Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology

Christopher Ochs[1,*], Zhe He[1], Yehoshua Perl[1], Sivaram Arabandi[2], Michael Halper[1], and James Geller[1]

[1]NJIT, Newark, NJ 07102; [2]ONTOPRO, Houston, TX 77025

ABSTRACT

An *abstraction network* is a compact network summarizing the structure and content of a given ontology. Abstraction networks have been shown to support orientation into and quality assurance of ontologies. Area and partial-area taxonomies are examples of abstraction networks that utilize the relationships of an ontology to group together classes with similar structure and semantics. These taxonomies can be derived in different ways, leading to different granularities of summaries. Such granularity is illustrated by applying various derivation methodologies to the Sleep Domain Ontology (SDO), hosted on BioPortal. The impact of different granularity levels is demonstrated with respect to orientation into and quality assurance of the ontology's structure and content.

1 INTRODUCTION

To be usable, an ontology must sufficiently cover the knowledge of its domain, which implies the need for many classes and relationships. When an ontology grows to be large and complex, errors and inconsistencies become almost unavoidable. Hence, quality assurance (QA) is essential; however, QA can be a difficult and time-consuming manual process in the context of a large ontology.

To aid in comprehension of ontology content and to support quality-assurance efforts, we have utilized *abstraction networks*. An abstraction network (AN) is defined as a high-level support network used to summarize and visualize the content and structure of an ontology. Typically, ANs are derived using the knowledge contained within an ontology itself. An AN is composed of nodes and links organized into a hierarchy. Nodes are used to summarize groups of similar concepts, while links summarize the hierarchical relationships between those concepts.

Many biomedical ontologies are released in the Web Ontology Language (OWL) format. In this paper, we expand on our previous research by deriving multiple different ANs for the same OWL-based ontology. Each AN is derived using different ontological elements. We show that using different derivation methodologies lead to different abstraction-network *granularities*. Comparing these ANs leads to the identification of the best among them for QA efforts. Our test bed is the Sleep Domain Ontology (SDO) of 1390 classes, available in BioPortal in OWL format and focused on sleep medicine (Arabandi *et al.* 2010).

* For correspondence: Christopher Ochs (cro3@njit.edu)

The SDO is an ontology of 1,390 classes developed as part of the PhysioMIMI project to support the merging of physiological and clinical data (Arabandi *et al.* 2010). It is available within BioPortal (Whetzel *et al.* 2011). The SDO was built by merging knowledge from several ontologies, such as the Ontology for General Medical Science (OGMS) (Goldfain) and Foundational Model of Anatomy (FMA) (Rosse *et al.* 2003), with sleep-domain knowledge. SDO was selected as a sample OWL ontology with many object properties specified by restriction and only a few specified by domain and range, which was obtained by reusing other ontologies, causing integration errors.

2 BACKGROUND

In previous research, we have developed complementary ANs called *area taxonomies* and *partial-area taxonomies* for the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Stearns *et al.* 2001) and the National Cancer Institute thesaurus (NCIt) (Fragoso *et al.* 2004), both based on a similar (though not identical) ontological model. The taxonomies were shown to support QA efforts (Min *et al.* 2006; Wang *et al.* 2007). We have also developed a taxonomy derivation approach with respect to OWL-based ontologies that use similar definitional elements. This methodology was successfully applied for orientation and QA of the Ontology of Clinical Research (OCRe) (Ochs *et al.* 2012) to obtain *domain-defined taxonomies*.

Within OWL, object properties represent potential relationships between class instances ("individuals"). Object properties can be given explicitly defined domains and ranges. Consider the following example from SDO (shown using Manchester OWL syntax):

```
ObjectProperty: hasBodyPosition
       Domain: Patient
       Range: BodyPosition
```

Property *hasBodyPosition* has a domain consisting of one class *Patient* and a range consisting of one class *BodyPosition*. The meaning is: *any patient can have any kind of body position* (i.e., *sitting*, *standing*, or *recumbent*, in SDO). Within OWL, domains and ranges may consist of any number of classes. Object properties may or may not be reified within the ontology. All instances of the property will conform to the specified domain and range.

