Abstract. Biomedical terminologies lack a clear-cut understanding of the semantics of part-whole relations. In order to design a conceptually adequate ontology for the biomedical domain, we investigate instead conceptually simpler though expressive spatial inclusion/containment relations. We outline a multi-faceted encoding pattern for biological concepts, which is centered around these relations.

1 Introduction

Tangible parts of organisms such as organs, tissues, cells, molecules constitute the location of biological processes and are the targets of experimental, diagnostic and therapeutic interventions. This notion of ‘biological structure’ plays a key role for ontologies in the life sciences [6, 16, 17, 9]. Many of them encode terminological knowledge in an informal, thesaurus-style way [12] such as the source vocabularies of the UMLS metathesaurus [18]. One of the few exceptions is the ‘Foundational Model of Anatomy’ (FMA) [17, 14], whose axiomatic basis is just going to be formalized [24].

None of the already mentioned conceptual models of human anatomy or biological structure make sufficient claims regarding these three controversial issues, mainly because they neglect, by and large, the dichotomy between concept-to-concept relations (Rel) and instance-to-instance relations (rel). Both FMA and GO have a clear commitment to item 1., at least regarding the transitivity property of the general part-of relation. The same applies to GALEN with regard to part-of subrelations [16]. A rudimentary commitment to item 2. is reflected by the introduction of several part-of subrelations in FMA and GALEN. Concerning the interpretation of part-of between

1 http://obo.sourceforge.net
2 http://www.informatics.jax.org
   http://www.flybase.org
   http://www.gramene.org

characterized by taxonomic (is-a) and partonomic (part-of) relations. While taxonomic hierarchies and their inherent reasoning patterns are quite well understood [5], we lack an equal form of consensus for part/whole-related reasoning [1].

With regard to the partonomic hierarchies in biomedical ontologies, we identify the following shortcomings:

1. There is no clear commitment to the algebraic foundation of mereological relations in terms of transitivity, reflexivity and symmetry.
2. The semantics often lacks stable descriptional commitments, i.e., it is largely a matter of taste of the developers of such terminologies whether parts and wholes are to be understood in a functional or in a topological sense.
3. While mereology requires part-whole relationships to hold between individual entities, OBO ontologies, but also FMA and UMLS establish Part-Of and Has-Part5 relationships between concepts, i.e., classes of individuals.2 A statement such as A Rel B (relating two classes, e.g., Cell Has-Part Axon, in GO) leaves unclear whether B should be interpreted (a) as some kind of range restriction of the filler of rel (the “obligatory parts” reading), or (b) as an existential condition for A with respect to rel (the “possible parts” reading). In addition, we have to make clear (i) whether Has-Part is the inverse relation of Part-Of or (ii) whether it is not.

None of the already mentioned conceptual models of human anatomy or biological structure make sufficient claims regarding these three controversial issues, mainly because they neglect, by and large, the dichotomy between concept-to-concept relations (Rel) and instance-to-instance relations (rel). Both FMA and GO have a clear commitment to item 1., at least regarding the transitivity property of the general part-of relation. The same applies to GALEN with regard to part-of subrelations [16]. A rudimentary commitment to item 2. is reflected by the introduction of several part-of subrelations in FMA and GALEN. Concerning the interpretation of part-of between
classes, the GALEN model adopted 3b(ii) (implicitly, because there are no explicit inverse relations), while FMA recently subscribed to 3b(i) [24] and GO adopted the interpretation 3a. The UMLS metasaurus may be interpreted as 3a or 3b(i).

This brief review of biomedical knowledge organization may illustrate the need to convene upon clear epistemological and semantic assumptions as an indispensable premise for the development, maintenance, and fusion of computationally usable ontologies for natural kinds [15]. We here provide a formal characterization of these relations in order to promote a clearer understanding of the conceptual foundations underlying the natural sciences.

2 Considering Mereological Relations

In the following we reconsider the three controversial issues from Section 1 and articulate our commitments, which are intended to reflect the particularities of ontologies in the biomedical domain.

2.1 Epistemic Status and Algebraic Properties

In classical (axiomatic) mereology [7, 23], generic parthood (between individual objects) is treated as reflexive, antisymmetric and transitive. The transitivity of part-of, however, has been refuted by linguists and cognitive scientists [8, 25], who distinguish several part-whole relation types (e.g., component-of, member-of) with diverging relation properties, e.g., non-transitivity. Yet, consensus has been achieved that parthood in its broadest, i.e., most general, sense must be transitive, indeed.

