Part-of Relations in Anatomy Ontologies: A Proposal for RDFS and OWL Formalisations

J.S. Aitken, B.L. Webber, and J.B.L. Bard

PART-OF RELATIONS IN ANATOMY ONTOLOGIES:
A PROPOSAL FOR RDFS AND OWL FORMALISATIONS

J.S. AITKEN, B.L. WEBBER

School of Informatics, University of Edinburgh,
Appleton Tower, Crichton St. Edinburgh EH8 9LE

J.B.L. BARD

Department of Biomedical Sciences, University of Edinburgh,
George Square, Edinburgh EH8 9XD

Abstract

Part-of relations are central to anatomy. However, the definition, formalisation and use of *part-of* in anatomy ontologies is problematic. This paper surveys existing formal approaches, as well as the use of *part-of* in the Open Biological Ontologies (OBO) anatomies of model species. Based on this analysis, we propose a minimal ontology for anatomy which is expressed in the Semantic Web languages RDFS and OWL-Full. The paper concludes with a description of the context of this work in capturing cross-species tissue homologies and analogies.

1 Introduction

The increasing number of anatomies being defined, published and linked to gene expression data provides new opportunities to explore tissue homologies across species, and their relationship to the genetic evidence. Anatomies are now available for the main model species (c elegans, drosophila, mouse, zebrafish etc), and are expressed in the format developed for the Gene Ontology [1] (GO). While GO was neither intended to encode knowledge of tissue homology, nor to include species-specific concepts, this knowledge can be very useful to biologists exploring gene expression data. However, automated techniques for manipulating this knowledge are needed.

Traditionally, homologous tissues are those which share a *common evolutionary ancestor* (bat wings and human forelimbs - both pentadactyl limbs); but, where there is no fossil record, it also means tissues with a *common developmental lineage* (the gut of mouse and gut of C. elegans). The other tissue relationship is analogy - where tissues have a similar function but different evolution/development (e.g. insect limbs and vertebrate limbs). We describe techniques for acquiring and representing homology knowledge from experts, encoding it through links between the tissues (terms) in OBO anatomies. Arising from this concern with anatomy, we also consider the structure and formal properties

of the OBO anatomies. For the proper treatment of reasoning about a single anatomy, and for making inferences about cross-anatomy links, the meaning of the anatomical terms and relationships requires such a clarification.

Anatomies differ of necessity due to the different developmental patterns and radically different structures of the species concerned. However, these differences are unnecessarily compounded in existing ontologies by the differences in terminology used for different species, and differences in the way the ontology relations are interpreted and used. For example, the OBO mouse anatomy uses only *part-of* while drosophila uses *isa*, *part-of* and *lineage*. As long as the anatomies are used by humans, mismatches in how relations are used and differences in modelling practice have few serious consequences. But, they will limit and impoverish any attempt to automate symbolic reasoning about anatomy. Anatomists are extending the types of relationships required to capture anatomical associations, but there is no consensus on a clear semantics for the essential notions of type, part and developmental order. Potentially useful distinctions such as sex, axis, stage, normal/modified gene complement are introduced in some anatomies, but are far from common usage.

This paper presents an analysis of the formal approaches to part-whole reasoning in anatomy. We then present a simple ontological framework, represented in RDFS and OWL-Full, which can be used to resolve the semantic and syntactic problems identified in GO. We then describe tools and techniques for acquiring homologies between the tissues defined in OBO anatomies.

2   Anatomy and Ontology

We introduce the Gene Ontology, with an emphasis on its formal aspects, and survey existing approaches to the formalisation of anatomical concepts. The proposed GO Schema (upper ontology) and homology mapping techniques are then presented in this context.

2.1   OBO and the Gene Ontology

OBO includes ontologies for the anatomies and developmental timelines of a range of plant and animal species. The semantics and syntax of OBO ontologies adhere to the open standard of the Gene Ontology.