An *area* is defined as the set of all classes that are explicitly defined or inferred as being in exactly the domains of a given set of object properties *O*. The list of names of the object properties is used to name the area. Areas are

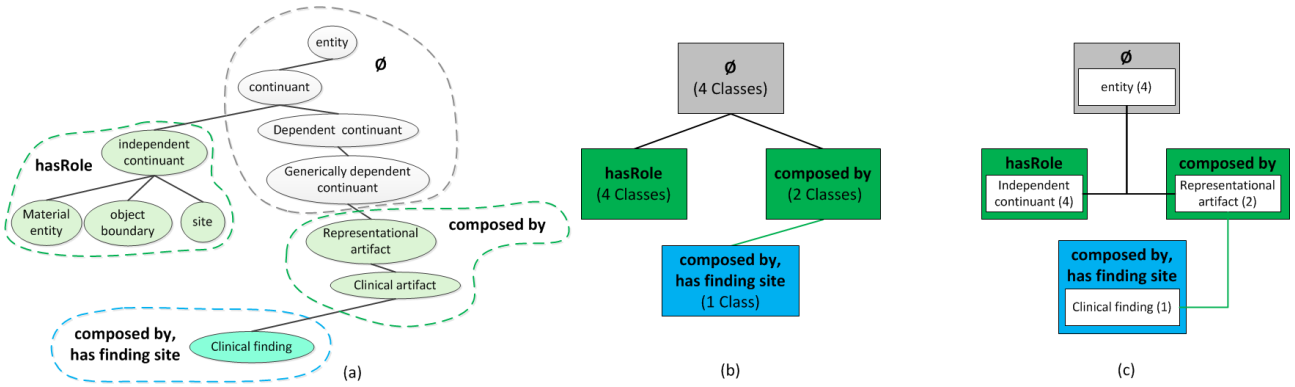

Fig. 1. (a) An excerpt of 11 classes and 3 object properties with defined domains taken from the Sleep Domain Ontology. **(b)** The domain-defined area taxonomy derived from the classes in (a)**. (c)** The domain-defined partial-area taxonomy derived from the classes in (a)**.**

connected by *child-of* links that are derived from the underlying ontology's class hierarchy. We define a *root* of an area as a class that has no parents in the same area (i.e., none of its parents share its set of object-property domains). An area may have more than one root. A root of an area defines a *partial-area*, which is the set of classes that includes the root and all its descendants in the area. As with areas, partial-areas are connected by *child-of* links derived from the underlying IS-A relationships. Specifically, partial-area *A* is *child-of* partial-area *B* if a parent of *A*'s root resides in *B*.

Figure 1(a) provides an excerpt of 11 classes taken from the SDO, along with three object properties, *hasRole*, *composed by*, and *has finding site*, having explicitly defined domains. Classes that are within the domain of a particular object property are shown in a dashed bubble. For example, *independent continuant* is explicitly in the domain of *hasRole*, while *material entity*, *object boundary*, and *site* are all implicitly in *hasRole*'s domain due to inheritance. Similarly, *Clinical finding* is explicitly defined as the domain of *has finding site*, but it also inherits the property *composed by* from its parent *Clinical artifact*.

Figure 1(b) shows the domain-defined area taxonomy for the excerpt in Figure 1(a). The four classes within the domain of *hasRole* are represented by the area with that name. *Child-of* links are shown as lines connecting the areas. Areas are organized into color-coded levels based on their numbers of object properties defined. Areas with a greater number of object properties are drawn lower down. Hence, the areas with the most complex classes of the hierarchy (those with the most object properties) will be at the bottom.

Figure 1(c) shows the domain-defined partial-area taxonomy for Figure 1(a). Partial-areas are represented using white boxes within areas and are labeled using their roots. The number of classes (including the root) in each partial-area is shown in parentheses. The lines are *child-of* links. Unlike in this example, typical areas contain more than one partial area.

3 DERIVING TAXONOMIES OF VARIOUS GRANULARITIES

We define the *abstraction ratio* of an AN to be the average number of ontology classes mapped to each AN node. This ratio indicates *granularity*. If there are few nodes in the AN (e.g., many classes are mapped to few nodes), we say that the AN has *coarse* granularity. Even though this AN summarizes the ontology, that summary may be too general for orientation and QA. Conversely, an AN's granularity may be considered too *fine* if it has too many nodes, meaning the benefits of the summary are effectively lost.

The granularity may be affected by the methodology used to derive the AN. Several different ANs can potentially be derived for the same ontology. What differs among the ANs is the algorithm used to define the nodes. Hence, granularity differences are expected. Finding the best AN for an ontology is based on the structure of the ontology and/or the purpose of AN use.

