ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Revising the Cell Ontology

Terrence F Meehan[1], Christopher J Mungall[2], Alexander D Diehl[3]

[1]Mouse Genome Informatics, The Jackson Laboratory, USA; [2]Lawrence Berkeley National Laboratory, USA; [3]Department of Neurology, University at Buffalo School of Medicine and Biomedical Sciences, USA

Abstract. The Cell Ontology (CL) is an ontology of in vivo cell types that is undergoing extensive revision to become a full member of the OBO Foundry. To help us achieve this, a series of goals was established at a CL development workshop in May 2010. Here we describe our ongoing efforts to meet these goals including the modification of the CL's domain, the import of over 400 cell types from the Foundational Model of Anatomy, the addition of both free text and logical definitions, and the incorporation of new terms in response to our user community. These enhancements increase the utility of the CL for both researchers and ontology developers while adhering to the principles of the OBO Foundry.

Keywords: Cell Ontology, CL, OBO Foundry

1 Introduction

The Cell Ontology (CL) is a candidate OBO Foundry ontology for the representation of cell types. First described in 2005 [1], the CL from its earliest incarnation has attempted to integrate cell types from the prokaryotic, fungal, animal and plant organisms. The original developers felt the advantages of having a common framework outweighed the difficulties in incorporating cell types from different phyla. As a core component of the OBO Foundry, the CL merges information contained in species-specific anatomical ontologies as well as referencing other OBO Foundry ontologies such as the Protein Ontology (PR) for uniquely expressed biomarkers and the Gene Ontology (GO) for the biological processes a cell type participates in [2,3].

An area of the CL that has benefited from the ongoing development of another ontology is the hematopoietic cell branch. In conjunction with a revision of immunological processes in the GO, about 80 new immune cell types were added to the CL in 2006 [4]. This brought the CL to the attention of the National Institute of Allergy and Infectious Disease, which sponsored a workshop in 2008 that brought together domain experts and biomedical ontologists to further improve immune cell representation in the CL [5]. Besides identifying missing cell types, participants agreed that creating logical definitions for cell types built from relationships to other OBO ontologies would increase accuracy and interoperability with other ontologies. Based on the experts' input, logical definitions (also known as computable definitions or cross-product definitions) were first created for dendritic cell types [6] and then subsequently for all hematopoietic cell types [7]. This approach has not only increased accuracy of the ontology but has also led to unexpected associations between cell types suggesting the ontology could be used for hypothesis generation. Building on this success, we have begun revising the whole of CL. To help with this, a workshop was convened with ontology experts in May 2010 and several goals were set for the further development of the CL. Here we describe our efforts to meet these goals and revise the CL for entry into the OBO Foundry.

2 Method

2.1 Cell Ontology Development Workshop

A workshop was held at The Jackson Laboratory on May 18-19, 2010. Participants included OBO Foundry ontology developers and users of the CL. The goal of this workshop was how to best resolve the problems of the

CL in relation to the OBO Foundry Principles [8] while meeting the requirements of our funding. A summary of the workshop can be found here: http://obofoundry.org/wiki/index.php/Cell_Ontology_Workshop_2010.

2.2 Editing and Generating Different Versions of the CL

The Cell Ontology has been developed as an OBO format ontology using OBO-Edit software [9]. CL ontology developers modify an editors' version of the ontology in the file, cell.edit.obo, which contains the CL as an unreasoned ontology that includes the minimum and necessary classes (i.e. MIREOTED classes [10]) from other OBO ontologies to allow for sufficient reasoning. An OWL version of the CL, cell.edit.owl, is generated from cell.edit.obo using the standard obolib-obo2owl converter (http://code.google.com/p/oboformat). This converter also performs macro-expansion of shortcut relations, as previously described [7] (also in http://www.berkeleybop.org/~cjm/obo2owl/obo-syntax.html).

An example of a shortcut relation is

lacks_plasma_membrane_part

which is used to define a cell type by an absence of cell surface marker, and has the following macro definition:

has_part exactly 0 (GO:'plasma membrane' and has_part some ?Y)

CL editors used the fully expanded ontology to find errors and make corrections to the cell.edit.obo file. A pre-reasoned version of the CL, cell.obo, is generated from cell.edit.obo and has all implied links asserted and MIREOTED classes removed for full compatibility with existing tools.

CL is available in two separate forms in either of two formats from the URLs below.

