Modularization for the Cell Ontology

Christopher J. Mungall[1], Melissa Haendel[2], Amelia Ireland[1], Shahid Manzoor[1], Terry Meehan[3], David Osumi-Sutherland[4], Carlo Torniai[2], and Alexander D. Diehl[5]

[1] Lawrence Berkeley National Laboratory
[2] Oregon Health and Sciences University
[3] The Jackson Laboratory
[4] The University of Cambridge
[5] University at Buffalo

Abstract. One of the premises of the OBO Foundry is that development of an orthogonal set of ontologies will increase domain expert contributions and logical interoperability, and decrease maintenance workload. For these reasons, the Cell Ontology (CL) is being re-engineered. This process requires the extraction of sub-modules from existing OBO ontologies, which presents a number of practical engineering challenges. These extracted modules may be intended to cover a narrow or a broad set of species. In addition, applications and resources that make use of the Cell Ontology have particular modularization requirements, such as the ability to extract custom subsets or unions of the Cell Ontology with other OBO ontologies. These extracted modules may be intended to cover a narrow or a broad set of species, which presents unique complications. We discuss some of these requirements, and present our progress towards a customizable simple-to-use modularization tool that leverages existing OWL-based tools and opens up their use for the CL and other ontologies

1 Introduction

Many bio-ontologies were initially conceived of as stand-alone monolithic entities, developed independently of other ontologies. However, a modular approach, whereby portions of other ontologies are reused and made interoperable has many advantages[23], and this was one of the reasons for the establishment of the OBO Foundry[14][1][25]. With a modular approach, more complex ontology classes are constructed combinatorially using simpler ontology classes as building blocks. These building-block classes may come from separate ontologies, or from orthogonal hierarchies within a single ontology. For example, a cell type such as *mature eosinophil* in the OBO Cell Ontology (CL)[2] can be functionally defined using the biological process class *respiratory burst* from the Gene Ontology (GO)[20]. Inter-ontology dependencies such as these can be bi-directional; for example, a GO process such as *eosinophil differentiation* can be defined in terms of the CL class *eosinophil*[21]. The CL is also *taxonomically modularized*[6]

[6] in this manuscript we use the term *taxon* in the sense of an organism/species taxonomy

in that it leaves representation of highly specialized cell types to species-specific ontologies. Table 1 shows the external ontologies that are of relevance to the CL modularization strategy.

Ontology	Scope and Relevance to CL
PR[22]	*The Protein Ontology* Proteins from multiple species - Used to define cell types based on presence of specific proteins
GO[12]	*The Gene Ontology* Biological Processes, Molecular Functions, and Cellular Components - Biological Processes used for defining cells by function, and Cellular Component for surface receptors
PATO[9]	*The Ontology of Phenotypic qualities* Qualities that can apply to antatomical entities or biological processes - used to define cells in terms of shape and other physical characteristics.
UBERON [11]	*The Uber anatomy ontology* Gross anatomical structures spanning metazoa (but like CL, with a significant vertebrate bias) - used to define cells by location in the organism
NCBITaxon	*Taxonomy of species* For taxonomic constraints. Only a very small subset is required.
MA[13]	*The adult Mouse Anatomy ontology* Species-specific gross anatomy
FBbt[8]	*The Drosophila anatomy ontology* Species-specific gross and cellular anatomy
WBbt[19]	*The C. elegans anatomy ontology* Species-specific gross, cellular and subcellular anatomy
FAO[3]	*The Fungal Anatomy Ontology* Multi-species gross, cellular, and subcellular anatomy
ZFA	*The Zebrafish Anatomy ontology* Species-specific gross and cellular anatomy
TAO[5]	*The Teleost Anatomy ontology* Multi-species gross and cellular anatomy
PO[16]	*The Plant Structure ontology* Multi-species gross and cellular anatomy
FMA[24]	*The Foundational Model of Anatomy - adult human* Species-specific gross, cellular, and subcellular anatomy

Table 1. Ontologies required by the Cell Ontology for importing modules and/or coordinating development.

Module extraction

When working in the context of multiple ontologies, it is important to be able to extract sub-modules from combinations of ontologies. For example, when working with the CL, it can be useful to extract the minimal subset of GO that is

required to perform automated reasoning over the CL and obtain results that are valid and complete. This subset can either be *imported* or *merged* into the source ontology. If the entire external ontology is imported or merged without first extracting a subset, the resulting ontology union can be difficult to work with and reason over.

