Toward a minimal information reporting standard about new cell types

Tiago Lubiana¹,², Paola Roncaglia³, Christopher J. Mungall⁴, Ellen M. Quardokus⁵, Joshua D. Fortriede⁶, David Osumi-Sutherland³ and Alexander D. Diehl⁷

¹ School of Pharmaceutical Sciences, University of São Paulo, São Paulo, SP, Brazil
² Ronin Institute for Independent Scholarship
³ European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
⁴ Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720 USA
⁵ Department of Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana, USA
⁶ Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45224 USA
⁷ Department of Biomedical Informatics, University at Buffalo Jacobs School of Medicine and Biomedical Sciences, Buffalo, NY, 14203 USA

Abstract

The classification of cell diversity using ontologies is a core step for managing biomedical data. While technical tools to represent knowledge about cell types are available, there is still a gap between classes mentioned in biomedical literature and what has been cataloged by the Cell Ontology. Here we introduce the Minimal Information Reporting About a Cell (MIRACL) standard to bridge the gap between natural language reports and formal ontologies. The standard currently outlines four free-text fields (label, diagnostic description, reference, and additional information) and three formal fields (scientific taxon, UBERON anatomical structure, and Cell Ontology superclass). The standard is open to contributors.

Keywords

Cell type, Cell Ontology, Minimum Information Standard, knowledge management, cell atlas

1. Introduction

Cell-type-oriented research has gained traction thanks to the recent blooming of projects like the Human Cell Atlas (HCA) (Rozenblatt-Rosen et al., 2017), the Human Protein Atlas (HPA) (Karlsson et al., 2021), the Human BioMolecular Atlas Program (HuBMAP) (Consortium and HuBMAP Consortium, 2019) and more (Ando et al., 2020). These efforts encompass diverse domains, from morphology to biophysics to molecular biology, and data integration requires advanced knowledge management. To tackle the organization challenge, programs like the Cell Annotation Platform (Osumi-Sutherland et al., 2021) and the CCF ASCT+B Reporter (Börner et al., 2021) are contributing and reusing data annotation from the Cell Ontology (Diehl et al., 2016) to standardize references to cell types.

¹EMAIL: tiago.lubiana.alves@usp.br (A. 1); paola.roncaglia.phd@gmail.com (A. 2); cjmungall@lbl.gov (A. 3); ellenmq@indiana.edu (A. 4); davidos@ebi.ac.uk (A.5) Joshua.Fortriede@cchmc.org (A.6) addiehl@buffalo.edu (A.7)
ORCID: 0000-0003-2473-2313 (A. 1); 0000-0002-2825-0621 (A. 2); 0000-0002-6601-2165 (A. 3); 0000-0001-7655-4833 (A. 4); 0000-0002-3293-5463 (A. 5); 0000-0002-7073-9172 (A. 6); 0000-0001-9990-8331(A. 7)
Despite the importance of curation, the community has yet to agree on common standards to report cell types in the scientific literature. Reporting standards are beneficial for the community, as they help the efforts of organizing life sciences knowledge. Minimum information standards are widely used for assays and data reporting, and submission of standardized data to public databases is a prerequisite for publication for e.g. DNA and protein sequences and protein structures. A similar approach may benefit the curation of cell types and states.

In this perspective, we introduce a discussion towards building a Minimal Information Reporting About a CelL (MIRACL) standard focusing on the curation provided by the Cell Ontology. We outline a set of core fields for cataloging new types and provide a schema and template for orienting reports of new cell classes.

2. Who can report a MIRACL sheet?

Cell types are currently described in various research contexts, few of which are solely descriptive, microanatomy works. Any research involving a cell type not cataloged in the Cell Ontology would benefit from providing a MIRACL sheet. We outline three different situations in which a MIRACL sheet would be most valuable, with examples:

- Claims of new cell types/subtypes (Villani et al., 2017) (Joseph et al., 2021; Stecco et al., 2018). Sometimes a close superclass is cataloged, but there is value in reporting specific subclasses, as researchers might want to represent different degrees of granularity.
- New information that might change the classification of a cell type, for example the identification of a cell type in a different region. (Elmentaite et al., n.d.)
- Mentions of cell types identified in other works that are not cataloged in the Cell Ontology. (Bigaeva et al., 2020) Scenarios may include proposing a synonym for a previously reported cell type. (Popescu and Faussone-Pellegrini, 2010)

