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

Rapid Development of an Ontology of Coriell Cell Lines

Chao Pang, Tomasz Adamusiak, Helen Parkinson, James Malone

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

Abstract. Many online catalogues of biomedical products and artifacts exist that are loosely structured but of great value to the community. These include cell lines, enzymes, antibodies, reagents, and laboratory equipment. Improving the representation of these products has several benefits; detailed querying power, product comparison and reporting of products used in experimental protocols. However, this formalization is often time-consuming, labor-intensive and expensive. We describe an approach to structuring these catalogues using semi-automated techniques to rapidly develop OWL ontologies. We demonstrate the approach using the Coriell Cell Line catalogue, and the resulting ontology of 28,000 classes which imports classes from other community ontologies such as Disease Ontology, Cell Type ontology and FMA. **Availibility:** http://efo.sourceforge.net/coriell.htm

Keywords: Coriell ontology, automated ontology engineering, cell line ontology

1 Introduction

The biomedical community has embraced the use of ontologies as a means of describing scientific data, such as experimental protocols (OBI) [1] experimental variables (EFO) [2] and phenotypes [3]. The development of these ontologies, however, is a costly activity. It requires considerable time and expertise to produce a large and/or complex ontology.

There is clearly value in producing robust expertly curated ontologies. The Gene Ontology (GO) [4] is the archetypal example of this, is developed by a team of experts and is continuously updated to include new biological knowledge. However, development in this form is clearly not repeatable across every area of biomedicine. There is evidence that methods and tools that expedite the process of ontology engineering are much needed [5].

Programmatic approaches can be powerful when transforming resources with some pre-existing structure into an ontological form [6]. Loosely structured data sources contain implicit knowledge – within the data or within the presentation layer, for example within categories in a drop-down list on a website. This is often the case for data stored in a database which is accessible through a web interface, which may contain some logic behind the sorting of options but which is otherwise unavailable. Similarly, such implicit knowledge may also be contained within the column headers of spreadsheets or within database table and field names. In such cases it may be possible to exploit implicit knowledge and develop models which enable a rapid transform into ontology classes.

In this paper we present our approach to the rapid development of the large Coriell cell line ontology based on a collection of semi-structured cell line descriptions from the Coriell cell line catalogue. The Coriell cell line catalogue contains ~27,000 mammalian cell lines and we demonstrate that by using a standardized modeling pattern and text mining approaches, a large ontology containing >28,000 classes can be rapidly produced which logically describes each cell line and their biological properties.

2 Methods

The principle methodology underlying this work is ontology normalization [7]. Specifically, that we manage multiple inheritance using class descriptions in OWL and infer structure using description logic reasoners such as HermiT [8]. By providing axioms on classes, the need to assert potentially conflicting or fragile subsumption hierarchies is removed. This approach also makes the biological knowledge used to create the hierarchy explicit and therefore renders implicit knowledge explicit in the ontology.

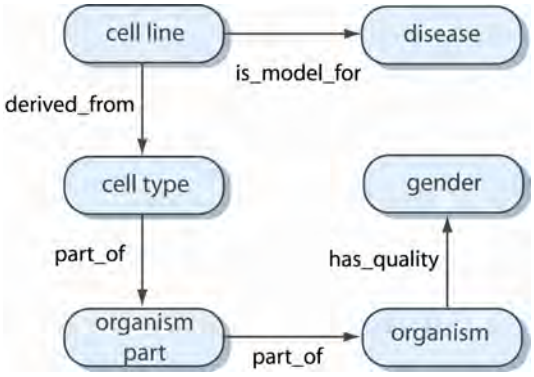

Figure 1. Part of the cell line model used to represent Coriell cell lines.

The first step in producing the ontology was to develop a standardized model for cell lines. By collaboration with the Cell Line Ontology [9] and the Cell Type Ontology [10] we created a model (Figure 1) which aligns these ontologies and which was used during development of the Coriell Cell Line Ontology.