2.2 Parthood vs. Location

Even restricting the use of mereological relations to the basic (most general) relation pair [part-of, has-part], with has-part = part-of\(^{-1}\), there is still a broad range of interpretation between a mainly functional and a strictly topological account. From a functional point of view, a liver tumor may not be viewed as being part of the liver, because it does not contribute to its function. From a topological view, it doubtlessly occupies a part of the liver. Non-topological views give leeway to discretionary conceptual modeling decisions. For instance, one could argue that a primary liver tumor is part of the liver (because it consists of degenerated liver cells), but a liver metastasis is not (because the cells have migrated into the liver but emanate from another organ or tissue). One could also contend that the oral cavity is not part of the head (but is part of the exterior space), but what about the lumina of the esophagus, the bronchi, the gastric glands? If a bacterium is in a cell, it is certainly not part of the cell. At which stage of the process of digestion (phagocytosis) do bacteria components become parts of the ingesting cell? Moving on to biological tissues, are the red blood cells in the liver capillaries part of the liver? Are the immunocompetent cells which have migrated from the blood into the connective tissue part of the connective tissue? An other example is the so-called endosymbiotic hypothesis which assumes that semi-autonomous cell components such as mitochondria and chloroplasts were originally endosymbiotic bacteria. This means that during evolution their ontological status has gradually changed from topological includees to proper cell parts.

We postulate that only a strictly topological interpretation avoids conflicting views. In [19] we argued to conceptualize biological objects based on their convex hull, because otherwise the cavernous character of most biological objects would permeate each millimeter of biological matter by extensions of the exterior space – which contradicts common conceptualizations. In [22] we refined this approach and proposed to drop entirely the notion of strict parthood in favor of a mereotopological view and so render a less controversial, yet still useful description of the physical structure of natural kinds.

To underline this shift in perspective, we generalize has-part to the relation includer-of (with the meaning of “includes”, “location of”) and part-of to containee-of, (with the meaning of “contained in”, “has location”) abbreviated as i and c, respectively, with i = c\(^{-1}\). We define i and c as being transitive, reflexive and antisymmetric.

2.3 Relations between Concepts

Both i and c establish relations between individual objects in the same way as the mereological relations part-of and has-part do. However, there seems to be a need for expressing mereotopological conceptual interrelationships in the biomedical domain by simple two-valued predicates which hold between concept classes as well. Therefore, we introduce the following concept-to-concept relations, in a similar way as proposed by Smith & Rosse [24]. The irreflexive, non-transitive and antisymmetric relation i (abbreviation for instance-of) represents class membership, i.e., it relates an instance with a corresponding class. On this basis, we define class inclusion via IS-A, a reflexive, transitive, and antisymmetric relation between concept classes A and B, as follows:

\[ IS\cdot A(A, B) \iff \forall x : (i(x, A) \rightarrow i(x, B)) \]  

We use the following notation: Concept (universal, class) variables are represented by upper-case letters, instances (particulars, individuals) by lower-case letters. Names of relations between concept classes have an upper-case initial letter, names of relations between instances are in lower case. We define the concept-to-concept relations SC (specific containee) and SI (specific includer) as follows:

\[ SC(A, B) \iff \exists y : i(y, B) \land c(x, y) \] (2)

\[ SI(B, A) \iff \forall y : i(y, B) \land i(y, x) \land i(x, A) \] (3)

If we want to express unilateral dependence, e.g., that cells exist with and without cell nuclei, we state: SC(Cell\{Nucleus\}, Cell), but not SI(Cell, Cell\{Nucleus\}). As cells include enzymes, we state SI(Cell, Enzyme), but as enzymes also occur outside of cells, we do not state SC(Enzyme, Cell). As the example suggests, the two definitions are not equivalent, i.e. (3) is not just a special case of (2).