While anyone is welcome to request a new term in the Cell Ontology GitHub Tracker using standardized term templates (https://github.com/obophenotype/cell-ontology/issues), ideally, we would like to engage the authors in the process, to minimize misinterpretation of data, and reduce overhead and time required for inclusion in the Cell Ontology. If authors of research articles shared this information in a standardized fashion, the knowledge produced could be readily integrated into the system (depicted in Figure 1) and magnify the impact of individual works.

![Figure 1: The flow from research to reuse, with an author-guided structuring of information catalyzing the biocuration process. Currently, the curation process requires Cell Ontology editors to sift through full texts to identify the core information needed to catalog cells. A MIRACL sheet would structure information at the publication level, thereby leveraging the expertise of the original authors. That, in](image-url)
turn, facilitates the process of curation and, subsequently, the reuse of new terms, for example, to label clusters in single-cell omics data.

3. What might a MIRACL sheet contain?

Cells are notoriously difficult to classify, as there is not a single biological feature that guides taxonomists. From simple colorimetric stains, to multiplexed immunohistochemistry, to patch clamps, to flow cytometry, to the multiple omics now performed at the single-cell level, scientists employ fundamentally different techniques to guide their perspectives on the cellular world. Any universal standard for cell type reporting faces the challenge of finding the commonalities across the domains of research, while preserving the power to represent the details of each project.

A MIRACL sheet is aimed at capturing the very basic information for descriptions of new cell types. As a middle stage between full ontological axiomatization and free text, the selected fields must facilitate the work of ontology editors without requiring too much technical training of contributors. We reason that 7 different fields of information should be available for any new description of a cell type:

- A label. An unambiguous name for the cell type being described. While listing synonyms is useful, a pragmatic decision on a label is necessary when building an ontology.
- A diagnostic description. A concise, free-text description on the core characteristics of the class, similar to the diagnosis section in species taxonomy. (Winston, 1999) It should suffice to distinguish the cell type from similarly existing types (e.g. the specific marker genes).
- A superclass. A Cell Ontology class, with name and identifier, corresponding to a broader category of the new cell type. For example, a report for a new type of neuron might include “neuron (CL:0000540)” as a superclass.
- A taxon. The scientific name for the taxon for which cells of the type are expected to be found. In most cases, it will be the name of the species in which the cells were found, but this depends on the generality claim being made.
- An anatomical structure. The UBERON (Mungall et al., 2012)ontology identifier for the anatomical structure(s) in which the cells of the type were found, usually a particular organ or tissue. If cells of the type are known to span multiple structures (e.g. a long neuron), or are believed to be present in other anatomical locations, details should be noted in the additional information field. Note that some cell types, especially immune ones, may not be tied to a single anatomical structure.
- A reference. A free-text field pointing to the reference(s) that support the new cell type.
- Additional information. Any additional core information about the cell or the classification that complements the previous assertions, as well as details and technicalities that might aid curation.

In Table 1, we present an example of what a MIRACL sheet would look like in the case of two different publications (Villani et al., 2017) based on the original texts and on searching the EMBL-EBI Ontology Lookup Service for CL and UBERON terms. (Jupp et al., 2015) In a report, we propose that the information is provided in an independent tab-separated value (TSV) spreadsheet as supplementary information.
### Table 1
An example of a MIRACL sheet

| Label                        | Diagnostic description                                                                 | Superclass                | Taxon         | Anatomical structure       | Reference                          | Additional information |
|------------------------------|----------------------------------------------------------------------------------------|---------------------------|---------------|---------------------------|------------------------------------|------------------------|
| AXL* SIGLEC6* dendritic cell | A dendritic cell that expresses AXL and SIGLEC6.                                        | dendritic cell (CL:0001056) | Homo sapiens  | blood (UBERON:000178)     | This article*                      | -                      |
| Lamina propria fibroblast    | A fibroblast in the lamina propria mucosa.                                              | fibroblast (CL:0000057)   | Homo sapiens  | lamina propria (UBERON:000030) | This article*, PMID: 31730855, 31474370, 30270042, 31348891, DOI:10.1101/2020.01.10.901579 | Articles used different names to refer to the class. |