Module extraction is also useful for downstream applications, such as using an ontology in annotation or analysis. Annotators may want a subset of the ontology that is of relevance to their taxon or domain of interest, and term enrichment tools benefit from using a subset as it decreases the size of the hypothesis space, resulting in improved p-values. Since its inception, the GO has catered to these use cases by providing manually created subsets or "GO slims"[12]. Using "slimming tools" (Ireland, unpublished), GO annotations can be mapped from a full ontology to a slim.

Manually creating subsets is a time-consuming task, and will not scale for all purposes, so automated techniques are extremely valuable. The problem of extracting minimal subsets that preserve reasoner results has received considerable attention in the Description Logic literature - see for example[17]. The majority of the discussion has been on the theory; some module extraction tools have been implemented for OWL ontologies, but they are not always easy to use.

The MIREOT (Minimal Information for Retrieval of External Ontology Terms) guidelines[4] and associated tooling[27] provide support for practical module extraction. One notable feature of MIREOT is that external ontology axioms are typically merged into the source ontology rather than imported, potentially leading to synchronization issues. MIREOT has been adopted by some ontologies, such as the Ontology of Biomedical Investigations (OBI) and eagle i[10]. The CL is currently using a MIREOT strategy with the ontology editor OBO-Edit[6] to create an extended version of the CL, which includes externally referenced classes. These are removed for the "basic" version of the CL.

Taxonomic module extraction

Another requirement is to extract sub-modules from unions of ontologies. For example, cross-species comparison of phenotypes requires reasoning over multiple ontologies[26]. For many purposes it can be useful to extract modules from the union of CL with species-specific anatomy ontologies.

Taxonomic modularization requires a slightly different strategy. This was first proposed and formalized by Kusnierczyk[18], and later implemented in the GO[7]. For example, the GO states that *lactation* occurs only in *Mammalia*, allowing a module extraction tool to automatically generate a *Drosophila* subset that excludes this term. However, other requirements, such as generating labels specific to certain taxa remain unmet.

As a multi-species ontology that is integrated with multiple species-specific anatomy ontologies (AOs) (see lower half of Table 1), the CL has particular requirements here. There is overlap between the general terms in CL and the

species-specific terms in these AOs, with the degree of overlap varying depending on the ontology.

For example, there is little overlap between plant and metazoan cell types, so it makes sense to manage these in separate ontologies. The Plant Ontology (PO), which combines cells and gross anatomy in a single ontology, is taking responsibility for plant cell types, leaving CL to focus on metazoa.

The situation is more complex when we consider the *Drosophila* Anatomy Ontology (FBbt). Managing all Drosophila cell types in CL would be difficult: this ontology has over 1,500 neuronal cell types, many of which are specific to this taxon, and this number is likely to grow. Representing these cell types in FBbt allows linkages between cell types and Drosophila gross anatomy to be maintained in a simple and logically coherent way. At the same time, we want to coordinate on a shared representation of core cell types such as "neuron" in the CL. We also want CL to have very specific cell types for mammals (note that the adult Mouse Anatomy ontology, MA, does not represent cell types). This tension between a shared general representation and individual specific representations creates challenges for ontology management. In addition, many users want to be able to obtain a single coherent ontology view of all cell types within a clade, or across all organisms, requiring intelligent combination of multiple ontologies.

The strategy thus far has been for CL to represent generalized cell types as far as possible, with taxonomic specificity indicated by constraints in the ontology, and for taxon-centric ontologies such as ZFA and FBbt to represent these cell types as they are instantiated in particular species, with OBO format "xrefs" (semantics-free cross-references) linking the two.

Towards an integrated tool

There is a lack of a single integrated tool that can fulfil all of these requirements. In an effort to redress this, we have specified a list of capabilities such a tool should have for working with the CL, and present initial progress towards the implementation of such a tool.

2 Cell Ontology Requirements

Axiom rewriting using subsets

A class subset S is a collection of classes $c^1, c^2, ..., c^n$ taken from an ontology O. An ontology O can be rewritten as an ontology O' such that O' contains no references to classes not in S, yet is still consistent with O. This process is colloquially known in GO as *slimming*. Note that we use the term ontology in the sense of any collection of axioms; this means that if we have a formalization of GO associations in OWL, we can use the same algorithm for mapping associations.