Our primary queries of interest are contained in this model, specifically: cell line name, cell type, disease, organism parts, organism and gender. The model therefore represents the key attributes of Coriell cell Lines and was evaluated against primary competency questions derived from use cases related to the development of a BioSample Database (www.ebi.ac.uk/biosamples/) at the EBI. These include queries by common cell types, by disease and tissues. We a use a 'short cut' relation, is_model_for, to associate cell lines with disease, this reflects the use of cell lines as models for particular diseases. Given the large size of the Coriell catalogue we developed a scaleable semi-automatic approach to creating the ontology. Information on each cell line was contained within 104 separate and redundant text files each describing different aspects of the Coriell products and derived from an SQL dump of a relational database. Five key files were selected which contained semi-structured descriptions covering the entities described in Figure 1 and which corresponded to our use cases. These files were merged, redundant information was removed and a single 'cell line' spreadsheet was produced using bespoke Perl scripts.

2.1 Lexical Entity Mapping

The cell line spreadsheet was used as an input for lexical entity mapping with the aim of generating list of classes from reference ontologies that matched the textual descriptions. The Perl OntoMapper [2] was employed as it has previously been used successfully in building similar application ontologies. The approach allows for fuzzy matching to identify classes from class labels and their synonyms. Given the nomenclature of areas such as disease and anatomy where synonymy is common, a fuzzy matching approach provided flexibility in mapping. A metric was assigned to each match and those with less than 100% confidence were manually inspected.

The reference ontologies (Table 1) were selected based on the content of the files and the model. Anatomy was particularly challenging as although the Coriell cell lines were primarily mammalian no single mammalian anatomy ontology exists which would provide the coverage necessary. Although some efforts are ongoing to develop an homology based anatomy ontology [11, 12] we used a pre-existing resource the Minimal Anatomy Terminology [13]. This species neutral ontology provides mappings to multiple anatomical ontologies and is subsumed by the Experimental Factor Ontology, with which we plan to merge the Coriell Cell Line Ontology in future. When a core mammalian anatomy ontology becomes available we will replace the MAT. Some human specific classes were also imported from FMA.

The disease information within the Coriell descriptions consisted of references to OMIM [14]. Since OMIM is not a disease ontology we exploited the links provided within the Human Disease Ontology (DO) to OMIM and imported DO classes. Where links were not made between OMIM and DO, a manual inspection using BioPortal [15] was required to extract the corresponding disease.

Domain	Reference Ontology	Term Number
Organism	NCBI Taxonomy, OBI	93
Anatomy	Experimental Factor Ontology, FMA	61
Cell Type	Cell Type Ontology	11
Disease	Human Disease, NCI Thesaurus	337
Gender	PATO	3

Table 1. Reference ontologies used in the Coriell cell line ontology.

2.2 Ontology Engineering Using OWL-API

The lexical mapping resulted in a set of files containing mappings between a label and the corresponding URI from the reference ontology, one file per domain. These mappings were used to construct the ontology programmatically – Figure 2 illustrates this process.

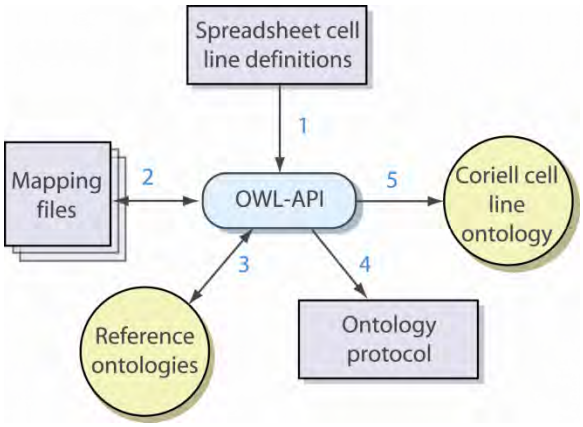

Figure 2. Methodology used for programmatic ontology creation

The process was implemented as follows:

(1) Input of cell line descriptions contained in the single merged spreadsheet.

(2) Files containing mappings from class label to reference ontology class IRI (Internationalized Resource Identifier) are matched.

(3) Class IRIs are used to import corresponding ontology classes from reference ontologies, along with axiomatic and annotation information within the class signature if present and parent classes.

(4) The EFO upper level is re-used here (a slim version of BFO) and determines where imported classes should be placed. For example, disease classes are imported under the *disease* parent, itself a child of *disposition*. This protocol also determines how axioms on classes should be formed.