* If this sheet was provided alongside the original publications (Villani et al., 2017)(Bigaeva et al., 2020)

### MIRACL Extensions: Profiles

While MIRACL’s fields are designed to be simple, there are guidelines that complement MIRACL (like the Petilla convention (Petilla Interneuron Nomenclature Group et al., 2008), Allen Institute’s cell type nomenclature (Miller et al., 2020), and Human Immunology Project Consortium’s tentative standard (Overton et al., 2019)) that might provide additional support for particular situations. Additionally, the Cell Ontology and other OBO Foundry ontologies have tools for formal representation of other aspects of cells. To formalize, extend and make the proposal computer processable, we have started a collaborative LinkML schema (https://lubianat.github.io/miracl/).

The current minimal schema allows for extensions in the form of checklists. For example, checklists could be created for broad categories of cell type such as stem cell, neuron, immune cell, extending the MIRACL sheet with fields that are applicable for describing cell types in a particular community. Additionally, the standard could be optionally combined with checklists particular to experimental modalities, e.g. marker genes for single-cell RNAseq data.

Extensions could also adapt MIRACL to species more distant from mammals. Single-cell technology provides an opportunity to better understand plants, fungi, and bacteria. (Cole et al., 2021) MIRACL could be used with adaptations for these taxa. For example, for submission of plant cell types to Plant Ontology (PO), the anatomical structure field would require a plant structure represented in PO, rather than an Uberon term.
4. Pilot test of MIRACL sheets with researchers

To gauge community response to MIRACL standards, a pilot test was performed. The volunteers were researchers who wished to request new ontology terms for anatomical structures or cell types. Those terms were determined to be missing from either the UBERON ontology or Cell Ontology as part of their work for HuBMAP in building tables of anatomical structures, cell types and biomarkers (ASCT+B). (Börner et al., 2021)

To minimize barriers for submitting information conforming to MIRACL standards, MIRACL sheets were provided as Google Sheets templates to authors of ASCT+B tables for thymus, spleen and lymph node (v1.1, see Acknowledgements for details). Most wet bench researchers were uncomfortable engaging with the GitHub environment, where requests for anatomy and cell type terms are usually made, and the MIRACL sheets provided a simple way to obtain and communicate information directly between the experts and ontology curators. The MIRACL tables curated by ASCT+B authors are provided in the Supplementary Data as examples.

Beyond the basic MIRACL information, HuBMAP ASCT+B authors have added 2 additional fields: “Gene Biomarkers” and “Protein Biomarkers”. Such extensions are welcome, as long as the core MIRACL fields are present. Additionally, the table was adapted for curating new anatomical structures for UBERON, highlighting the usefulness of the scaffold for different ontologies.

We noticed that the most difficult ontology concept for most researchers was superclass, while all other fields of data in the MIRACL sheet were easily provided. The superclasses provided by the authors were further curated by ontology editors, and the final ontology placement may be different. Thus, if navigating the ontology proves time-consuming, authors might provide broad MIRACL superclasses (e.g. “neuron” or “lymphocyte”), as these already serve as pointers to ontology editors.

5. Discussion and conclusion

This work presents the initial step for a standard to report new cell types. By providing well-crafted standards, researchers would streamline access of readers to an article's main discoveries while helping to curate entities in the Cell Ontology. The curation, in turn, empowers future data interoperability, as other researchers working on the same cell types may use the terms to annotate their data accurately. A MIRACL sheet thus may catalyze the curation process, facilitating the attribution of credit for new cell type descriptions and extending the reach of individual works. Importantly, the MIRACL format has already proven to remove an involvement barrier for some researchers. In the authors’ experience, scientists unfamiliar with GitHub repositories, and/or under tighter time constraints, are often reluctant to engage with the GitHub issue tracking system, but are happy to provide information in a MIRACL template to create or modify existing terms. This standard therefore can maximize output and minimize time needed for inclusion in the Cell Ontology - a gain for the entire community.