We formally prove this stipulation by counterexample. Let

\[ i(x_1, A) \land i(x_2, A) \land \forall(x) : x = x_1 \lor x = x_2 \lor \neg i(x, A) \] (4)

\[ i(y_1, B) \land i(y_2, B) \land \forall(y) : y = y_1 \lor y = y_2 \lor \neg i(y, B) \] (5)

i.e., the class A only has the instances x1 and x2, the class B only has y1 and y2 as instances. Let c(x1, y1) and c(x2, y2). Then SC(A, B) holds for these assignments. Apart from reflexivity, let these be the only cases in which c holds, i.e.,

\[ \forall(x) : x = x_1 \lor x = x_2 \lor (x \neq x_1 \land x \neq x_2 \land \forall(y) : \neg c(x, y)) \] (6)

Since i = c\(^{-1}\), we can rewrite (3):

\[ \forall(y) : i(y, B) \rightarrow \exists(x) : i(x, A) \land c(x, y) \] (7)

We instantiate y with y2. Since there is no x for which c(x, y2) holds, the right-hand side of (7) is false for these assignments. Therefore, \neg SI(B, A). Hence (3) is not a special case of (2) q.e.d. (the inverse
can be shown analogously). Thus, the two definitions are indeed distinct.

Furthermore, from the combination of expression (1) with (2) and (3) we deduce that the relations $SC$ and $SI$ propagate via $Is-A$:

$$SC(A, B) \land Is-A(B, C) \rightarrow SC(A, C)$$

$$SI(B, A) \land Is-A(A, D) \rightarrow SI(B, D)$$

We introduce the inverse relations of both $SC$ and $SI$ as follows:

$$OC(A, B) \leftarrow_{def} SI(B, A)$$

$$OI(B, A) \leftarrow_{def} SC(A, B)$$

OC can be paraphrased as is-obligatory-containee-for, and $OI$ as is-obligatory-includer-for. Note that the propagation via $Is-A$ does not hold for these relations. As an example, Hemoglobin is an obligatory containee of Red Blood Cell, but not of its parent Cell. Instead, these relations propagate in the opposite direction, as a simple consequence of being the converse relations. In our example, Cell Membrane is an obligatory containee of Cell, therefore it is also an obligatory containee of any subclass of Cell, such as Red Blood Cell.

$SC$ and $SI$ are both transitive. From Formula (2), together with

$$SC(B, C) \leftarrow_{def} \forall y : t(y, B) \rightarrow \exists z : t(z, C) \land c(y, z)$$

we get:

$$\forall x : t(x, A) \rightarrow \exists y : t(y, B) \land t(z, C) \land c(x, y) \land c(y, z)$$

Because of the transitivity of $c$ we infer:

$$\forall x : t(x, A) \rightarrow \exists z : t(z, C) \land c(x, z)$$

which is equivalent to the expression $SC(A, C)$. In an analogous way we can prove the transitivity of $SI$. As a consequence, the inverse relations $OC$ and $OI$ are transitive as well. Fig. 1 depicts an example.

![Concept-to-Concept relations. Example Cell Nucleus.](image)

### 3 Includer/Containee Extension of Taxonomies

In the following, we demonstrate how the proposed solution can be encoded as an extended taxonomy. In order to avoid concept-to-concept relations $R(A, B)$ we introduce the following reificator nodes

$$Is-A(B, ASC) :=_{def} SC(B, A)$$

$$Is-A(B, ASI) :=_{def} SI(B, A)$$

These additional nodes can be described as follows (cf. Fig. 1, in which $A$ corresponds to Cell Nucleus):

- $ASC$ (specific containee nodes of $A$) subsumes those classes $A, B_1, B_2, \ldots, B_n$ whose instances must be related to an instance of $A$ by the relation $c$. Here, the (SC-)containees depend on the includer, or the includer is obligatory for (SC-)containees to exist. As an example, Nucleolus is a specific containee of Cell Nucleus (it occurs only there). It is, however, not an obligatory containee, because cell nuclei without nucleoli exist.

- $ASI$ (specific includer node of $A$) subsumes those classes $A, B_1, B_2, \ldots, B_n$ whose instances must be related to an instance of $A$ by the relation $i$. Here, the (SI-)includers depend on the containee, or the containee is obligatory for the (SI-)includers. As an example, Tissue is a specific includer of Cell Nucleus, because there are cell nuclei in any instance of tissue.

The problems stated in the previous section are addressed by concept encoding patterns which we have introduced previously [11, 21, 20]. It makes taxonomic reasoning immediately available for mereotopological reasoning. One reason for this is that classification-based reasoning is well-understood and powerful implementations (e.g., based on description logics [2]) exist.