2.3 Obol Analysis and Import of Foundational Model Anatomy Classes

The FMA contains a number of cell type classes that are locationally qualified, for example "mesothelial cell of visceral pleura" and "endothelial cell of hepatic sinusoid". To support our goal of making the Cell Ontology the central repository of cell types, we set about generalizing these classes and placing them in CL. We used the Obol tool [11] to parse the labels of the FMA classes into logical definitions, typically consisting of a generic cell type and a gross anatomical location qualified by the part_of relation (for example, "mesothelial cell" and part_of some "visceral pleura"). We mapped the generic cell type to a CL class, and the location to an Uberon class. The mappings were done on the basis of existing dbxrefs maintained in CL or Uberon. Where no generic cell class existed in CL, we manually created one in OBO-Edit.

3 Results

3.1 Providing Different Versions of the CL

The first priority set by members of the CL workshop was to extend our use of logical definitions for hematopoietic cell types to the whole of CL. A logical definition is constructed in a modular fashion by using relationships to classes from other ontologies. These computable definitions can be expressed in ontology formats and languages such as OBO or OWL, and are treated as equivalence relationships between the defined class and some conjunction of classes. For example, the class "pancreatic centro-acinar cell" can be defined as equivalent to the class of things that both are "epithelial cells" and are part of the "pancreatic acinus".

	OBO Format	OWL RDF/XML
Standard	http://purl.obolibrary/obo/cl.obo	http://purl.obolibrary/obo/cl.owl
Basic	http://purl.obolibrary/obo/cl-basic.obo	http://purl.obolibrary/obo/cl-basic.owl

Table 1: Download options for CL
All forms are pre-reasoned. The standard form includes MIREOTed classes, which could cause problems for some tools, we we also provide a basic form that has all MIREOTed classes and references to these classes removed.

We have recently published our work on generating logical definitions for the vast majority of hematopoietic cell types [7] and are continuing to use logical definitions as we add classes. Currently 586 of 1559 CL classes have a logical definition (Table 1). Of these definitions, 442 use macro relations that expand to more complex expressions that can be used by OWL reasoners [7] and are available in the OWL serialization of the ontology. We found that the use of macro relationships was critical in maintaining the logical structure of the CL because some inferences are incomplete with the rule based reasoner (RBR) in OBO-Edit. For example, the cell class "erythroid lineage cell (CL:0000764)" was originally defined with:

"myeloid cell *has_plasma_membrane_part* transferrin receptor protein 1 (PR:000001945)"

while one of its descendent classes "erythrocyte (CL:0000232)" was partially defined as:

"erythroid lineage cell *lacks plasma membrane part* transferrin receptor protein 1"

The reasoners in OBO-Edit are unable to detect this logical inconsistency. However, with an OWL translation of the ontology and the use of macro relations (see Methods), OWL reasoners in Protégé identify the cardinality violation. We then changed the ontology accordingly.

We adjusted our workflow to take advantage of these macro relationships (Figure 1). For a typical user, we felt a pre-reasoned ontology that contained all links inferred by an ontological reasoner as fully asserted links and did not contain classes from other ontologies was important for ease of use. This version is available as http://purl.obolibrary.org/obo/cl-basic.obo and is identical to the "cell.obo" file deposited in the obo library CVS repository, which is maintained for historic reasons. The basic version is also available as OWL. We also make available a complete pre-reasoned version http://purl.obolibrary.org/obo/cl.{obo,owl}.

This version includes the full set of MIREOTed classes and equivalence axioms linking to these classes. The CL editors edit the cell.edit.obo file, which is not pre-reasoned. This file currently resides in the sourceforge CVS repository and is in the process of moving to googlecode. Other versions of the Cell Ontology including an OWL version are provided as described in the methods.

3.2 Incorporating Foundational Model Anatomy Cell Types into the CL

A second important goal for the development of the CL was a reiteration that CL is the ontology for all *in vivo* cell types in the OBO Foundry. Thus, a cell type that is represented in an anatomical OBO ontology should have a species-neutral equivalent in the CL and a mapping between classes. As such, we modified the ontology so that a CL class is a superclass of all the classes stated as dbxrefs. For example, the CL class "photoreceptor cell (CL:0000210)" that describes any animal cell that is able to detect light and transduce a signal, has dbxrefs to equivalent classes in both fly anatomy and human anatomy ontologies. The dbxrefs are translated to *is_a* (SubClassOf) relationships between the CL and the respective classes in the different ontologies. Existing dbxrefs in CL that did not fit this criterion were moved to the comments section for the cell type. Once this was done, we examined the other OBO anatomical ontologies and decided to work first with the Foundational Model Anatomy ontology (FMA) for several reasons. First, the number of cell types represented in the ontology was in the hundreds compared to the thousands represented in some other ontologies. Second, the cell types in FMA contained many general cell types like "endo-epithelial cell" that were not present in the CL. Third, the FMA is a mature ontology that has been in development for over ten years. While the FMA is still being actively developed, major revisions in the ontology were not planned during the months it required us to import the cell classes.