For the SDO, we first utilize our previously developed domain-defined taxonomy derivation methodology which was applied to OCRe (Ochs *et al.* 2012). Figure 2 shows the domain-defined partial-area taxonomy obtained from SDO's *Entity* hierarchy (1,275 classes). The taxonomy contains 13 partial-areas separated into an equal number of areas. The abstraction ratio is 98.08 (=1,275/13) classes per partial-area. Three partial-areas, *Entity*, *Representational artifact*, and *Independent continuant*, together constitute nearly the entire hierarchy (1,217 classes). The ten other areas together contain only 58 classes, 25 of which are in the partial-area *Procedure*. Hence, the granularity of the top part of the taxonomy is too coarse for either orientation into the classes of those areas or for QA since it over-summarizes the content.

Domain-defined taxonomies will only provide sufficient granularity when enough object properties have explicitly defined domains. Within the SDO's *Entity* hierarchy only 16 of the 50 object properties have explicitly defined domains. Therefore, we have to use other ontological knowledge to derive taxonomies, namely, object properties used as restrictions.

To overcome the lack of granularity, we define two new abstraction-network derivation methodologies leading to *restriction-defined taxonomies* and (*domain* ∪ *restriction*)*-defined taxonomies*. These use different structural knowledge of the ontology, and each redefines how areas

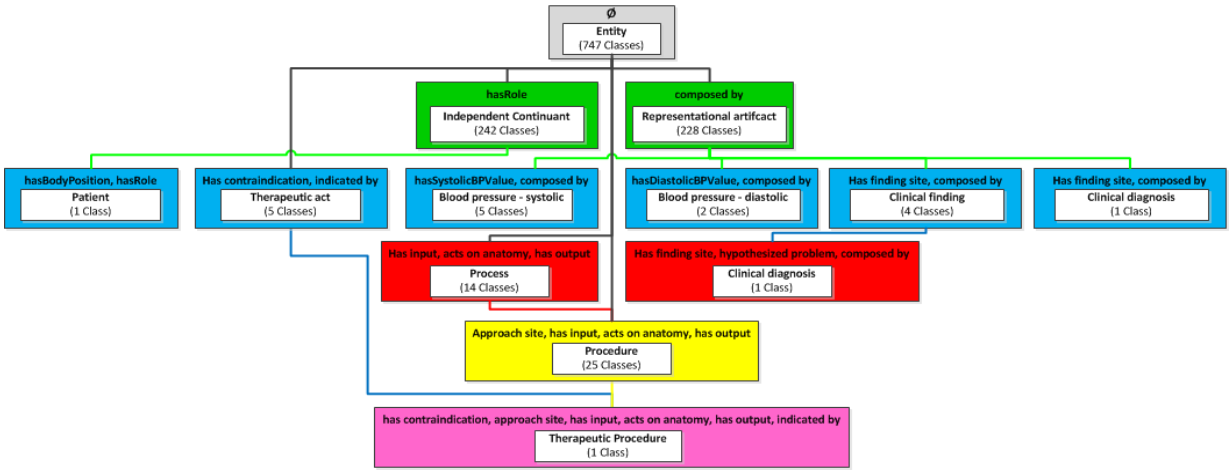

Fig. 2. Domain-defined partial-area taxonomy for the SDO's *Entity* hierarchy.

are derived. The partial-area derivation methodology remains unchanged, but the resulting two corresponding partial-area taxonomies can be different due to the different object properties used to define the partial-areas.

By comparing ANs of various granularities for the same ontology, one can choose the AN that best fits an application such as QA. An AN with too coarse a granularity does not offer many options with respect to QA. One of finer granularity does, but if the AN is too detailed and therefore not compact, the benefits of using the AN will be lost, as it will become similar in size to the ontology itself.

