An Upper-Level Ontology for Chemistry

Colin BATCHELOR \(1\)

\(a\) Royal Society of Chemistry, Cambridge, UK

Abstract. Chemical entities are the foundation of biochemistry and biology, but until now there have been few coherent attempts to produce a top-level ontology for chemistry to connect ontological descriptions of reality at the molecular level, such as ChEBI, with upper-level ontologies such as BFO, or indeed familiar laboratory-scale concepts such as mixtures. We work out relationships between chemical types that are compatible with the OBO Relation Ontology, describe macroscopic chemical systems in terms of grains and collectives, and propose a top-level ontology for chemically-relevant continuants and discuss it in relation to BFO and BioTop.

Keywords. upper-level ontology, chemistry, OBO Relation Ontology

1. Introduction

Chemoinformatics, the study of information systems that handle chemical entities, has spawned a multimillion-pound industry supplying services to the pharmaceutical and other industries. The field has been slow to adopt the methods of formal ontology. Exceptions include work by Dumontier and coworkers \([1,2]\), who use the connectivity of atoms to infer classifications for molecules, work at NIST \([3,4]\) on a chemical taxonomy of HIV inhibitors written in OWL, and ChEBI (Chemical Entities of Biological Interest)\([5]\) an ontology containing around 15 000 chemical entities. Most if not all biomedical ontologies refer explicitly or implicitly to chemical entities, so the OBO Foundry project \([6]\), a suite of interoperable biomedical ontologies, needs a soundly-constructed ontology of chemical entities to build on. However progress on aligning ChEBI with other biomedical ontologies such as the Gene Ontology (GO)\([7]\) is slow \([8,9]\). One reason is that it is far from clear from the definitions in ChEBI whether the terms refer to molecules in the universe, to names of molecules, or to molecular structures.

The case for an upper-level ontology for chemistry is therefore twofold; first, defining what the objects referred to in an ontology actually are allows the curators to set its scope and determine the genera for high-level genus–differentia definitions; second, it allows the types in the ontology to be reused safely by other ontologies with overlapping scope. Schulz et al. \([10]\) have developed an upper-level ontology for molecular biology by taking the GENIA ontology \([11]\) as a start point, and performing an ontological analysis of it which is based around choosing a set of foundational relations and developing formal properties and classifying entities based on those.

\(1\) Corresponding Author: Royal Society of Chemistry, Thomas Graham House, Milton Road, Cambridge, UK CB4 0WF; batchelorc@rsc.org
In this paper we present an upper-level ontology for chemistry based on an analysis of the foundational and disguised foundational relations in ChEBI. The paper is structured as follows: In section 2 we analyse ChEBI and discover. In section 3 we rewrite the relations so that they are compatible with the OBO Relation Ontology [12], taking into account Rector et al.’s [13] distinction between determinate and granular parthood, and determine the types that belong in this upper-level ontology. In section 4 we align the terms with a top-level ontology, in this case BFO [14]. Finally in Section 5 we discuss related work and assess the implications for ChEBI, BioTOP [10] and the InChI chemical identifier [15].

We present types as follows: existing ChEBI entities will be in italics, proposed top-level types will be in bold and relationships will be in monospaced text. CamelCase names are taken from BioTop. BFO types will be in CamelCase preceded by “snap:”.

2. Examination of ChEBI

We shall take a realist perspective[16] and think in terms of ChEBI containing types that are actually or potentially instantiated in the universe. So the instances we are seeking to describe in this paper with a chemical ontology include for example, single atoms in the interstellar medium, a water molecule interposed between a particular pair of nucleotides in a specific RNA motif, an individual cluster of atoms under a scanning electron microscope tip, a sodium ion in a glass of brine and so forth.

This clashes with the use of the word “instance” in the definition of the is_a relationship given by ChEBI:

Implies that ‘Entity A’ is an instance of ‘Entity B’. [...] chloroform (CHEBI:23143) is an instance of the class of chloromethanes (CHEBI:23148), which is itself an instance of the class of chloroalkanes (CHEBI:23143), and so forth.

Elsewhere ChEBI explicitly describes two superclasses: organic functional classes (CHEBI:33244) and natural product classes (CHEBI:33243), which reflect two parallel systems for classifying organic compounds, one according to the biological context where they were first discovered, the other according to their chemical composition. They will have to be renamed and have definitions written for them if ChEBI is to be made compatible with other realist ontologies. It is possible to imagine a shelf which has on it bottles containing natural products, but certainly not bottles that contain natural product classes.