The Gene Ontology is a curated, controlled vocabulary for the species-independent annotation of genes and gene products. GO conforms to a central proposition of ontology development - it is a product of a community effort, so can be said to represent a consensus. GO is composed of three ontologies, biological process, molecular function, and cellular

component, which are meant to be orthogonal and independent of each other.[1] The concepts in these ontologies are known as 'terms', each with a unique ID.

GO terms are intended to be at the class (or type) level rather than describing a single instance. The *true path rule* should hold for all GO terms: the pathway from a child term to its top-level parents should always be true.[1] In addition, citations and evidence supporting GO attributes and GO annotations must be provided.

The *isa* and *part-of* relations are used to make links between terms– GO ontologies are directed acyclic graphs. In such graphs, an entity can have multiple links to parents via *isa* and *part-of*. The *isa* relation has the interpretation of subclass: *B isa A* means all *B's* are *A's* (however, *isa* sometimes taken to mean 'instance of'[1] and under this interpretation *isa* is not transitive). *part-of* relates an entity and its components, and is intended to be transitive. However, the definition (as specified in natural language in [1]) allows this relation to hold if some instances of the parent have an instance of the child as a component. Interpreted formally, the property of transitivity does not follow from the definition. Assuming that both *isa* and *part-of* are to be interpreted as transitive, then a *part-of* association stated about an entity must hold for all subtypes of that entity. However, without a clear model for the interpretation of relations, or any automated reasoning to compute the deductive inferences, it appears unsafe to assume that anatomy developers have accounted for the interaction of *isa* and *part-of*. A critique of GO[2] notes the overloading of *isa* to denote both *type-of* and *instance-of*, and the confusion in practice of *isa* and *part-of* relations.

On the syntactic level, ontologies that are written in the GO format can be stored in several formats, including: flat file, XML/RDFS, or OWL (being an extension of XML/RDFS). The flat file format uses indentation by white space and the symbols % < to denote the hierarchy, but is unable to store the textual annotations that accompany GO terms. These are stored in the definitions file. The XML/RDFS syntax removes the reliance on white space, and permits a full record of the GO term to be made. The XML/RDFS syntax makes use of the RDF mechanism of URIRefs - references to named concepts which are defined in web-accessible files. In fact, RDF cannot be used without the RDF schema and this schema provides both a subclass relation, and the facility to introduce and define new types of relation.

As noted above, there is a need to model relationships such as lineage, requiring an extensible syntax for expressing the ontology. We explore these semantic and syntactic issues in a new RDF Schema for GO presented in Section 3.

2.2 Axiomatising part-of for Anatomy

Winston, Chaffin, and Herrmann [3] identify six senses of *part-of* in their analysis of the semantics of meronymic relations underlying English usage. The senses are: component-integral, stuff-object, portion-mass, place-area, member-collection, feature-activity. The first five of these senses are used in the GALEN ontology of human anatomy,[4] which we discuss in more detail below. The six senses are differentiated from each other by three properties that can be associated has holding between parts and wholes: *functional* - the parts play a functional role; *homeomerous* - the parts are similar to each other and to the whole; *separable* - in principle, the parts can be disconnected from the whole.

The formal properties of a generic *part-of* relation are analysed in the theories of mereology and topology. This *part-of* relation is often simply denoted (P $part$ $whole$). When axiomatised in first-order logic, it is usually defined to to transitive and symmetric: [5]

1. $(P\ x\ x)$
2. $(P\ x\ y)\ \land\ (P\ y\ z)\ \rightarrow\ (P\ x\ z)$
3. $(P\ x\ y)\ \land\ (P\ y\ x)\ \rightarrow\ x = y$

This formal analysis allows two *part-of* relations to be differentiated: a *proper part*, PP, is any part excluding the whole,[5] and a *direct part*, DP, is a proper part which is an immediate part of the whole:[6]