3.1 Restriction-defined Taxonomies

Many ontologies (such as the SDO) do not rigorously define domains and ranges for every object property. OWL allows object properties to be used in restrictions on classes. The major difference from specifying domains and ranges is that a restriction is *local*, i.e., the restriction only applies within the context of the given owl:Restriction. As an example, consider the following:

```
Class: BilateralUpperLimbMovementDuringSleep
   SubClassOf:
      UpperLimbMovementDuringSleep
      includes some
            RightUpperLimbMovementDuringSleep
      includes some
            LeftUpperLimbMovementDuringSleep
```

This states that the class *BilateralUpperLimbMovementDuringSleep* is defined in terms of two restrictions that utilize the object property *includes*. One restriction is that *BilateralUpperLimbMovementDuringSleep* includes *RightUpperLimbMovementDuringSleep*; the second is that it includes *LeftUpperLimbMovementDuringSleep*. Both restrictions use the constraint *some*, which requires that at least one instance of the object property used with *BilateralUpperLimbMovementDuringSleep* conform to the restriction. An alternative would be *all*, which means when the object property *includes* is used with *BilateralUpperLimbMovementDuringSleep*, all instances must conform to the restriction. Using an object property as a restriction allows for more flexibility. *Includes* is a high-level property used in 82 different restrictions in SDO.

Taxonomies can be derived using the defined restrictions when there are a sufficient number of them, yielding *restriction-defined taxonomies*. The SDO has 44 object properties used in restrictions on classes, making them a viable choice for defining taxonomies with finer granularity.

In this context, we redefine the area taxonomic element as follows. In a restriction-defined taxonomy, we define an area to be the set of classes that are explicitly defined or inferred to be bound by restrictions that use the object properties. A restriction can be either allValuesFrom or someValuesFrom; the methodology does not distinguish between the two. *Child-of* links are derived as with the domain-defined partial-area taxonomy. Essentially, we are treating the class that has the restriction as belonging to the domain of the object property. Additionally, any descendants of the class with the restriction are considered to be implicitly in the object property's domain.

3.2 (Domain ∪ Restriction)-defined Taxonomies

If taxonomies with higher granularity than the domain-defined and the restriction-defined taxonomies are desired, they can be derived using both object properties with explicitly defined domains and with restrictions together. We refer to these as (*domain* ∪ *restriction*)-*defined taxonomies*, with the set union symbol "∪" denoting that the object properties can be either of the two varieties.

In the (domain ∪ restriction)-defined taxonomy, an area is defined as the set of classes that are defined or inferred to belong to the same set *O* of object-property domains *or* are used in the same set of restrictions. Of the three partial-area taxonomies, the (domain ∪ restriction)-defined taxonomies have the finest granularity because they combine the knowledge used to derive both the domain-defined taxonomies and the restriction-defined taxonomies.

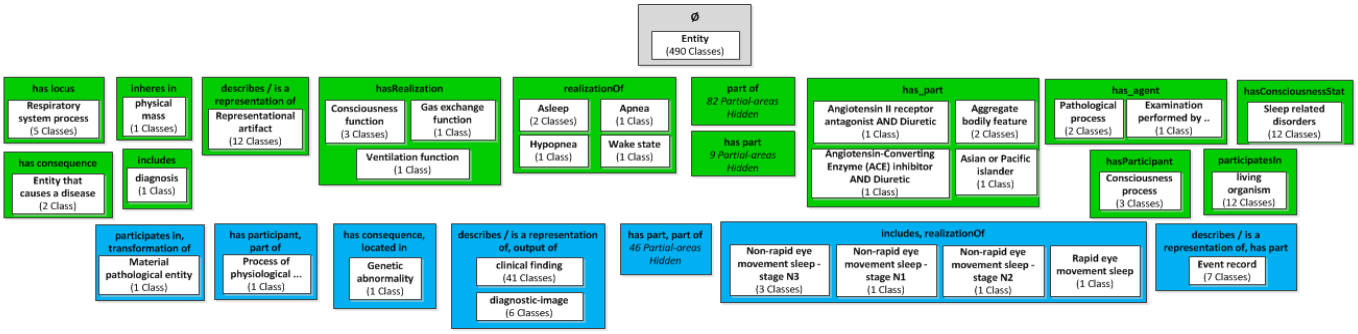

Fig. 3. The first three levels of the Restriction-defined Partial-area Taxonomy for SDO's Entity hierarchy.