Even though the benefits for standardization are solid, such a standard will only be valuable if it is adopted by the community. Similar to the publishing of accession numbers for nucleic acid sequences and genotypic information for cell lines, MIRACL's success depends on being adopted by journals as a requirement for reporting new cell types, and that in turn depends on a community agreement on standard operating procedures. If researchers submitted MIRACL sheets to the Cell Ontology before journal or archive submission, or during review, the cell types described in the manuscript would be curated and
given unique identifiers that the authors (and the community) could use promptly - similarly to how e.g. accession numbers for DNA sequences can be used even before publication.

In that direction, we invite the reader to contact us (https://obophenotype.github.io/cell-ontology/contact_us/), share opinions, and contribute toward a Minimum Information About a New Cell Type standard. A blank MIRACL template in tabular format is provided in this Google Sheet.

6. Acknowledgments

We’d like to thank all colleagues contributing to the Cell Ontology for valuable discussions leading to this article. We also want to thank the ASCT+B authors that tested the MIRACL sheet prototype and provided feedback on the usability. (Brusko et al., 2021)(Jorgensen et al., 2021a)(Jorgensen et al., 2021b)

7. Competing interests

The authors declare no competing interests.

8. Funding

This work has been made possible in part by a grant from CZI (Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation). TL is supported by a grant from the São Paulo Research Foundation (#19/26284-1). PR and DOS were supported by the Chan-Zuckerberg Initiative award for the Human Cell Atlas Data Coordination Platform to EMBL-EBI. JF is supported by an NHLBI/NIH award for the LungMap Data Coordinating Center (U24HL148865).

9. References

Ando Y, Kwon AT-J, Shin JW. 2020. An era of single-cell genomics consortia. Exp Mol Med 52:1409–1418.
Bigaeva E, Uniken Venema WTC, Weersma RK, Festen EAM. 2020. Understanding human gut diseases at single-cell resolution. Hum Mol Genet 29:R51–R58.
Börner K, Teichmann SA, Quardokus EM, Gee JC, Browne K, Osumi-Sutherland D, Herr BW 2nd, Bueckle A, Paul H, Haniffla M, Jardine L, Bernard A, Ding S-L, Miller JA, Lin S, Halushka MK, Boppana A, Longacre TA, Hickey J, Lin Y, Valerius MT, He Y, Pryhuber G, Sun X, Jorgensen M, Radtke AJ, Wasserfall C, Ginty F, Ho J, Sunshine J, Beuschel RT, Brusko M, Lee S, Malhotra R, Jain S, Weber G. 2021. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. Nat Cell Biol 23:1117–1128.
Brusko M, Beuschel RT, Radtke AJ. 2021. HuBMAP ASCT+B Tables. Thymus v1.1. doi:10.48539/HBM392.LPKF.942
Cole B, Bergmann D, Blaby-Haas CE, Blaby IK, Bouchard KE, Brady SM, Ciobanu D, Coleman-Derr D, Leiboff S, Mortimer JC, Nobori T, Rhee SY, Schmutz J, Simmons BA, Singh AK, Sinha N, Vogel JP, O’Malley RC, Visel A, Dickel DE. 2021. Plant single-cell solutions for energy and the environment. Communications Biology 4:1–12.
Consortium H, HuBMAP Consortium. 2019. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. doi:10.1038/s41586-019-1629-x
Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y,
Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, Van Slyke CE, Vasilevsky NA, Haendel MA, Blake JA, Mungall CJ. 2016. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* **7**:1–10.