For this section we shall determine what top-level classes are entailed by a chemical reading of ChEBI and its relationships and definitions, and define them properly, in the light of BFO, in a later section. We refer to the OBO format version of ChEBI Release 43 throughout.

2.1. Explicit Parthood

ChEBI has one explicit parthood relation, is_part_of, which is defined as follows:

Used to indicate relationship between part and whole. [...] tetracyanonickelate(2−) (CHEBI:30025) is part of potassium tetracyanonickelate(2−) (CHEBI:30071)
This example is not compatible with the part_of relationship in the OBO Relation Ontology, because that would entail that all tetracyanonickelate (2−) ions were part of some potassium tetracyanonickelate (2−) species, but of course it is quite possible to have tetracyanonickelate ions that are free in solution or part of some other complex. Indeed, we will see later on that compounds similar to “potassium tetracyanonickelate” may have different ontological interpretations depending on context.

Inspection of ChEBI leads to at least five interpretations of is_part_of:

1. is necessarily part of a molecular entity: carbonyl group (CHEBI:23019) and carbonyl compounds (CHEBI:36586)
2. is possibly part of a molecular entity: electron (CHEBI:10545) is part of muonium (CHEBI:30213)
3. is possibly part of a mixture: kanamycin A (CHEBI:17630) is part of kanamycin (CHEBI:6104)
4. is possibly part of a salt: lead(2+) (CHEBI:30179) is part of lead diacetate (CHEBI:31767)
5. is connected in the ChEBI database structure to another entry: biological role (CHEBI:24432) is part of ChEBI ontology (CHEBI:23091)

This contradicts the principle of univocity,[17] that each relationship should only have one interpretation. This is also possible because the domain and range of the is_part_of relationship have not been specified in advance. We will defer in-depth discussion of these to a later section, but point out that the fifth interpretation can in no sense be related to instances in reality and therefore we will not consider it further.

2.2. Disguised Parthood

There are also three ‘disguised’ parthood relations that prove straightforwardly to be incompatible with a realist approach to ontology development—is_substituent_group_from, has_functional_parent and has_parent_hydride. First, is_substituent_group_from can be interpreted as:

1. is possibly part of a molecular entity: methyl group (CHEBI:32875) is_substituent_group_from methane (CHEBI:16183)

This is difficult to interpret in terms of the all–some structure of RO relations. Does it make sense to say that all methyl groups are derived formally from some other methane molecule? For compatibility with RO it is best to drop this relation altogether and think of the relation between methyl groups and methane molecules in terms of parthood as we shall see later.

Second, has_functional_parent from a realist perspective looks like this:

1. shares a substructure with a molecular entity: D-glucoronate 1-phosphate (CHEBI:28547) has_functional_parent D-glucuronate (CHEBI:15748)

The ChEBI definition is:

Used to denote the relationship between two molecular entities (or classes of entities), one of which possesses one or more characteristic groups from which the other can be derived by functional modification.
Again, while it is true to say that all molecules of \( \text{D-glucoronate 1-phosphate} \) could in principle have been formed by phosphorylating some molecule of \( \text{D-glucoronate} \) it seems best again to drop this relationship.

Last, has_parent_hydride is a special case of has_functional_parent, though it does have a range if not a domain:

Denotes the relationship between an entity and its parent hydride (defined by IUPAC as “an unbranched acyclic or cyclic structure or an acyclic/cyclic structure having a semisystematic or trivial name to which only hydrogen atoms are attached”). The definition of parent hydride here is in fact closest to that of CHEBI:33245, organic fundamental parents, but this type is problematic as it has been added as a guide to systematic nomenclature; the children of this type have little in common.