4 THE SDO'S REFINED TAXONOMIES

Our goal with this research is to provide summaries of ontologies that assist with orientation into the content and structure of those ontologies, as well as support QA. Generally, partial-area taxonomies are used for QA purposes as they summarize both structure and semantics. Therefore, we derive both a restriction-defined partial-area taxonomy and a (domain ∪ restriction)-defined partial-area taxonomy for SDO's *Entity* hierarchy, having 50 object properties (16 with defined domains and 44 used as restrictions, with 10 properties used in both ways). We utilized our Biomedical Layout Utility for OWL ("BLUOWL"), an early prototype tool, for derivation and display. BLUOWL is based on our BLUSNO utility, used previously for SNOMED taxonomies (Geller *et al.* 2012). Due to lack of space, the complete figure for *Entity*'s restriction-defined partial-area taxonomy is given in (http://cs.njit.edu/~oohvr/SABOC/figures.php). The top three levels of this taxonomy are shown in Figure 3.

This partial-area taxonomy has a significantly finer granularity than the domain-defined one shown in Figure 2. The taxonomy is composed of 262 partial-areas within 61 areas organized into 12 levels. The abstraction ratio is 4.87 (=1,275/262) classes per partial-area. Within the taxonomy, 25 areas are multi-rooted. The areas with the most partial-areas are {*has part*} (82 partial-areas) and {*has part, part of*} (46 partial-areas), with all of their classes coming from the FMA ontology.

Figure 4 shows the entire 13-level (domain ∪ restriction)-defined partial-area taxonomy for the *Entity* hierarchy. There are 267 partial-areas organized into 67 areas, 25 of which are multi-rooted. For readability, *child-of* links between partial-areas are not displayed. The granularity is slightly finer than for the previous taxonomy. The abstraction ratio is 4.78 (=1,275/267).

5 DISCUSSION

Two applications of taxonomies and other abstraction networks are their support for ontology QA and orientation into the ontology's content. We will illustrate these two applications in the context of the (domain ∪ restriction)-defined taxonomy of Figure 4. Orientation concentrates on identifying large groups of structurally and semantically similar classes, which can be achieved by reviewing all large partial-areas with, say, 10–50 classes: *Clinical finding* (41), *Apnea hypopnea index* (30), *Polysomography (PSG) finding (21), Sleep-related disorder* (12), and *Representational arti fact (12)*. For a more refined orientation, one may view the 14 medium-sized partial-areas with 5–10 classes, e.g., *Sleep hypopnea finding* (8) and *Respiratory flow amplitude finding* (6). Together, the combined list of 19 (5+14) partial-areas gives a summary of the main kinds of classes and their frequency in the SDO, beyond the two very large areas containing the partial-areas *Entity* and *Independent continuant*, and the three very large areas {*part of*}, {*has part*}, and {*has part, part of*}.

The 128 partial-areas outside of the above five large areas cover 354 classes. The 19 medium and large partial-areas, with 5-50 classes, accounting for 204 (58%) of these 345 classes, provide orientation into the content of the SDO by highlighting important sets of concepts.

For the SDO, the domain-defined taxonomy (Figure 2) has too coarse a granularity; most classes were mapped to three nodes. The restriction-defined taxonomy R is much more refined. This is illustrated by the partial-area *representational artifact* (228) in Figure 2 broken in R into many partial-areas, leaving only 12 classes in this partial-area (see Level 1 of Figure 3). Also, anatomy classes imported from the FMA are moved to the areas {*part of*}, {*has part*} and {*part of, has part*} of R since these two object properties are used in restrictions (see Figure 3).

The inclusion of object properties with explicitly defined domains in the (domain ∪ restriction)-defined partial-area taxonomy DR in Figure 4 resulted in several structural changes from the restriction-defined partial-area taxonomy R. The most significant change is that many partial-areas of R moved down to a lower level in Figure 4, due to the introduction or inheritance of object properties with explicitly defined domains. *Representational artifact* (*12*), for example, moved from Level 1 of R to Level 2 of DR due to the display of the object property *composed by* which has an explicitly defined domain. Altogether, 97 partial-areas (out of 262) containing 297 classes (out of 1275) were moved down to lower levels, in the transition from the taxonomy R to the taxonomy DR in Figure 4.