4. $(PP\ x\ y)\ :=\ (P\ x\ y)\ \land\ \neg(P\ yx)$
5. $(DP\ x\ y)\ :=\ (PP\ x\ y)\ \land\ \neg(\exists z\ (PP\ x\ z)\ \land\ (PP\ z\ y))$

It is notable that in cases where *part-of* is observed to lack the property of transitivity, the explanation is often that several senses of *part-of* are being used — there is no inherent opposition between the conceptual and formal approaches. In the formal analysis, *part-of* is a relation between instances of objects. However, in GO, *part-of* relates classes of objects, so the meaning of the class-level relation needs to be defined in terms of the conventional instance-level relation.

As anatomy ontologies may specify tens of thousands of terms, the acquisition and maintenance of anatomical knowledge, and the efficiency of automated reasoning become major considerations. For these reasons, description logics and frame-based approaches have been adopted. A key issue in the practical application of anatomical knowledge is the propagation of properties up and down the *part-of* structure, and the need to control the generalisation/specialisation of these properties. As the models of anatomy that have been developed cannot be separated from the language they are expressed in, we discuss both together, beginning with description logics (DL).

Recent approaches to exploiting the subsumption mechanism of DL to perform *part-of* reasoning include that of Schulz and Hahn,[7] who have translated the anatomical terms in UMLS into description logic. In order

to use subsumption in this way, a single anatomical entity is modelled by additional concepts that denote the structure of the entity and the set of parts that correspond to the entity. A property such as *perforation-of* attributed to *Colon Structure* will correctly generalise to *Intestine Structure*, as these entity structures are in a *isa* relationship.[7] A property which should not be generalised, such as *inflammation-of*, is defined to hold of the entity, as opposed to the entity structure, and no *isa* link holds.

Description logics may allow a relationship to be declared to be transitive, in which case it is not necessary to use subsumption reasoning to get the transitivity property. For example, GRAIL,[4] the DL used in the GALEN project, allows *part-of* relations to be declared transitive. As noted earlier, GALEN uses five senses of *part-of*, but it also specialises them further. For example, the component-integral sense is represented by *component-of*, and this has the specialisation *func-component-of* to distinguish functional components. One of the reasons for introducing this refinement concerns the inheritance of the *function-of* property along *component-of*: this property inherits along *component-of* but not along *func-component-of* to limit inheritance as being valid up to a *"certain (often arbitrary) level of anatomical aggregation"*[4].

GALEN also includes an *arbitrary part* relation to describe structures whose parts are not delimited in a conventional anatomical sense. The need to distinguish anatomical parts from arbitrary parts is also noted in the Digital Anatomist Foundational Model.[8] In this frame-based approach, the *part-of* relation has the attributes of being shared or unshared, and of being arbitrary or anatomical. This formalisation avoids the need to define a set of different *part-of* relations with different properties: *part-of-shared-arbitrary*, *part-of-shared-anatomical* and so on. Further, Protege's frame language permits the definition of constraints that hold at the class level (and therefore apply to all sub-concepts), and constraints that hold of the concept but do not get inherited by more specialised concepts (the so-called own-slots).

Other extensions of DLs that are potentially relevant include those which permit rule-like reasoning (Horn clauses), and those which define 'plural quantifiers' that are able to discriminate between parts or wholes with different properties (e.g. parts physically connected to the whole) to specify how properties are generalised or specialised [9].