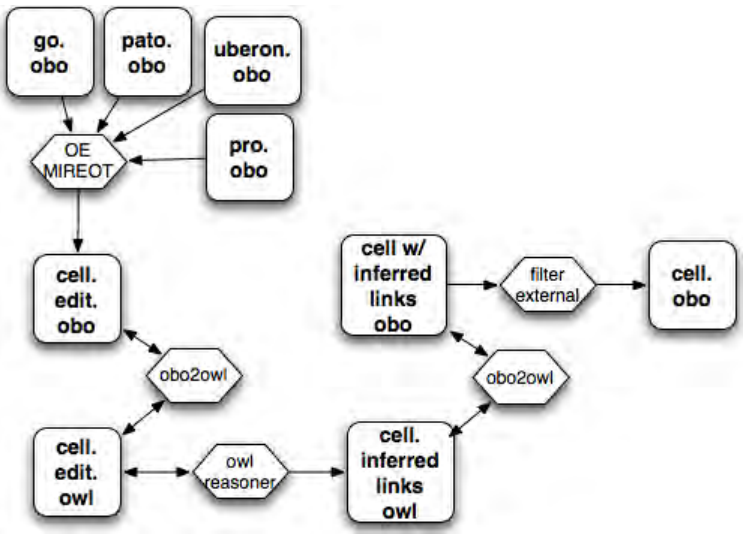

Figure 1. Dataflow for CL. External ontologies are incorporated using the oboedit MIREOT implementation. The obolib-obo2owl tool is used to interchange back and forth between obo and owl. Owl reasoners are used to generate an OWL file with inferred links materialized. The file is converted back to obo, where the file is filtered for the basic version.

We employed a term structure parsing tool called Obol to identify cell types in the FMA [11]. This approach gives unique non-trivial definitions to classes based on implicit rules inherit in the class names. The Obol analysis was able to generate logical definitions for 221 FMA cell types based on the syntax of: "<generic> of <anatomical entity>". In 189 cases, a successful match was made from the "generic" FMA cell type to a pre-existing CL class, The missing 32 generic FMA cell types were then manually added to CL. This information was also supplied to the developers of the Uberon ontology [12]. Uberon is a multi-species anatomy ontology created to facilitate comparison of phenotypes across multiple species. As such, the Uberon developers referenced the species-neutral anatomical structures in their ontology to the corresponding structures in the FMA. Once these mappings were complete, the 221 cell types parsed by the Obol analysis were added automatically to the CL complete with logical definitions using Uberon classes and dbxref to the FMA class.

Obol analysis of the FMA was unable to parse 539 classes beyond the classification of "cell type". Of these, 213 classes had dbxrefs references in the CL and 88 were neuronal cell types, which were put aside for a future workshop (see "Discussion"). The remaining 238 classes were reviewed in detail and ultimately over 200 new cell types were added with free text definitions to the CL. Cell types ranged from those type that failed to parse in the Obol analysis due to historical names like "Boettcher cell" to cell types whose names reflected their highly specialized nature like "type II cell of carotid body". While laborious, addition of these cell types added great value to the CL including many cell types performing unique biological roles that will aid in logical definitions of the Gene Ontology. An important addition to the CL was cell types that reflect lineage development such as "endo-epithelial" defined as "An epithelial cell derived from endoderm." 83 cell types have this term as an ancestor, which means all these cell types can be traced back to the endoderm lineage (Figure 2). This greatly extends the representation of developmental lineages in the CL and will serve as a cornerstone as we further develop this aspect of the ontology. By placing FMA cell types in the context of the CL and using logical definitions, we could find equivalent cell types. For example, Type F (FMA:83409) and Type PP (FMA:62938) enteroendocrine cells were found to be equivalent as both cell types are defined by secreting pancreatic polypeptides.

Developers of the FMA are considering obsoleting the cell-type classes in favor of the CL classes. Until this is done, bridging between FMA and CL will be provided by our dbxrefs where a reference to a FMA class represents the human equivalent of cell-type represented by a species neutral CL class.