Considering the issue of which AN will be best for QA purposes, we realize that the abstraction ratios of the ANs R and DR are almost equal, so compactness is not a determining factor. Having the full set of object properties for the 97

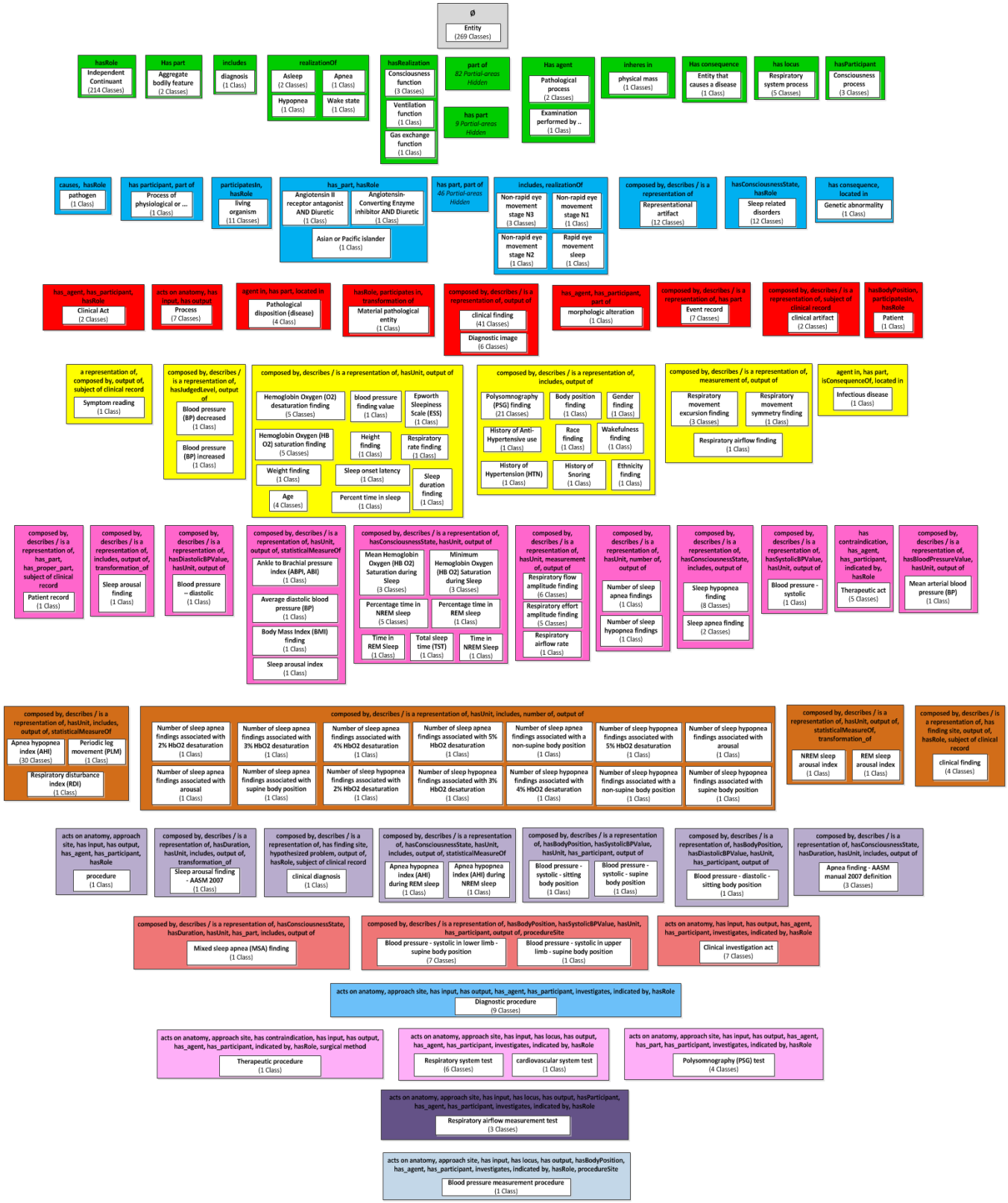

Fig. 4. The entire (domain ∪ restriction)-defined partial-area taxonomy for SDO's *Entity* hierarchy.

partial-areas that moved one or more levels down when going from the taxonomy R to the taxonomy DR led us to prefer the taxonomy DR for QA. That taxonomy is a more complete and accurate summary of the ontology's structure, enabling a more detailed QA review as demonstrated below.

Our initial scan of SDO identified 12 candidates for further QA analysis. A few patterns including dissimilar partial area groupings and duplicate object properties were noted within these candidates.

