gene
- 2.1.4.2. Gene as individual and gene as a class
- 2.2. The concept of "taxon"
- 2.2.1. Lines of thought about the taxon concept
- 2.2.1.1. Cladistics and the PhyloCode
- 2.2.1.2. Phenetics and Numeric Taxonomy
- 2.2.2. Bacteria, cell lines and the plurality of the concept of taxon
- 2.3. The concept of "cell"
- 2.3.1. Historical perspective
- 2.3.2. Boundaries of the concept in the context of this thesis
- •
- 2.4 The concept of "cell type"
- 2.4.1. Historical perspective and the morphological cell type
- 2.4.2. Modern perspectives of classification of cells
- 2.4.2.1. Classification
- 2.4.2.1.1 Evolutionary cell types
- 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
- 2.5. 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 trail of biocuration
- 0 Master's projects
- 0.1 fcoex & FCBF
- 0.2 Alexa
- 0.3 PubScore

This manuscript is a template (aka "rootstock") for <u>Manubot</u>, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (.md files in the /content directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined italics and bold

Strikethrough

- 1. Ordered list item
- 2. Ordered list item
 - a. Sub-item
 - b. Sub-item
 - i. Sub-sub-item
- 3. Ordered list item
 - a. Sub-item
- List item
- List item
- · List item

subscript: H₂O is a liquid

superscript: 2¹⁰ is 1024.

unicode superscripts 0123456789

unicode subscripts 0123456789

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to <u>editing</u> and <u>version</u> control.

Line break without starting a new paragraph by putting two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6



Horizontal rule:

Heading 1's are recommended to be reserved for the title of the manuscript.

Heading 2's are recommended for broad sections such as Abstract, Methods, Conclusion, etc.

Heading 3's and Heading 4's are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

<u>Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah</u>

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

```
Figure 3

Figure 4

Table 1

Equation 1

Equation 2
```

Quotes and code

Quoted text

Quoted block of text

Two roads diverged in a wood, and I—I took the one less traveled by, And that has made all the difference.

Code in the middle of normal text, aka inline code.

Code block with Python syntax highlighting:

```
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures



Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.



Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.



Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.



Figure 4: A vector .svg image loaded from GitHub. The parameter sanitize=true is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	Jane	John	Alice	Bob
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	Digits 34-66	Digits 67-99	Ref.
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	piday.org
е	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	nasa.gov

 Table 3: A table with merged cells using the attributes plugin.

	Colors		
Size	Text Color	Background Color	
big	blue	orange	
small	black	white	

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$$
(2)

Special

▲ WARNING The following features are only supported and intended for .html and .pdf exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as .docx.

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot attributes plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

Manubot Manubo

Available background colors for text, images, code, banners, etc:

white lightgrey grey darkgrey black lightred lightyellow lightgreen lightblue lightpurple red orange yellow green blue purple

Using the Font Awesome icon set:



Light Grey Banner useful for *general information* - <u>manubot.org</u>

1 Blue Banner

useful for important information - manubot.org

♦ Light Red Banner useful for *warnings* - <u>manubot.org</u>

References

1. Sci-Hub provides access to nearly all scholarly literature

Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene

eLife (2018-03-01) https://doi.org/ckcj

DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. Reproducibility of computational workflows is automated using continuous analysis

Brett K Beaulieu-Jones, Casey S Greene

Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/

DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. Bitcoin for the biological literature.

Douglas Heaven

Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888

DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. Plan S: Accelerating the transition to full and immediate Open Access to scientific publications

cOAlition S

(2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. Open access

Peter Suber *MIT Press* (2012)

ISBN: 9780262517638

6. Open collaborative writing with Manubot

Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter

Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. Opportunities and obstacles for deep learning in biology and medicine

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, ... Casey S. Greene

Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn

DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. Open collaborative writing with Manubot

Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) https://doi.org/c7np

DOI: <u>10.1371/journal.pcbi.1007128</u> · PMID: <u>31233491</u> · PMCID: <u>PMC6611653</u>