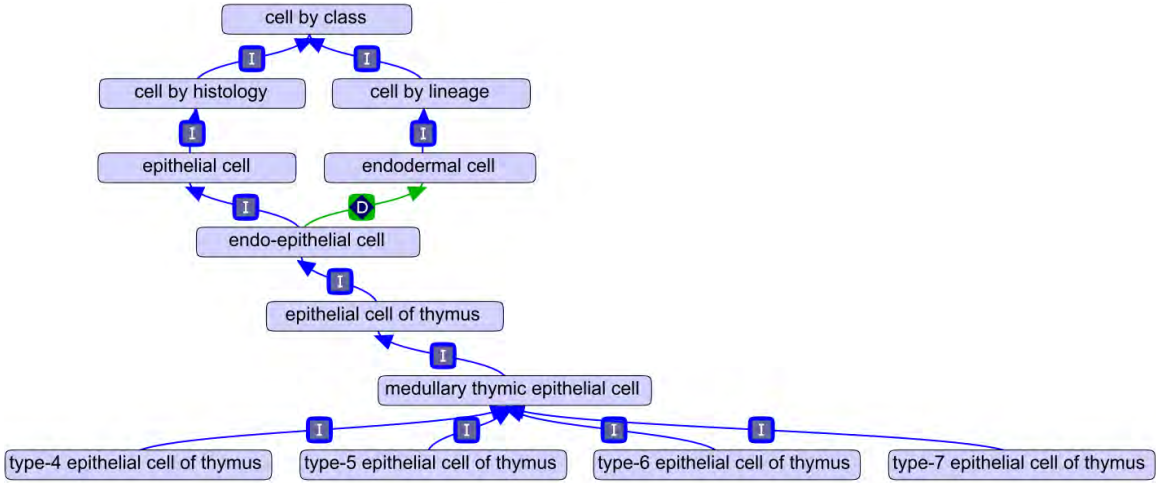

Figure 2. Endo-epithelial cell type and a subset of its descendents.

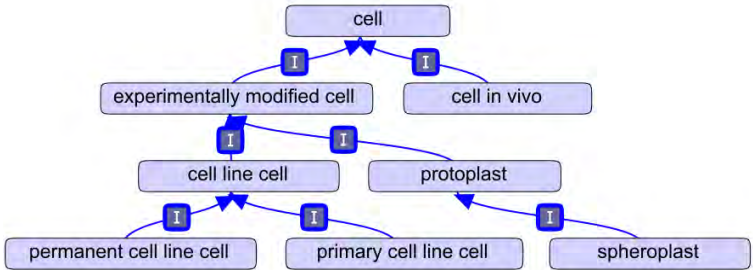

Figure 3. Descendants of experimentally modified cell will be removed from CL

3.3 Removal of Experimentally Modified Cell Types

Another goal established at the Cell Ontology workshop is that experimentally modified cell types should be moved into OBI. When the CL was implemented, *in vitro* cell types were included in its domain such as "primary cell line" and "permanent cell line" (Figure 3). Since then, the Ontology of Biomedical Investigations (OBI) has been undergoing active development and many of the workshop participants felt *ex vivo* cell types and cell lines fall under OBI's domain. OBI is a candidate OBO Foundry ontology that represents the design, materials, implementation, and data of biological investigations [13]. As cell lines are experimentally derived entities, they arguably fall within OBI's domain. Biologists agree for multiple reasons that cell lines often bear little resemblance to the cell types they are derived from. For example, principal component analysis of a large number of gene expression arrays demonstrates that cell lines fail to cluster with their tissues of origin and instead tend to cluster among themselves [14]. While there was general agreement that cell line classes be moved to OBI, many participants felt that mappings between cell lines and the CL should be made. Towards this goal, we have been working with the developers of a cell line ontology that is based on the Cell Line Knowledgebase [15]. We have begun adding missing cell types to the CL from which these cell lines are derived. These additions include logical definitions to Uberon classes to provide a generalized anatomical context. A derives_from relationship is used in the cell line ontology to the appropriate CL class. More about this work appears in a related ICBO submission.

3.3 Other Developmental Goals for the CL

Several other goals were established for the development of the CL. One goal is to provide free text definitions and biomedical references for all CL classes. Currently 1263 classes of

1559 have some sort of free text definition with the majority of those containing references to the biomedical literature (see Table 1). Another goal was to improve response time for requested changes in the CL. Excluding neuronal classes (see "Discussion"), we have worked through a three-year backlog of tracker items on the CL SourceForge tracker and now respond to most term requests within a week. Other accomplishments achieved since the workshop include the import of our sub-ontology of hematopoietic cells (called Hemo_CL) [7] into the CL, providing better documentation of the CL structure and development, and increasing our outreach to potential users of the CL ontology.