The first, a dissimilar partial area grouping, was noticed at the third (blue) level where the area {*has_part, hasRole*} has three partial-areas, one of which (*Asian or Pacific Islander*) does not match the other two partial-areas about *Angiotensin*. They fall under very different hierarchies – the first one is a subclass of population, and the other two are classes under the medication hierarchy. Upon investigation, the class *Asian or Pacific Islander* was introduced for cases where the records did not distinguish between the two races and the actual race is not known. The semantics of such a situation fits the OR logical operator, as the term describes, and does not fit a part relation (Winston *et al.* 1987). An individual of this race is not part Asian and part Pacific Islander, but is one of the two. We just do not have the knowledge. Thus, the *has_part* object property is removed from this class and it will be in the *Independent continuant* partial-area in the {*hasRole*} area like all of its sibling races.

We note that this modeling error was discovered only due to the dissimilar grouping in the area {*hasRole, has_part*} of DR. This area is defined by the object property *hasRole* with an explicit domain and the object property *has_part*, which is used in a restriction. Hence, this area does not appear in the domain-defined partial-area taxonomy of Figure 2. Neither does it appear in the restriction-defined partial-area taxonomy of Figure 3. The only taxonomy where this dissimilarity appears is in the (domain ∪ restriction)-defined partial-area taxonomy of Figure 4. This example demonstrates why we have chosen this taxonomy for performing the QA of SDO, due to its full display of all object properties.

The classes in the partial-area *living organism* (level 3) were found to have duplicate properties – "*participatesIn*" (from BioTop) and "*participates in"* (from RO). On examination, neither of these two properties has a description associated. However, based on the usage of these properties, it appears that the two are equivalent. Neither property has a domain or range specified, but the RO version has a subproperty and an inverse property associated with it. The BioTop version of the property is used only once (in the definition of living organism). Therefore, SDO was refactored to replace this relation with the one from RO.

These findings are the result of an initial examination of SDO. It is important to note that these are not 'issues' or 'errors,' but candidates that require further analysis. These are currently being examined and will form the basis for follow-up research.

CONCLUSIONS

We introduced two new derivation methodologies for area and partial-area taxonomies, resulting in the restriction-defined taxonomies and the (domain ∪ restriction)-defined taxonomies. These methodologies can be applied to OWL-based ontologies that have object properties used in restrictions on classes in addition to object properties with defined domains. Using different taxonomy derivation methodologies on the SDO results in taxonomies of differing granularity. The taxonomies of finer granularity can be used to provide summaries of and aid orientation into the SDO's content and structure as well as support QA of the SDO, as demonstrated in this paper.

ACKNOWLEDGMENTS

We would like to thank Drs. Natasha Noy and Mark Musen of the NCBO for their continued guidance.

REFERENCES

"Sleep Domain Ontology." Retrieved March 4, 2012, from http://bioportal.bioontology.org/ontologies/1651.

Arabandi, S., *et al.* (2010). "Developing a Sleep Domain Ontology." AMIA Clinical Research Informatics Summit.

Fragoso, G., *et al.* (2004). "Overview and utilization of the NCI thesaurus." Comp Funct Genomics **5**(8): 648-654.

Geller, J., *et al.* (2012). "New Abstraction Networks and a New Visualization Tool in Support of Auditing the SNOMED CT Content." AMIA Annu Symp Proc: 237-246.

Goldfain, A. "Ontology for General Medical Science (OGMS)." Retrieved January 8, 2013, 2013, from http://code.google.com/p/ogms/.

Min, H., *et al.* (2006). "Auditing as part of the terminology design life cycle." J Am Med Inform Assoc **13**(6): 676-690.

Ochs, C., *et al.* (2012). "Deriving an Abstraction Network to Support Quality Assurance in OCRe." AMIA Annu Symp Proc: 681-689.

Rosse, C., *et al.* (2003). "A reference ontology for biomedical informatics: the Foundational Model of Anatomy." J Biomed Inform **36**(6): 478-500.

Stearns, M. Q., *et al.* (2001). "SNOMED clinical terms: overview of the development process and project status." Proc AMIA Symp: 662-666.

Wang, Y., *et al.* (2007). "Structural methodologies for auditing SNOMED." J Biomed Inform **40**(5): 561-581.

Whetzel, P. L., *et al.* (2011). "BioPortal: Enhanced Functionality via New Web services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications." Nucleic Acids Research (NAR) **39**(Web Server issue): W541-545.

Winston, M. E., *et al.* (1987). "A taxonomy of part-whole relations." Cognitive science **11**(4): 417-444.