OBO anatomies typically contain none of the fine-grained distinctions used in these models of human anatomies. As noted, *part-of* may be used exclusively, and may be used in place of *isa*. Where both are used, the type hierarchy based on *isa* is often incomplete, containing concepts whose parent is only identified by a *part-of* link. That is, the class hierarchy does not form a connected DAG. Similarly, the *part-of* model may be fragmented, having anatomical entities that are not part of any
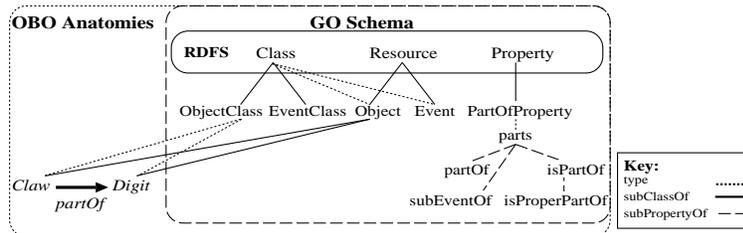
Figure 1: *part-of* vocabulary in the GO Schema

other anatomical entity. Furthermore, parts often have no associated type. No consistent modelling style or convention is imposed on the directed acyclic graph model.

Despite these formal shortcomings, the OBO anatomies are extensive, encode valuable knowledge, are intuitive to biologists and are linked to gene expression data, and so constitute valuable resources. Enhancing the machine processability of anatomies by clarifying the semantics of the terms, and assigning proper definitions are the immediate tasks that we address by constructing a schema for GO. This schema can be used in addition to the existing GO XML DTD (completing the anatomies in terms of the *part-of* and *isa* structure must be left to the curators of the OBO anatomies).

3   The GO Schema

The schema we propose clarifies the meaning of GO relationships, and provides an extendible framework at both semantic and syntactic levels. This schema can be thought of as a minimal upper ontology. We view this as an intermediate step towards a fully formal ontology supported by inferencing capabilities, which may be in a description logic based language such as DAML-OIL or OWL.

Following the GO standard, OBO anatomies currently have an XML/RDF syntax, but none of the RDF Schema features are used. RDFS provides a well-defined *subClassOf* relation (*isa*), and a *subPropertyOf* relation.[a] The latter could be used to specialise a generic *lineage* relation to *descends-in-the-male* - which is one of the ways the C Elegans anatomy models lineage. RDFS-enabled tools can either fully understand the *descends-in-the-male* relation, or can simply treat it as *lineage*. RDFS also now has a great deal of code support in Java (e.g. Jena [10]).

---

[a]Note that OWL also uses these relations, and that OWL-Full has the same expressivity as RDFS - differing primarily by the namespace to be used.

The proposed GO Schema extends the existing RDFS/OWL-Full classes (*Class*, *Resource*, *Property*) by making a fundamental distinction between *Event* and *Object*. *Events* are things that can be said to occur or 'happen', while *Objects* are things that exist over time. This distinction is common to the process ontology PSL[11] and approximately corresponds to occurrents and continuants,[2] where occurrents cannot be said to exist at a single moment in time, while continuants can preserve their identity from moment to moment. Figure 1 illustrates how the GO Schema extends RDFS to create a richer vocabulary in the ontological layer. OBO anatomies are expressed as assertions using the new vocabulary.

Based on the event/object distinction, two types of *part-of* relation are defined: *partOf* and *subEventOf*. *partOf* relates *ObjectClasses* to their parts, which must be subclasses of *Object*. Similarly, *subEventOf* relates composite *EventClasses* to their constituent *EventClasses*. A lineage predicate, *successorOf*, is introduced to relate later to earlier developing tissues (*ObjectClasses*). While this is not a *part-of* relation, we include it due to its prevalence in anatomy models. As yet we do not define further specialisations of these relations, as might be based on the functional/homeomerous/separable properties.

On adopting the schema, concepts in OBO anatomies and in the GO cellular component ontology become subclasses of *Object*, while (the majority of) the GO process and molecular function concepts become subclasses of *Event*.[2] The *go:part-of* relation is replaced by *partOf* and *subEventOf* as appropriate.