### 2.3. Subsumption Relations

The is_a relationship is used for a variety of purposes in ChEBI:

1. an amount of a compound has a biological role:  
   \( \text{tris (CHEBI:9754)} \) is a buffer (CHEBI:35225)
2. an amount of a compound has an application:  
   \( \text{sodium dodecyl sulfate (CHEBI:8984)} \) is a detergent (CHEBI:27780)
3. connecting a less-abstract class with a more-abstract class:  
   \( \text{propane (CHEBI:32879)} \) is a alkanes (CHEBI:18310)
4. connecting macroscopic entities with atoms:  
   \( \text{metals (CHEBI:33521)} \) is a atoms (CHEBI:33250).
5. connecting elements with atoms:  
   \( \text{main group elements (CHEBI:33318)} \) is a atoms (CHEBI:33250)

The clearest problems with the first two cases are that the relationships would be better represented as has_role and has_application, and secondly that they can only be understood in terms of amounts of compounds, inconsistently with the rest of the ontology, which is mainly described in terms of individual molecules. It is samples of tris that act as buffers, rather than individual tris molecules, but every propane molecule is an alkane molecule. The third case is absolutely standard in ontologies and it is possible to imagine genus–differentia definitions for almost all of the structurally-based examples in ChEBI. However, ChEBI has not followed this route, preferring a multiple-genus approach. We shall consider the case of \( \text{polypodine B (CHEBI:28485)} \), which has no fewer than nine is_a parents, which we list in Table 1. The easiest parent to discuss is \( \text{phytoecdysteroids (CHEBI:26118)} \), which says something about the role the compound plays in the context of plant metabolism but is inessential to its structural description. The other parents have genus–differentia names but not definitions, so a \( 3\beta\)-hydroxy steroid is a compound with the steroid skeleton and a hydroxy group bound to the skeleton at the position numbered 3 with a particular orientation (\( \beta \)) relative to the rest of the molecule. The definition of the steroid skeleton and the canonical numbering are given by Moss.[18]

This demonstrates that simply stating the determinate parts of a molecule is not enough to specify it unambiguously. We must also specify their relative locations. This requires a new parthood relationship which we will define in detail in the next section.
The fourth and fifth cases are tricky to disentangle: metals and main group elements themselves are undefined within ChEBI, but the definition of atom reads:

An atom is the smallest particle still characterizing a chemical element.

This definition relies on the notion of a chemical element, which is defined in the Gold Book [19] as

1. A species of atoms; all atoms with the same number of protons in the atomic nucleus.
2. A pure chemical substance composed of atoms with the same number of protons in the atomic nucleus. Something this concept is called the elementary substance as distinct from the chemical element as defined under 1, but mostly the term chemical element is used for both concepts.

There is a straightforward fix for the oddity of elements being is_a children of atoms. If the descendants are renamed with names ending in “atom”—“main group atom”, “s-block atom” and so forth, then the problem goes away. The existence, however, of the terms metals (CHEBI:33521) and nonmetals (CHEBI:25585) is pathological. Metallicity and nonmetallicity are not properties of atoms but rather of large assemblies of atoms, so would belong elsewhere in ChEBI if at all.

### 3. Relation Definitions

We need to distinguish between the parthood relations that hold both (a) on the level of molecules and a sample in a test-tube, and (b) only on the level of the sample in the test-tube. Grenon et al. [14] propose that each material application of BFO should be restricted to a given level of granularity. For this reason we seek to distinguish the relations that hold within a level of granularity from those that hold between levels. Rector et al. [13] propose a distinction between granular and determinate parts which is useful to us here. A determinate part, for example a finger on a hand, is one in which the part is directly part of the whole and in which on being removed necessarily
damages or diminishes the whole. A granular part is one, like a cell within a finger, where the grains are parts of the whole by virtue of being grains in a collective and in which removing one granular part does not necessarily damage or diminish the whole.

Another way of looking at the distinction is that determinate parts are often named (ring, middle, index, pinky in the case of fingers), or numbered as in the case of atoms in a molecule (the InChI identifier is written in terms of a canonical numbering for each atom) or amino acid residues in a peptide chain. No such naming or numbering happens to hairs on one’s head or the molecules in a macroscopic sample.

The determinacy of the determinate parthood relation is stronger for chemical entities than is generally the case in biomedicine. While my hand can lose a finger while remaining identifiably a hand, a molecule of C\textsubscript{60} cannot lose carbon atoms without ceasing to be C\textsubscript{60}. The same applies to nearly all of the entities in ChEBI defined in terms of a specific formula. But the granular parthood relation is exactly the same—a lump of gold can lose gold atoms without ceasing to be gold. A lump of rocksalt can lose both Na\textsuperscript{+} and Cl\textsuperscript{−} ions without ceasing to be rocksalt.