As an example, we construct a containment hierarchy of the classes $A, B, C$, in which $A$ is a specific containee of $B$, and $B$ is a specific containee of $C$. We define $C$ as a specific includer of $B$, and the latter as a specific includer of $A$. We emulate this hierarchy by a taxonomy: For each of the classes $A, B, C$, we first introduce two reificator nodes for the relations $i$ and $c$:

$$\forall x : t(x, ASC) \rightarrow \exists y : t(y, A) \land c(x, y)$$

$$\forall x : t(x, ASI) \rightarrow \exists y : t(y, A) \land i(x, y)$$

$$\forall x : t(x, BSC) \rightarrow \exists y : t(y, B) \land c(x, y)$$

$$\forall x : t(x, BS1) \rightarrow \exists y : t(y, B) \land i(x, y)$$

$$\forall x : t(x, CSC) \rightarrow \exists y : t(y, C) \land c(x, y)$$

$$\forall x : t(x, CS1) \rightarrow \exists y : t(y, C) \land i(x, y)$$

$C_{ASC}$ subsumes all specific containees of $C$ (i.e., those which have $C$ as an obligatory includer). $C_{ASI}$ subsumes all specific includers of $C$, i.e., those concepts which have $C$ as a obligatory containee. Cascading subsumption of classes emulates transitivity of $c$ and $i$ even for languages which do not support the transitivity of relations (e.g., the standard description logics ALCH [3]). The same applies to $i$: In this case, the $SI$ node of a class is subsumed by $SI$ nodes of its obligatory containees.

$$Is-A(A, ASI)$$

$$Is-A(A, ASC)$$

$$Is-A(B, BSI)$$

$$Is-A(B, BSC)$$

$$Is-A(A, CS1)$$

$$Is-A(A, CSI)$$

In this manner the following properties are propagated via taxonomic subsumption: $A$ inherits the properties (19) and (21), $B$ inherits (21) and (18), and $C$ inherits (20) and (18). The resulting complex

$^{6}$ Ontological dependence in the sense of generic dependence [23].
graph is still acyclic with respect to the Is-A relations which emulate i and c hierarchies, given that there are no cycles in either of these subgraphs. Note that the situation given by these examples is highly idealized in the sense that both branches, c and i, are present throughout the hierarchy. According to our proof in Section 2.3, Is-A(A, B_{SC}) does not imply Is-A(B, A_{SI}), which is widely reflected by numerous examples in biology and medicine as argued above.

Using the auxiliary nodes SC and SI for each class, it is not yet possible to directly address obligatory containees (OC) and obligatory includers (OI). There are cases in which this is useful, particularly for concept definitions in which role values have to be restricted. For a more thorough analysis in a description logics context cf. [10].

Using the SC and SI auxiliary nodes, we make the following stipulations:

- **Reflexivity**: Each class A is subsumed by its corresponding A_{SC} and A_{SI} nodes as a consequence of the reflexivity we assume to hold for c and i.
- **Taxonomic Subsumption**: For each class A, which is subsumed by a superclass B, B_{SC} subsumes A_{SC}, and B_{SI} subsumes A_{SI}.
- **Mereotopological Hierarchies as Taxonomies**: For classes in an includer-containee hierarchy (I for includer, C for containee), we subsume C_{SC} by I_{SC} in order to express the dependence of the containee class on the includer class. In an analogous way, we subsume I_{SI} by C_{SI} in order to express the ontological dependence of the includer class on the containee class.

The flexibility of this approach is shown in Fig. 2:

(i) A Eukaryotic Cell is a Cell;
(ii) A Cell is a specific includer of Water, but Water is not a specific containee of Cell;

(iii) A Cell Nucleus is a specific containee of an Eukaryotic Cell;
(iv) Eukaryotic Cells without Cell Nuclei may exist;
(v) A Cell Nucleus obligatorily includes a Nuclear Membrane;
(vi) A Nuclear Membrane is obligatorily a containee of a Cell Nucleus;
(vii) A Nuclear Membrane is an Intracellular Membrane.

In this graph, each square depicts the domain concept class as a white node in the center, together with the auxiliary nodes SC and SI in the corners. All arrows represent Is-A relations. The thin arrows within the square visualize the taxonomic links between each domain concept class and its SI and SC nodes.