Elmentaite R, Kumasaka N, King HW, Roberts K, Dabrowska M, Pritchard S, Bolt L, Vieira SF, Mamanova L, Huang N, Goh Kai’En I, Stephenson E, Engelbert J, Botting RA, Fleming A, Dann E, Lisgo SN, Katan M, Leonard S, Oliver TRW, Hook CE, Nayak K, Perrone F, Campos LS, Domingue-Conde C, Polanski K, Van Dongen S, Patel M, Morgan MD, Marioni JC, Bayraktar OA, Meyer KB, Zilbauer M, Uhlig H, Clatworthy MR, Mahbubani KT, Saeb Parsy K, Haniffa M, James KR, Teichmann SA. n.d. Cells of the human intestinal tract mapped across space and time. doi: 10.1101/2021.04.07.438755

Jorgensen M, Radtke AJ, Beuschel RT. 2021a. HuBMAP ASCT+B Tables. Spleen v1.1. doi: 10.48539/HBM582.WJWX.929

Jorgensen M, Radtke AJ, Rodriguez N. 2021b. HuBMAP ASCT+B Tables. Lymph Node v1.1. doi: 10.48539/HBM573.SHCQ.259

Joseph DB, Henry GH, Malewska A, Reese JC, Mauck RJ, Gahan JC, Hutchinson RC, Malladi VS, Roehrborn CG, Vezina CM, Strand DW. 2021. Single-cell analysis of mouse and human prostate reveals novel fibroblasts with specialized distribution and microenvironment interactions. *J Pathol*. doi: 10.1002/path.5751

Jupp S, Burdett T, Leroy C, Parkinson HE. 2015. A new Ontology Lookup Service at EMBL-EBI. *SWAT4LS* **2**:118–119.

Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, Sjöstedt E, Butler L, Odeberg J, Dusart P, Edfors F, Oksvold P, von Feilitzen K, Zawahlen M, Arif M, Altay O, Li X, Ozcan M, Mardonoglu A, Fagerberg L, Mulder J, Luo Y, Ponten F, Uhlén M, Lindskog C. 2021. A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**. doi: 10.1126/sciadv.abh2169

Miller JA, Gouwens NW, Tasic B, Collman F, van Velthoven CT, Bakken TE, Hawrylycz MJ, Zeng H, Lein ES, Bernard A. 2020. Common cell type nomenclature for the mammalian brain. *Elife* **9**. doi: 10.7554/elife.59928

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**:1–20.

Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, Lein E, Teichmann SA. 2021. Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol* **23**:1129–1135.

Overton JA, Vita R, Dunn P, Burel JG, Bukhari SAC, Cheung K-H, Kleinstein SH, Diehl AD, Peters B. 2019. Reporting and connecting cell type names and gating definitions through ontologies. *BMC Bioinformatics* **20**:182.

Petilla Interneuron Nomenclature Group, Ascoli GA, Alonso-Nanclares L, Anderson SA, Barrionuevo G, Benavides-Piccione R, Burkhalter A, Buzsáki G, Cauli B, Defelipe J, Fairén A, Feldmeyer D, Fishell G, Fregnac Y, Freund TF, Gardner D, Gardner EP, Goldberg JH, Helmstaedter M, Hestrin S, Karube F, Kissvárday ZF, Lambolez B, Lewis DA, Marin O, Markram H, Muñoz A, Packer A, Petersen CCH, Rockland KS, Rossier J, Rudy B, Somogyi P, Staiger JF, Tamas G, Thomson AM, Toledo-Rodriguez M, Wang Y, West DC, Yuste R. 2008. Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nat Rev Neurosci* **9**:557–568.

Popescu LM, Faussone-Pellegrini M-S. 2010. TELOCYTES - a case of serendipity: the winding way from Intersitial Cells of Cajal (ICC), via Interstitial Cajal-Like Cells (ICLC) to TELOCYTES. *J Cell Mol Med* **14**:729–740.

Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. 2017. The Human Cell Atlas: from vision to reality. *Nature* **550**:451–453.

Stecco C, Fede C, Macchi V, Porzionato A, Petrelli L, Biz C, Stern R, De Caro R. 2018. The fasciacytes: A new cell devoted to fascial gliding regulation. *Clin Anat* **31**:667–676.

Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, Griesbeck M, Butler A,
Zheng S, Lazo S, Jardine L, Dixon D, Stephenson E, Nilsson E, Grundberg I, McDonald D, Filby A, Li W, De Jager PL, Rozenblatt-Rosen O, Lane AA, Haniffa M, Regev A, Hacohen N. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**. doi:10.1126/science.aah4573

Winston JE. 1999. Describing Species: Practical Taxonomic Procedure for Biologists. Columbia University Press.