Note that the axioms in the target ontology need not be a subset of the axioms in the source ontology - some axiom rewriting may be required. Consider the case where X is a subclass of Y and Y is a subclass of *part_of* some Z, and

$S = X, Z$, then a simple subsetting operation will lose the axioms connecting X to Z. The following procedure should be used to extract a subset S from ontology O:

Create a target ontology O' that is an exact copy of the source ontology O
Remove all axioms from O' where that axiom references a class not in S (i.e. all classes in the signature of the axiom must be in S).
Reason over O to find all inferred axioms[7] A.
For each axiom in A, add that axiom to O', provided this is not redundant with anything in O'. An axiom is redundant if it exactly matches an existing axiom, or it is entailed by O'.

For OBO format (obof) ontologies, the ontology should first be converted to OWL, after which it can be converted back to obof; this ensures correct interpretation when implementing the above procedure.

If the source ontology contains equivalence axioms (*intersection_of* tags in obof) that reference a class not in the subset, this procedure will rewrite them as plain SubClassOf axioms (*is_a* or *relationship* tags in obof). This is the desired behavior, as writing the IDs but keeping equivalence axioms would result in incorrect inferences.

Ontology property subsets

Some ontologies use a large number of properties (relations), some of which may be organized in a hierarchy. For example, the FMA has many different relations, and distinguishes between 3 sub-properties of *part_of* (systemic, regional and constitional). Sometimes it is desirable to map these to the generic relation.

Here we can specify an ontology *property subset*, excluding the sub-properties of *part_of*. Then when we use the procedure above, axioms are automatically "mapped up" to the generic relation.

For example, if the source ontology contains an axiom X subclass of *regional_part_of* some Y, and the regional part relation is a sub property of *part_of*, then a reasoner can infer X subclass of *part_of* some Y. If the property subset contains only the generic relation, then the target ontology would have only the latter axiom and not the former.

Annotation Axiom Rewriting

When constructing a union of the general Cell Ontology and a species-specific ontology such as FBbt, we are faced with a problem that the resulting ontology will result in classes with non-unique labels, since we will have both CL:0000540 (neuron) and FBbt:00005106 (neuron). One highly impractical solution is for

[7] Whilst strictly speaking "inferred axiom" is an oxymoron, the OWL literature uses "axiom" in place of "sentence" and frequently distinguishes between inferred and asserted axioms

each anatomy ontology to ensure their primary labels are globally unique - for example, FBbt:00005106 would be labeled "*Drosophila* neuron". Another approach would be to merge selected class pairs as part of the process of creating the union - for example, merging FBbt:00005106 into CL:0000540. One must then decide how to deal with the axioms of the merged classes. If the axioms are combined it can generate problematic statements such as "every (generic) neuron is part of a Drosophila nervous system" - obviously false for a zebrafish Purkinje cell. The opposite approach, discarding axioms, loses potentially useful information.

The accepted solution is to create a *bridge ontology* connecting the ontologies, and include annotation assertions in this bridge ontology for multi-context labels. For example, the bridge ontology would assert that FBbt:00005106 would have an "OBO Foundry unique label" of "Drosophila neuron" or "neuron (Drosophila)" where the taxon is included in the label. This would only be necessary for taxon-specific subclasses of generic classes, but it may be simpler to apply this uniformly across the species-specific anatomy ontology.

The modularization procedure can then merge the generic Cell Ontology with the cell subsets of the species-centric anatomy ontologies and rewrite the primary label axiom to use the OBO Foundry label, adding an axiom annotation to the axiom stating the source of the rewriting.

Taxonomy reasoning based module extraction

Many ontologies such as GO and CL are intended to be applicable across taxa. This means that these ontologies typically contain modules that are useful to one community and not another; for example the class *mammary gland epithelial cell* in the CL would not be useful for gene expression queries for chicken. The taxonomic constraint strategy used by GO[7] has been adapted for UBERON and will be used for the CL, replacing the *sensu* designators that are currently in use.

One of the main use cases for taxon-based module extraction from the CL would be to provide modules that exclude non-taxonomically relevant classes. For example, in a generic cell ontology it is useful to have a generic "erythrocyte" class and two subclasses, depending on whether the erythrocyte is nucleate or enucleate. However, most species have one or the other form, so when creating a taxon module the irrelevant classes can be discarded. For example, for a mouse module, only "enucleate erythrocyte" is required. It may also be desirable to give this the label "erythrocyte" when used in a mouse context.