We need to distinguish salts (and their molten phases, ionic liquids) from mixtures. It is not clear from [13] whether salts count as mixtures. As a mixture is understood in chemistry, however, it is an amount of matter that has components, or in the language of [13], “ingredients”, which are collectives in which the grains retain their identity. The ingredients of a mixture can in practice be recovered by mild techniques such as chromatography. Absent from the definition in [13] of a mixture being an amount of matter that has ingredients, is this notion of the grains retaining their identity. A macroscopic sample of a component of a mixture can generally exist independently but a macroscopic sample of lead (2+) would undergo a spectacular Coulomb explosion. Other examples of mixtures include alloys and solutions, gels, foams and soft matter in general.

If we examine our four surviving is_part_of cases from above for determinate and granular parts we find:

1. carbonyl compound molecules have as determinate parts carbonyl groups
2. muonium atoms have as determinate parts electrons (subatomic particles)
3a. kanamycin mixtures have as an ingredient kanamycin A
3b. kanamycin mixtures have as granular parts kanamycin A molecules
3c. kanamycin mixtures have as granular parts molecules of kanamycins, where kanamycins (CHEBI:24951) is a class that has subclasses kanamycins A, B and C, which are compounds, kanamycin (CHEBI:6104), which is the mixture we have already discussed, and the kanamycin derivatives 2\textsuperscript{nd}-nucleotidyikanamycins (CHEBI:27557) and acetylkanamycins (CHEBI:22201).
4. lead diacetate salts have as granular parts lead(2+) ions.
We might also add, though these relations are not explicit in ChEBI:
5. carbon dioxide molecules have as determinate parts oxygen atoms
6. [bmim][PF\textsubscript{6}] ionic liquids have as determinate parts bmim\textsuperscript{+} ions.
7. pure gold has as granular parts gold atoms

So our macroscopic entities, salts, solutions, mixtures, and pure substances all have granular parts. In addition, mixtures and solutions have ingredients. Molecular entities, on the other hand, have only determinate parts. Note, however, that the chemical formula may be insufficient to tell you whether it refers to a molecular or macroscopic entity. Take NaCl, for example. In both liquid and solid NaCl, the sodium ions have no particular affinity for a single chloride ion, but in the solid state are surrounded by six
Table 2. Domains and ranges of relationships for upper-level chemical classes.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>has_determinate_part</td>
<td>molecule</td>
<td>molecular part, atom</td>
</tr>
<tr>
<td></td>
<td>molecular part</td>
<td>molecular part, atom</td>
</tr>
<tr>
<td></td>
<td>atom</td>
<td>subatomic particle</td>
</tr>
<tr>
<td>has_substituent</td>
<td>molecule</td>
<td>molecular part, atom</td>
</tr>
<tr>
<td></td>
<td>molecular part</td>
<td>molecular part, atom</td>
</tr>
<tr>
<td>has_granular_part</td>
<td>pure substance</td>
<td>molecule</td>
</tr>
<tr>
<td></td>
<td>pure substance</td>
<td>atom</td>
</tr>
<tr>
<td></td>
<td>salt</td>
<td>molecule</td>
</tr>
<tr>
<td></td>
<td>salt</td>
<td>atom</td>
</tr>
<tr>
<td></td>
<td>mixture</td>
<td>molecule</td>
</tr>
<tr>
<td></td>
<td>mixture</td>
<td>atom</td>
</tr>
<tr>
<td>has_ingredient</td>
<td>mixture</td>
<td>pure substance, salt</td>
</tr>
</tbody>
</table>

nearest neighbours. There are no identifiable molecules and they can only be understood as macroscopic entities. Contrast this with a sample of gaseous NaCl, which consists of diatomic molecules. Loss of a sodium or chlorine atom from a given molecule would result in it no longer being an NaCl molecule. This suggests very strongly that ChEBI should be restricted to describing molecular-level entities.

The relationships in RO are all defined in the all–some direction, which makes it a trivial task to determine an RO-compatible mapping. Currently is_part_of means in about a dozen cases ‘necessarily part of’, and in all others ‘possibly part of’. The solution is to reverse them all and replace them with the appropriate specialization of the RO has_part relation. All but one of our parthood relations have the same formal definition as the RO has_part relation, except that the domains and ranges, which are listed in Table 2, are different for each one.

The exception is has_substituent. The domain of has_substituent, \( C \) in the definition, is molecules and molecular parts. Allowable substituents \( S \) are molecular parts and atoms.