(5) The Coriell cell line ontology in OWL is output.

(6) The ontology was manually reviewed for correctness, checked for consistency using HermiT 1.3.1 and defined classes were added to ensure axioms were used, formulated correctly and meaningful as well as to add structure to the hierarchy.

3 Results

The Coriell cell line ontology produced contains 27,002 cell line classes, covering 11 cell types, 61 anatomical terms and 93 organisms. 657 OMIM numbers were attached to cell lines and 393 OMIM numbers were mapped to 337 unique Disease Ontology classes. 7,688 cell lines were confirmed to model disease and a small number modeled multiple diseases, for example ND00139 which models Parkinson's Disease and Lewy Body Disease.

Following the creation of the ontology some refinements to the imported structure were required.

3.1 Organism Taxonomy

Organism classes imported from the NCBI taxonomy [16] have very long chains of parent classes. For example Homo sapiens has 28 classes in a subclass hierarchy between it and the parent class organism. We then retrospectively removed some of these nodes, applying the following design principle; 1. Remove intermediate classes when the child class does not have more than 2 siblings, 2. When the deletion leads to >3 child classes, the parent class is retained. This strategy removed a large number of classes which were not required by our query use cases and these could easily be added back in if needed in future.

3.2 Anatomy

There were 81 unique terms describing anatomy, 45 mapped exactly to pre-existing terms in the MAT. Unmapped terms describe classes other than anatomy such as fibroma, leiomyoma (which are disease classes) and were removed. Buttock-thigh and Thorax/abdomen could be separated into two single terms but it is not clear which part the terms were describing and these were also removed. 9 terms were unmapped which did not appear to fit into anatomy, such as Keloid breast organoid, so were removed. Among the remaining terms unmapped from the entity recognition step, 12 terms are mapped to FMA, 9 terms to EFO, 2 terms to SNOMED CT ontology, 2 terms to

NCI Thesaurus and one term is not mapped to any ontology. The mixing of terms from disease and anatomy domains was found to be common in many parts of the Coriell Catalogue and manual effort was spent assessing these terms prior to building the ontology.

3.3 Cell Type

22 unique cell type terms were mapped to the Cell Type Ontology. 11 terms are with 100% similarity. Partial mappings were refined manually e.g. *smooth muscle* is not a cell type and was modified to *smooth muscle cell*. Myeloma is not a cell type, but a cancer of plasma cells and was changed to plasma cell. Another 11 unmapped terms were not cell type terms and were removed.

3.4 Disease

We use the structure described in Disease Ontology in the Coriell cell line ontology and imported 337 disease terms. DO, it is not axiomatised except for the use of subclass relationships. EFO, however, provides more information for the class relationships (e.g. disease to anatomical parts). For disease we therefore added axioms to allow construction of defined classes based on e.g. disease e.g. '*liver disease cell lines*'. Firstly DO classes mapped from the cell line description spreadsheet were imported including any DO metadata. Then imported classes were axiomatised using additional logical restrictions from EFO (e.g. an axiom linking disease to anatomical part). The axiomatic information imported from EFO does not affect the DO child and parent classes and therefore the canonical structure from DO (and DO IRIs) is preserved.

3.5 Adding Defined Classes to Infer Structure

Use of normalisation methodology results in an asserted flat cell line hierarchy, i.e. the only asserted parent class of each cell line is the cell line class. For browsing purposes, however, it is often useful to produce an organizational hierarchy and as such we created some under cell line using defined classes in OWL, i.e. classes with necessary and sufficient restrictions describing members.