This kind of automated taxonomic extraction is possible, provided the ontology has enough axioms to support this. For the above example, the ontology would have to state that (1) "erythrocyte" is a the disjoint union of "enucleate erythrocyte" and "nucleate erythrocyte" (alternatively, this could be inferred if these classes are defined using GO) and (2) no mammal erythrocyte is a nucleate erythrocyte (i.e. a standard taxonomic constraint). The taxonomy ontology will tell us that mouse is a subclass of mammal. The module extraction procedure for a taxon t is then to add an axiom for every class c in the ontology, stating

that c is in-taxon some t. We then eliminate any unsatisfiable classes, and merge equivalent classes, using the more generic label as primary.

Taxonomic bridge ontologies

Cross-species ontology integration can be assisted by means of *bridging axioms* – for example, ZFA:0009248 (neuron in ZFA) is a subclass of the generic CL:0000540 (neuron in CL).

Maintaining these bridging axioms explicitly can be difficult since the resulting ontology has a highly latticed structure, so an alternative approach is to use a feature specified in obo format 1.4 called "*xref macros*". Here header directives can be used to indicate how xrefs for a particular ontology are to be translated. For example, use a *treat-all-xrefs-as-has-subclass* header directive in CL, all FMA xrefs in CL can be expanded to:

```
Class: FMA_54527
SubClassOf: CL_0000540
```

We can even make a stronger taxonomically-qualified equivalence axiom by including a *treat-all-xrefs-as-genus-differentia* directive together with appropriate IDs:

```
Class: FMA_54527
EquivalentTo:
  CL_0000540 and
  part_of some NCBITaxon_9606
```

i.e. any CL neuron in a human is equivalent to a FMA neuron. These header macros are applied on a per-ontology basis.

A modularization tool for CL should be capable of generating the logical axioms from the xrefs, and placing these in the requisite bridge ontologies, from where they can be subsequently merged.

Creation of taxon-union importing ontologies

We plan to publish modules that import subsets of both CL and external anatomy ontologies for different taxonomic clades. This is already possible to a certain extent with UBERON - figure 1 shows the OWL import chain for a pan-eukaryotic anatomy ontology which selectively imports pan-anatomy ontologies for different clades.

The modularization tool should be able to use taxon ontologies to dynamically build these importer ontologies.

Fig. 1. Import chain for uberon pan-eukaryote anatomy (http://purl.obolibrary.org/obo/uberon/mod/uberon-combined-eukaryote.owl). This selectively imports bridge modules, species specific anatomy ontologies and recursively imports taxonomically more restricted import ontologies. The bridge modules are generated from xrefs stored in the main UBERON file. The zoomed area shows how the metazoa module imports the vertebrate module plus selected invertebrates, and the vertebrate module imports the amniote module plus selected anamniotes/

3 Implementation Progress

The CL modularization tool is being developed as part of the OWLTools library (http://code.google.com/p/owltools/), and will be released in the fall of 2011. OWLTools is layered on top of the OWLAPI[15], so it can take advantage of standard OWL reasoners and generic modularization tools. It also takes advantage of the new obo2owl implementation (http://code.google.com/p/oboformat/), and should thus be capable of working with ontologies whose source is either OBO format or OWL.

4 Conclusions

OWL modularization tools provide powerful and formally sound means of extracting modules from multiple ontologies that are amenable to reasoning. These tools would become even more useful for the bio-ontologies community if embedded in software that is aware of common metadata tags used in OBO ontologies and of taxonomic constraints. We have outlined some specific requirements for a generic tool that is currently being developed to perform these tasks.