(i) is reflected by the subsumption of E by C, E_{SC} by C_{SC}, and E_{SI} by C_{SI}. Due to (ii), there is only one Is-A link from C_{SI} to W_{SI}, (iii) is also reflected by one Is-A link, but here involving the SC node, and (iv) is expressed by the absence of links between the SI nodes. The mutual dependence (v) and (vi) is reflected by Is-A links between both auxiliary nodes. Finally (vii) - the subsumption of Nuclear Membrane by Intracellular Membrane is expressed in the same way as done with Eukaryotic Cell and Cell. The resulting "circuit diagram" allows includer(containee) related inferences such as Is-A(N, I_{SI}), Is-A(M, C_{SC}), Is-A(E, W_{SI}), etc.

Finally we will analyze how "possible" associations can be expressed with our coding scheme. As we, intentionally, do not use modal logics we cannot express this notion by a specific operator. Depending on the language in which we implement our ontology, e.g. description logics, we subscribe to an open-world assumption. This means that the absence of information is interpreted as lack of knowledge, but not as negative indication (such as in databases) [3]. Given two classes A and B, we can therefore relate any pair of instances a and b (a ∈ A; b ∈ B) by a relation such as i or c as long this does not violate existing constraints. This is a way in which we can express the notion of "possible" relationships as already discussed considering the meaning of "part-of" in the Gene Ontology. In doing so, the knowledge engineer has to ensure a partitioning of the ontology in a way such that it obviates unwanted conclusions. For instance, we may want to express that the classes A and B are inclusionally disjoint (ID), i.e., there is no object in the world which is included in any instance of both A and B:

\[ ∀ x, y : i(x, A) ∧ i(y, B) → ¬∃ z : c(z, x) ∧ c(z, y) \]  

Using this formalism in order to declare SC nodes as mutually exclusive, inclusional disjointness can be expressed on a high level, and is then propagated through all subsumed nodes. As an example, we would state: ID(Trunk, Head). Given Is-A(Thorax, Trunk_{SC}) together with Is-A(Brain, Heads_{SC}) the assertion i(b, Brain) ∧ i(t, Thorax) ∧ i(t, Brain) ∧ i(t, Heads) would then be inconsistent.

4 CONCLUSIONS

We proposed a semantically explicit representation for the compositional structures of spatially relevant entities in bio-ontologies. By stipulating artificial concepts we reify the mereotopological relations includer-of and container-of, thus obtaining a domain representation as an extended taxonomy.

In order to get a realistic estimation of the size dimension of such encoding patterns, we calculated the number of nodes and edges for the 2003 UMLS [18] release of the Foundational Model of Anatomy with 53,400 concepts, related by 110,446 taxonomic and 37,388 mereologic links. The latter ones are, in their entirety, bilateral Part-Of / Has-Part concept-to-concept relations. As a consequence, each
of the original taxonomic links is represented by three Is-A links, each mereological association by two Is-A links, and each domain concept is represented by three nodes and two Is-A links. The result is a lattice of 160,200 nodes, woven together by 331,338 + 74,776 + 106,800 = 512,914 Is-A links. The numbers implied by this encoding are prohibitive for any attempt of manual knowledge editing. The use of an intermediate representation format for the knowledge engineer is therefore mandatory. For this purpose we use Protégé [13], together with appropriate export routines that construct the complex hierarchies [4].

In the context of biomedicine, one should emphasize that this apparently huge domain model still constitutes a lower growth bound, because it comprises "only" the rough structure of the human body, excluding molecular, functional and dysfunctional aspects. The size of the UMLS metathesaurus – which has recently gone through the roof of one million domain concepts – points at the real size dimension advanced reasoning systems will be exposed to whenever a serious biomedical application is considered. This is also the main reason why we diverge from the conclusions drawn in many theoretical studies dealing with the proper representation of parts and wholes [1]. We, intentionally, do not give mereo(topological) knowledge a dedicated 'first-class citizen' status at the representation level because this would require to implement entirely new types of reasoning engines ("partifiers" rather than just classifiers). Given the huge size of biomedical knowledge bases, we have to rely on really robust reasoning engines, currently only available as terminological classifiers (if at all). Therefore, we chose to encode partonomies as taxonomies. Therefore, we subscribed to parsimonious variants of description logics rather to much more expressive ones (e.g., including modal extensions) and chose taxonomic subsumption as a simple reasoning mode.

References