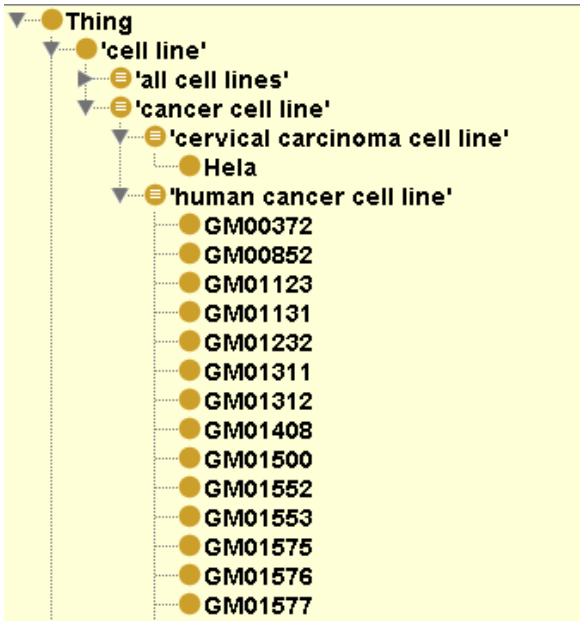

Figure 3. Inference of cell line hierarchy shown in Protégé.
The defined class *human cancer cell line* has necessary and sufficient conditions which result in various cell lines inferring as subclasses following the use of a description logic reasoner. This is very useful in rapidly creating dynamic hierarchies which can be changed very easily and for managing multiple inheritance.

For example, human cancer cell line (Figure 3) has the following necessary and sufficient restriction using Manchester OWL syntax [17]:

```
'cell line'
 and (is_model_for some cancer)
     and (derives_from some
     ('cell type'
          and (part_of some
          ('organism part'
               and (part_of some 'Homo sapiens')))))
```

The nesting reflects an important distinction between separate statements; in effect, we are saying for a specific *organism*, for which a specific *organism part* is part, and from which a specific *cell type* was taken. For the example in Figure 3, the defined class restricts membership to those classes where cancer is the modeled disease and which are derived from humans (more specifically cell types that are part of an organism part which are part of humans).

We have also used disjoints in some areas of the ontology, for example by making Homo

sapiens disjoint from other siblings under organism, we are able to ask the query for things which are not Homo sapiens because they have been explicitly stated as such. The following returns a class of cell lines that are not derived from human:

```
'cell line'
and (derives_from some
('cell type'
        and (part_of some
        ('organism part'
             and (part_of some
             (organism
                  and (not ('Homo sapiens')))))))
```

3.6 Rapid Generation and Rapid Regeneration

The ontology was developed over 3 months by one person working full time. The majority of this time was spent developing the code to produce the ontology and a repeat exercise using similar methods would take a great deal less. We made several changes to the ontology as we progressed and refined the model slightly; the programmatic method used meant regenerating the new OWL ontology took minutes. Moreover, rapidly adding new content programmatically is also possible.

4 Discussion

One of the central claims of this work is that the ontology was rapidly developed using the methods described. Over the 3 months that this work was conducted, we estimate 2 months comprised investigation of the catalogue content and Perl scripting to merge and format the initial input files. A further month's programming resulted in an ontology of ~28,000 classes. Generalizable components of the methodology include: design of reusable design patterns, re-use of ontology development code and exploitation of the MIREOT process for term imports.

There is a trade-off between hand crafted curation by individual experts and the rapid development of a very large resource. Our approach is of most benefit when a semi-structured data exists and existing Foundry type ontologies are available e.g. for cell types. As a one-off SQL dump was used for development updates need to be managed in future and a dynamic method for accessing new data is desirable.

One of the criteria for inclusion in the OBO Foundry effort [18] is that every class is given a textual definition. The effort required to manually produce good textual definitions for an ontology the size of the Coriell cell line ontology is significant. Given the axiomatisation of the ontology, however, efforts such as producing natural language from OWL statements may offer an effective and rapid method to producing textual definitions [19]. If such an approach can be applied we will seek to include the artifact into the OBO Foundry in the future. We are also currently working with the Cell Line Ontology to ensure our respective models are synchronized and to merge the Coriell cell line ontology with the CLO which is currently derived from the American Tissue Culture Collection (ATCC). Other work includes mapping to all resources which contain cell line references and addition of these to the ontology, re-running of imports to detect changes in source ontologies, term requests from e.g. the cell type ontology to classify cells by anatomical part and addition of information manually where possible e.g. much text containing phenotypic descriptions was unstructured and could be mined added. A complete evaluation of additional meta data vs. that of the CLO is also desirable in order to prioritise where to add curation effort and which additional data could added to the core we have built. This work has allowed us to refine the cell line model within EFO to be consistent with the CLO and this will be revised in future releases of EFO. Future work also includes the release of the Coriell ontology to Bio2RDF for linked open data access. Finally our programmatic approach is fully compatible with manual curation and ontology development, and a combined approach is likely to produce rich, well structured ontologies for community use.