4 Discussion

Here we describe our improvements in the representation of cell types in the Cell Ontology that enhances its use in data annotation and integration. By implementing these changes, the CL is taking on a core role in the OBO Library by serving as a conduit to link anatomical ontologies to the GO, the PR and other OBO ontologies. We have done this in a manner that adheres to the principles of the OBO Foundry, namely by working collaboratively with other OBO Foundry members to provide a clearly delineated ontology expressed in both OBO and OWL formats that is available for use by all.

Ontology	cell.edit.obo v1.1	cell.edit.obo v1.62
Cell type classes	988	1559
Classes with free text definitions	565	1263
Classes with logical definitions	0	586
Relationships used with CL classes	2	10
Number of External OBO classes MIREOTED into the CL	0	1005
Number of xrefs	121	564

Table 2. Summary of changes in the CL from September, 2009 through January, 2011

Perhaps the most important goal for development of the CL is to continue outreach to both ontology developers and to end-users. Much of the work described herein stems from CL development workshops that involved biomedical researchers. Thus, the CL has become more reflective of the needs of our users. For example, we have recently been asked to join the FANTOM5 consortium, which seeks to map transcription initiation in over 200 human cell types. The organizers felt the CL's representation was broad and deep enough to help structure the terabytes of data the project is expected to generate, and we have committed to adding additional cell types needed for the FANTOM5 work. Outreach also helps coordinate the CL with other OBO Foundry ontologies. We have recently held a workshop to extend the representation of neuronal cell types in collaboration with the International Neuroinformatics Coordinating Facility (INCF). The workshop included experimental neuroscientists including members of the INCF Neuron Registry Task Force and developers of anatomical ontologies for Human, Drosophila, rodent, and the pan-species Uberon ontology. Despite differences in neuronal representation by these ontologies, we believe a unified approach to neurons is possible based on our past experience with having ontology developers and biologists work together and we have now begun curating neurons based on this approach. This in turn will help integrate data in the respective ontologies.

In summary, we continue to develop the CL to enhance its usefulness for researchers and ontology developers, and for consideration as a full member of the OBO Foundry.

Acknowledgments

The Cell Ontology project is supported by an NHGRI-funded, ARRA administrative supplement grant HG002273-09Z to the parent grant, HG002273, to the Gene Ontology Consortium. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol. 2005;6(2):R21.
2. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, et al. Cross-product extensions of the Gene Ontology. J Biomed Inform. 2011 Feb;44(1):80-86.
3. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. Nucleic Acids Res. 2011 Jan;39(Database issue):D539-545.
4. Diehl AD, Lee JA, Scheuermann RH, Blake JA. Ontology development for biological systems: immunology. Bioinformatics. 2007 Apr 1;23(7):913-915.
5. Diehl AD, Augustine AD, Blake JA, Cowell LG, Gold ES, Gondré-Lewis TA, et al. Hematopoietic cell types: prototype for a revised cell ontology. J Biomed Inform. 2011 Feb;44(1):75-79.
6. Masci AM, Arighi CN, Diehl AD, Lieberman AE, Mungall C, Scheuermann RH, et al. An improved ontological representation of dendritic cells as a paradigm for all cell types. BMC Bioinformatics. 2009;10:70.
7. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, et al. Logical development of the cell ontology. BMC Bioinformatics. 2011;12:6.
8. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 2007 Nov;25(11):1251-1255.
9. Day-Richter J, Harris MA, Haendel M, Lewis S. OBO-Edit--an ontology editor for biologists. Bioinformatics. 2007 Aug 15;23(16):2198-2200.
10. Courtot M, Gibson F, Lister AL, Malone J, Schober D, Brinkman RR, et al. MIREOT: The minimum information to reference an external ontology term. Applied Ontology. 2011 Jan 1;6(1):23-33.
11. Mungall CJ. Obol: integrating language and meaning in bio-ontologies. Comp. Funct. Genomics. 2004;5(6-7):509-520.
12. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. 2009 Nov;7(11):e1000247.
13. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010;1 Suppl 1:S7.
14. Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. 2010 Dec 23;11(12):R124.
15. Sarntivijai S, Ade AS, Athey BD, States DJ. A bioinformatics analysis of the cell line nomenclature. Bioinformatics. 2008 Dec 1;24(23):2760-2766.