\( C \) has_substituent \( S \) at_position \( P \) \( \equiv \) for all \( c, s, t \), if \( c \) instance_of \( C \) at \( t \) and \( s \) instance_of \( S \) at \( t \) then there is some \( s \) that is part_of \( c \) at position \( P \).

It is useful here to introduce the distinction between an open-world name and a closed-world name. Many chemical names are polysemous and have both an open-world and a closed-world reading. This has been discussed at length from a natural-language-processing perspective by Corbett et al. [20]. Systematic and semi-systematic chemical names specify a genus and differentiae. Often these differentiae themselves have genus–differentia form, so the name has to be interpreted recursively. In the closed-world reading, which corresponds to an EXACT name in [20], any locations with unspecified substituents have a hydrogen atom attached to them. In the open-world reading, which corresponds to a CLASS name in [20], the substituents which are not mentioned in the name are left unspecified. Thus a type that has a name in the closed-world reading is an is_a descendant of the type that has the same name in the open-world reading.
In the above case a propane molecule (in the closed-world sense) is an alkane molecule which has_part carbon atom with a cardinality of exactly 3 and has_part hydrogen atom with a cardinality of exactly 8.

polypodine B would be the intersection of steroids (CHEBI:35341) with eight substituent–position pairs as follows:

A steroid that has an oxo substituent at position 6, a beta-hydroxy substituent at position 2, a beta-hydroxy substituent at position 3, a beta-hydroxy substituent at position 5, an alpha-hydroxy substituent at position 14, a hydroxy substituent at position 20 and a hydroxy substituent at position 26.

If we assume that all positions left unspecified in that definition are occupied by hydrogen atoms, then we have the type for the molecule with the closed-world name “polypodine B”. If we leave them unspecified, then we have a type for the molecules with the open-world name “polypodine B”, which chemists might refer to as “the polypodine B”, “a polypodine B”, or, although these are not attested, “polypodine Bs” or “polypodines B”.

4. Alignment with a Top-Level Ontology

There is no shortage of top-level ontologies to choose from, of which DOLCE [21], GFO [22] and BFO [14] are recent examples. We choose BFO for this alignment because it is the top-level ontology for the OBO ontologies of which ChEBI is one.

In the last section we identified top-level classes molecule, molecular part, atom, subatomic particle, pure substance (for want of a better name), salt and mixture. Our careful insistence on a realise perspective has ensured that we haven’t interpreted them as molecular structures, which would in terms of BFO be generically-dependent continuants, or names. They are all, in BFO terms, snap:IndependentContinuants. They cannot inhere in anything else. molecules and pure substances are all snap:Objects—they are spatially extended, self-connected and self-contained. salts and mixtures are clearly not snap:ObjectAggregates, because all of the boundaries within them (between, for example, solutes and solvents, or positively-charged and negatively-charged ions) are connected to some other boundary within the system. There are, so to speak, no gaps.

molecular parts, for example the methyl group in a toluene molecule are snap:FiatObjectParts, in that there is no physical discontinuity between the methyl group and the benzene ring. The methyl group is a good example of ‘carving nature at the joints’, as the processes of adding a methyl group (methylation) and removing one (demethylation) are carried out by both enzymes and small molecules in nature and in the laboratory.

atoms and subatomic particles are somewhat trickier. While it is true that within an atom there are nodal points and planes where the probability of finding an electron is mathematically zero, this is not the sense of a gap in a snap:ObjectAggregate, not least because it is the same electron on either side of the nodal plane. One cannot somehow convert the nodal plane into a bigger, three-dimensional, gap and obtain two fractional electrons. Hence they are either, according to whether they are bound, snap:Objects or snap:FiatObjectParts. We therefore call them snap:IndependentContinuants.²

Now that we have the alignment with BFO we can write definitions for our top-level classes.

²I am grateful to Barry Smith for this suggestion [23], which is also a recommendation of Schulz et al. [24].
1. **atom** A snap:IndependentContinuant which has as determinate parts a single atomic nucleus and one or more electrons.
2. **subatomic particle** A snap:IndependentContinuant which does not have as granular parts atomic nuclei or electrons.
3. **molecule** A snap:Object which has as determinate parts two or more atoms.
4. **molecular part** A snap:FiatObjectPart which consists of at least one atom which is part of a molecule.
5. **collective** A snap:Object that has granular parts.
6. **pure substance** A collective of many atoms or molecules of a single type.
7. **salt** A collective of positively-charged atoms or molecules and negatively-charged atoms or molecules.
8. **mixture** A collective that has as ingredients pure substances.