We argue that for the kinds of anatomy ontologies that biologists have created, it is important to distinguish the direct parts of a tissue, and to know whether *part-of* should be interpreted as transitive or not. Consequently, the GO Schema defines a number of primitive relations, *isPartOf* and *isProperPartOf*, which are used to define the semantics of *partOf*. *isPartOf* and *isProperPartOf* correspond to the *P* and *PP* relations introduced earlier and defined by formulae 1-4. The semantics of *partOf* are defined below, where we use the transformation from RDFS to a first-order logic described in:[12]

(?Subject ?Predicate ?Object)^{RDFS}
↔ (PropertyValue ?Predicate ?Subject ?Object)^{FOL}
(?Subject type ?Object)^{RDFS} ↔ (Type ?Subject ?Object)^{FOL}

```
(=> (and (PropertyValue partOf ?P ?W)
         (PropertyValue classDefinition ?P ?W))
    (forall (?w)
      (=> (Type ?w ?W)
          (exists (?p)
             (and (Type ?p ?P)
                  (PropertyValue isProperPartOf ?p ?w))))))
```

Essentially, *(partOf Part Whole)*, plus the *classDefinition* qualifier, means that every instance of *Whole* has some instance of *Part* as a part. Note that *Whole* and *Part* are classes, as all concepts are types in GO. The semantics of *subEventOf* and *successorOf* are defined in the same manner as for *partOf*. Note that *Type* relates an instance to the class it belongs to.

The *partOf*, *subEventOf* and *successorOf* relations can be qualified in several ways: they can be *classDefinitions*, meaning the whole necessarily has parts of the specified type, or can be *termDefinitions*, meaning that the part-of assertion does not necessarily apply to sub-types of the whole (i.e. it is not inherited). The part-of relation can be declared to be a direct part relation by the qualifier *directPartDefinition*, and if the part always exists as part of the whole, then the *partDefinition* qualifier can be used. All qualifier relations hold between the part class and the whole class[b]. The *directPartDefinition* qualifier strengthens the part-of definition:

```
(=> (and (PropertyValue partOf ?P ?W)
         (PropertyValue directPartDefinition ?P ?W))
     (forall (?w)(forall (?p)
      (=> (and (Type ?w ?W) (Type ?p ?P)
               (PropertyValue isProperPartOf ?p ?w))
                (not (exists (?z)
                  (and (PropertyValue isProperPartOf ?p ?z)
                       (PropertyValue isProperPartOf ?z ?w))))))))))
```

The qualifier relations are used in conjunction with the basic *partOf* relation, and therefore we do not need to introduce a whole set of *part-of* relations corresponding to combinations of properties. This approach is in accord with that of [8] who reify the *part-of* slot in order to state the shared/anatomical properties. The approach allows the different versions of part-of to be stated using the RDF syntax of subject-predicate-object, and for these to be stated for each part and whole assertion (as opposed to defining a single part-of relation for all parts/wholes in the anatomy).

The motivation for introducing the direct-part property is to be able to distinguish direct and inferred parts in anatomy models and in queries over those models. The not-inherited property, *termDefinition*, reflects the current state of OBO anatomies where the interaction of *part-of* and subclass has not received the necessary attention. Thus, *part-of* assertions in current OBO anatomies can be immediately and safely translated into *partOf* plus *directPartDefinition* and *termDefinition*. Further modelling input will be required to remove the qualifications, if they can correctly be removed. The use of relations that hold between classes puts

---

[b]See `http://www.aiai.ed.ac.uk/resources/go/go-schema` for the RDF Schema and OWL formalisation.

the ontology into the OWL-Full category. The qualifiers *classDefinition* and *partDefinition* can be translated into equivalent OWL-DL expressions, while the direct-part relation can be represented as a specialisation of part-of, but the axiomatisation cannot be captured in OWL-DL.

The GO Schema plus the associated axioms provide a way to represent and interpret OBO anatomies in an unambiguous way. The RDFS/OWL-Full approach allows new sub-types of *part-of* to be introduced in an extendible way. Further, the approach allows discussion of the meaning of relationships (the axioms) without impacting on the syntax of the anatomies. This seems very appropriate as the level of consensus in the OBO community is rather low, and discussion and clarification seems desirable.