References

1. M. Ashburner, CJ Mungall, and SE Lewis. Ontologies for biologists: a community model for the annotation of genomic data. In *Cold Spring Harbor symposia on quantitative biology*, volume 68, pages 227–235, 2003.
2. Jonathan Bard, Seung Y Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005.
3. M.C. Costanzo, M.S. Skrzypek, R. Nash, E. Wong, G. Binkley, S.R. Engel, B. Hitz, E.L. Hong, and J.M. Cherry. New mutant phenotype data curation system in the Saccharomyces Genome Database. *Database*, 2009(0), 2009.
4. M. Courtot, F. Gibson, A.L. Lister, J. Malone, D. Schober, R.R. Brinkman, and A. Ruttenberg. Mireot: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011.
5. Wasila M. Dahdul, John G. Lundberg, Peter E. Midford, James P. Balhoff, Hilmar Lapp, Todd J. Vision, Melissa A. Haendel, Monte Westerfield, and Paula M. Mabee. The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. *Syst Biol*, 59(4):369–383, 2010.
6. John Day-Richter, Midori A Harris, Melissa Haendel, Gene Ontology OBO-Edit Working Group, and Suzanna Lewis. OBO-Edit–an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, Aug 2007.
7. J. Deegan, E. Dimmer, and C.J. Mungall. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC bioinformatics*, 11(1):530, 2010.
8. R. A. Drysdale, M. A. Crosby, W. Gelbart, K. Campbell, D. Emmert, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhang, P. Zhou, M. Zytkovicz, M. Ashburner, A. de Grey, R. Foulger, G. Millburn, D. Sutherland, C. Yamada, T. Kaufman, K. Matthews, A. DeAngelo, R. K. Cook, D. Gilbert, J. Goodman, G. Grumbling, H. Sheth, V. Strelets, G. Rubin, M. Gibson, N. Harris, S. Lewis, S. Misra, and S. Q. Shu. FlyBase: genes and gene models. *Nucleic Acids Res*, 33(Database issue):D390–5, 2005. 1362-4962 Journal Article.
9. GV Gkoutos, C Mungall, S Doelken, M Ashburner, S Lewis, J Hancock, P Schofield, S Khler, and PN Robinson. Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*, 2009.
10. M. Haendel, M. Wilson, C. Torniai, E. Segerdell, C. Shaffer, R. Frost, D. Bourges, J. Brownstein, and K. McInnerney. Eagle-i: Making invisible resources, visible. *Journal of Biomolecular Techniques: JBT*, 21(3 Suppl):S64, 2010.
11. Melissa A. Haendel, Georgios G. Gkoutos, Suzanna E. Lewis, and Chris Mungall. Uberon: towards a comprehensive multi-species anatomy ontology, August 2009.
12. M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics

resource. *Nucleic Acids Res*, 32 Database issue:D258–61, 2004. 1362-4962 Journal Article.

13. T.F. Hayamizu, M. Mangan, J.P. Corradi, J.A. Kadin, and M. Ringwald. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biology*, 6(3):R29, 2005.
14. D. P. Hill, J. A. Blake, J. E. Richardson, and M. Ringwald. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, 12(12):1982–91, 2002. 1088-9051 Journal Article.
15. M. Horridge. The OWL API: A Java API for Working with OWL 2 Ontologies. In *6th OWL Experiences and Directions Workshop (OWLED 2009)*, 2009.
16. P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, et al. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7):388–397, 2005.
17. B. Konev, C. Lutz, D. Walther, and F. Wolter. Semantic modularity and module extraction in description logics. In *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 55–59. IOS Press, 2008.
18. Waclaw Kusnierczyk. Taxonomy-based partitioning of the Gene Ontology. *Journal of Biomedical Informatics*, 41(2):282–292, 2008.
19. R.Y.N. Lee and P.W. Sternberg. Building a cell and anatomy ontology of Caenorhabditis elegans. *Comparative and Functional Genomics*, 4(1):121–126, 2003.
20. Terrence Meehan, Anna Maria Masci, Amina Abdulla, Lindsay Cowell, Judith Blake, Christopher Mungall, and Alexander Diehl. Logical development of the cell ontology. *BMC Bioinformatics*, 12(1):6, 2011.
21. Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-Product Extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44(1):80 – 86, 2011. Ontologies for Clinical and Translational Research.
22. Darren A Natale, Cecilia N Arighi, Winona C Barker, Judith Blake, Ti-Cheng Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, and Cathy H Wu. Framework for a protein ontology. *BMC Bioinformatics*, 8 Suppl 9:S1, 2007.
23. Alan L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 121–128, Sanibel Island, FL, USA, 2003. ACM.
24. C. Rosse and J.L.V. Mejino. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics 36:478-500.*
25. Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007.
26. Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation. *PLoS Biology*, 7(11), 2009.

27. Z. Xiang, M. Courtot, R.R. Brinkman, A. Ruttenberg, and Y. He. OntoFox: web-based support for ontology reuse. *BMC research notes*, 3(1):175, 2010.