Acknowledgments

We thank the Functional Genomics Production Team, the Coriell Institute for Medical Research, Alan Ruttenberg and Science Commons for providing the Coriell SQL dump. Lynn Schriml and colleagues from the Disease Ontology for OMIM mappings and Sirarat

Sarntivijai, Oliver He, Alexander Diehl and Terry Meehan for discussions on the cell line model. **Funding**: The European Molecular Biology Laboratory, and EC (HEALTH theme no. 200754 Gen2Phen).

References

1. The OBI Consortium (2010) Modeling experimental processes with OBI. *J. Biomedical Semantics*, 1(Suppl 1):S7.
2. Malone, J. Holloway, E. Adamusiak, T. Zheng, J. Kolesnikov, N Zhukova, A. Kapushesky, M. Brazma, A. Parkinson, H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112-1118.
3. Mungall, C. Gkoutos, G. Smith, C. Haendel, M. and Lewis, S. and Ashburner, M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biology* ((1))R2.
4. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.,* 25(1):25-9.
5. Falconer, S. Noy, N. and Storey, MA. (2007): Ontology mapping – a user survey. In Shvaiko, P. Euzenat, J. Giunchiglia, F. and He, B. eds.: *Proc. of the Workshop on OntologyMatching (OM2007) at ISWC/ASWC2007*, Busan, South Korea.
6. Antezana, E. et al. (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biology, 10(5):R58.
7. Rector, AL. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proc. of 2nd Int. Conf. on Knowledge Capture 2003.*
8. Motik, B. Shearer, R. and Horrocks, I. (2009) Hypertableau reasoning for description logics. *J. of Artificial Intelligence Research*, 36:165–228.
9. Sarntivijai, S. et al. (2011) Cell Line Ontology: Redesigning Cell Line Knowledgebase to Aid Integrative Translational Informatics. *ICBO 2011*, Buffalo. *Accepted.*
10. Meehan, TF. Masci, AM. Abdulla, A. Cowell, LG. Blake, JA. Mungall, CJ. Diehl, AD. (2011) Logical development of the cell ontology. *BMC Bioinformatics*, 12:6.
11. Dahdul, WM. et al. (2010) The Teleost Anatomy Ontology: Anatomical representation for the genomics age. *Systematic Biol.* 59(4): 369–383.
12. Travillian, RS. Adamusiak, T. Burdett, T. Gruenberger, M. Hancock, J. Mallon, A-M. Malone, J. Schofield, P. and Parkinson, H. (2010) Anatomy ontologies and potential users: Bridging the gap. Proc. of the Workshop on Ontologies in Biomedicine and Life Sciences, Mannheim, Germany, 2010.
13. Bard, JBL. Malone, J. Rayner, TF. and Parkinson, H. (2008) Minimal anatomy terminology (MAT): a species-independent terminology for anatomical mapping and retrieval. In Proc. of ISMB 2008 SIG meeting on Bio-ontologies, Toronto.
14. OMIM, Online Mendelian Inheritance in Man (2011) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine, (date accessed: January 12, 2011). URL: http://www.ncbi.nlm.nih.gov/omim/
15. Noy, NF. et al. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 1;37(Web Server issue):W170-3.
16. Sayers, EW. et al. (2009) Database resources of the National Center for Biotechnology Information. *Nuc. Aci. Res.* 37(Database issue):D5-15.
17. Horridge, M. Drummond, N. Goodwin, J. Rector, A. Stevens, R. and Wang, H. (2006): The Manchester OWL syntax. *In OWLed 2006.*
18. Smith, B. et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25:1251-1255.
19. Stevens, R. Malone, J. Williams, S. Power, R. and Third, A. (2011) Automating generation of textual class definitions from OWL to English. *J. Biomedical Semantics,* 2(suppl 2):S5.