A method for the step-wise refinement of GO to a DL representation is described in [13]. We concur that a clarification of the meaning of terms by creating definitions is desirable, and is a requirement for automated support for ontology curation. Our approach differs in that we allow the axiomatisation of relations to use full first-order logic (of which DLs constitute subsets) while remaining within the syntax of RDFS. However, there is as yet no inference support for the proposed formalisation. We argue that moving immediately to DAML-OIL or OWL-DL will preclude other options for the formal language, for example, DLs which provide plural quantifiers, DLs which support Horn clause reasoning, and frame-based approaches. As discussed earlier, *part-of* reasoning in 'plain' description logics has proved in the past to raise complex modelling issues.

4 Tools and Techniques for Acquiring Homology Mappings

The formal issues discussed above are important in the context of our work on defining homology and analogy mappings between anatomies, as we aim to make inferences on a more substantial basis than can be done by relying on concept names alone. However, we must use the existing OBO resources and we now describe the mapping process that links these anatomies.

4.1 Homology Data

The homology mappings are associations between (the existing) OBO anatomies. One of the key ideas is to identify the cell types that justify the homology. These cell types must be common to the tissues linked by homology. The cell type is drawn from a cell ontology that includes all the cell types for all the major phyla. This ontology includes type and lineage information. The major cataloguing categories are class

(function, morphology, lineage, gender-specific, number of nuclei, ploidy and stem cells) and organism (animal, fungus, plant, prokaryotic, spore).

We identify four types of homology: *tissue-homology*, *cell-function-homology*, *analogy* and *association*. The homology relationship holds between two tissues, typically selected from the anatomy of different model species. Database identifiers as well as tissue names are recorded for all homology mappings. Our approach requires that one or more cell types be associated with the homology mapping. The author of a homology mapping should be recorded, as should the date, and a textual annotation may be made. It is assumed that only one homology relation holds between any two tissues.

An example of a cell function homology is listed below. The mapping is from `WBdag:3681` to `FBbt:5612`, that is, from the pharynx in C Elegans to embryonic esophagus in Drosophila. The relationship is symmetric: it can be interpreted as holding from Drosophila to C Elegans. The cell ontology term *lining_cell* (`CL:213`) is given as a basis for the homology. In addition, a textual annotation provides additional information, and where possible, a reference to the literature.

```
mapping:cell-function-homology
WBdag:3681        name:"pharynx"
FBbt:5612         name:"embryonic esophagus"
CL:213            name:"lining_cell"
author:"J. Bard"  date:"26.02.03"
annotation:"Porteriko & Mango (Dev Biol 233, 482, 2001)
say that the CE pharynx links the buccal cavity to the
midgut and hence = pharynx unlikely to be a lineage
homology as D eosophagus is ectodermal"
```

4.2 Methodology

To assist the recording of homologies, we have implemented an acquisition tool which displays two ontologies and permits the user to select terms and enter the homology data. Figure 2 shows the user interface.

The tool allows the data to be recorded, and ensures that sufficient information is entered. The anatomist is responsible for exploring the two ontologies and identifying the most plausible homologies. This is necessarily a manual process which relies on the knowledge and expertise of the biologist.

In an independent exercise we shall consider each anatomy and associate one or more cell types with each tissue at a leaf node in the anatomy. The cell types characterise the essential properties that distinguish one tissue from another. Then, these cell types will be propagated upwards through the anatomy.

Once a set of homology mappings has been acquired we shall perform two types of evaluation. Firstly, for each tissue marked as being part
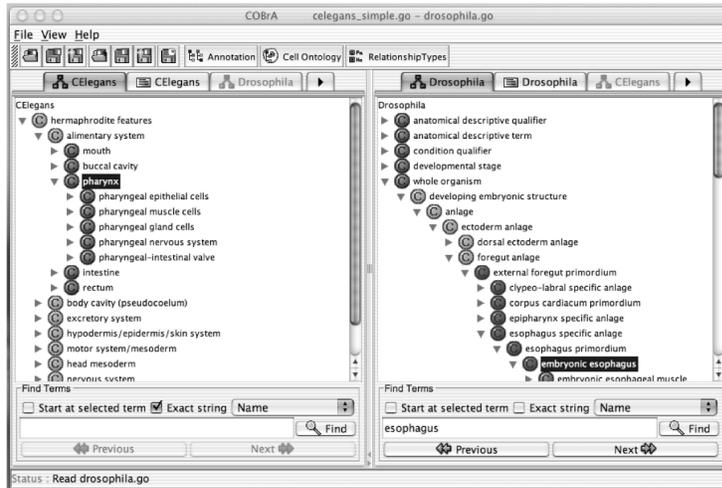
Figure 2: The COBrA Homology Acquisition Tool

of a homology relation, we will determine whether the cell types from the homology generation exercise match those from the tissue-cell type annotation exercise or not. Any differences will be resolved, and a final homology to cell type association determined. The use of cell types to provide definitional knowledge of homology is a key innovation of our technique. The second evaluation will be a critiquing exercise which will involve biologists in the respective fields. Together, these evaluations will provide a degree of confidence in the knowledge acquisition process.

5 Conclusions

We have presented an analysis of semantic and syntactic problems in the language used for OBO anatomies. The proposed solution allows a consensus to be reached on the relations required to describe anatomy, and their properties. We also show how a cell ontology can be used to define homology mappings between tissues in OBO anatomies. A further use of the cell type ontology is to provide concepts that can be used to define the properties of tissues. The cell ontology provides the means to define species-specific concepts (parts in an anatomy) in terms of more primitive species-independent concepts, and so can play an important role in deepening the ontological modelling of anatomy.

Acknowledgments

References

1. Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, 11:1425–1433, 2001.
2. B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. Proc. AMIA 2003 *in press*.
3. M.E. Winston, R. Chaffin, and D. Herrmann. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11:417–444, 1987.
4. J. Rogers and A. Rector. Galen's model of parts and wholes: Experience and comparisons, 2000. Annual Fall Symposium of American Medical Informatics Association, LA.
5. A. Varzi. Parts, wholes, and part-whole relations: The prospects of meretopology. *Data and Knowledge Engineering*, 20:259–286, 1996.
6. P. Borst, J. Benjamins, B. Wielinga, and H. Akkermans. An Application of Ontology Construction. *Proc. ECAI*, pages 5–16, 1996.
7. U. Hahn, S. Schulz, and M. Romacker. Part-whole reasoning: A case study in medical ontology engineering. *IEEE Intelligent Systems*, 14(5):59–67, 1999.
8. N.F. Noy, M.A. Musen, J.L.V Mejino, and C. Rosse. Pushing the Envelope: Challenges in a Frame-Based Representation of Human Anatomy, 2002. Stanford Medical Informatics technical report SMI-2002-0925.pdf.
9. A. Artale, E. Franconi, N. Guarino, and L. Pazzi. Part-Whole Relations in Object-Centered Systems: An Overview. *Data and Knowledge Engineering*, 20:347–383, 1996.
10. HP Labs Jena Toolkit http://www.hpl.hp.com/semweb/jena.htm.
11. C. Schlenoff, M. Gruninger, F. Tissot, J. Valios, J. Lubell, and J.Lee. The process specification language (PSL) overview and version 1.0 specification, 2000. NIST Internal Report 6459.
12. D.L. McGuinness, R. Fikes, J. Hendler, and L.A. Stein. Daml+oil: An ontology language for the semantic web. *IEEE Intelligent Systems*, 17:72–80, 2002.
13. C.J. Wroe, R. Stevens, C.A. Goble, and M. Ashburner. A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. *Pacific Symposium on Biocomputing*, 8:624–635, 2003.