5. Related Work

In addition to ChEBI, there are three serious chemical ontologies of which we are aware. The first, CO, is by Dumontier and co-workers [1,2], who have been using OWL to represent how atoms are connected to each other and have worked out necessary and sufficient conditions for defining particular molecular parts, such as a carbonyl group or a nitro group. The top-level term in [1] is OrganicGroup, but no alignment of this to a higher-level ontology is attempted. The second is ChemTop, which is simply the chemical subset of BioTop.[25] The third is the chemical ontology implicit in ChemBLAST [3], which, like CO, has no explicit commitment to a top-level ontology.

5.1. Implications for ChEBI

The current molecular structure tree, therefore, needs to be replaced by five trees rooted in atom, molecule, molecular part, subatomic particle and collective. It would be safest, in order to prevent the sort of confusion seen above, if this last were moved to a separate ontology.

The main implication of this work for ChEBI is that the current is_part_of relation should be replaced by appropriate has_part relations.3

The first thing to note is that the high-level classes within ChEBI do not map neatly to the high-level classes we have found here. Secondly, ChEBI is not is_a-complete. To take an eye-catching example, amino-acid residues (CHEBI:33708) has no is_a parents, even though it is_substituent_group_from amino acids (CHEBI:33709), so there is no inferrable connection between it and groups (CHEBI:24433).

The full alignment task, which we do not attempt here, is to determine the upper-level classification of each type in the ontology by inspection. Before we can use a reasoner to work with genus–differentia definitions we must first have a genus for each type. Often there are cues in the term name. We list some of these cues in Table 3.

It is also worth noting in passing some other problems with ChEBI as it is presently constituted, notably that many entity names with a closed-world reading are plural. This implies that ChEBI has been constructed in terms of instances rather than types. It

3Note added after acceptance: A similar changeover has now been scheduled for the October 2008 release of ChEBI.
Table 3. Some cues for classifying entities in ChEBI

<table>
<thead>
<tr>
<th>Top-level class</th>
<th>Cue</th>
</tr>
</thead>
<tbody>
<tr>
<td>atom</td>
<td>name is a chemical element or ends “elements”</td>
</tr>
<tr>
<td>mixture</td>
<td>definition includes “mixture” or “racemate”</td>
</tr>
<tr>
<td>molecule</td>
<td>default</td>
</tr>
<tr>
<td>molecular part</td>
<td>name ends in “group(s)” or “residue(s)”</td>
</tr>
<tr>
<td>pure substance</td>
<td>difficult to determine automatically. is_a descendant of native element minerals (CHEBI:46730)</td>
</tr>
<tr>
<td>salt</td>
<td>is_a descendant of salts (CHEBI:24866) or minerals (CHEBI:46662) but not of native element minerals (CHEBI:46730)</td>
</tr>
<tr>
<td>subatomic particle</td>
<td>is_a descendant of subatomic particle (CHEBI:36342)</td>
</tr>
</tbody>
</table>

also leads to ungrammatical-looking statements such as metals is_a atoms. It would be clearer to replace plural names, for example pyrroles, with something like “pyrrole compounds”—an alternative, “pyrrole molecule”, could be interpreted as either a molecule of pyrrole (closed-world) or a molecule of a pyrrole (open-world).

5.2. Implications for the Use of InChI

Prasanna and co-workers[4] use InChIs to represent fully-specified molecules, under-specified molecules and molecular parts. This is intrinsically ambiguous because the bonding between atoms is only implied in the InChI by the presence or otherwise of hydrogen atoms. Thus the InChI

InChI=1/C6/c1-2-4-3-5-6

which represents either an implausible molecule which is only likely to last long in interstellar space or a fragment of a molecule, could, given the molecular part interpretation, indicate either a singly-bonded cyclohexane ring or an aromatically-bonded benzene ring.

This means that even the combination of this upper-level ontology for chemistry and the InChI identifier is insufficient to fully describe objects of chemical discourse without a domain ontology such as ChEBI. However, it would be useful to indicate whether a given InChI is to be understood as a grain or a collective.

5.3. Implications for BioTop

We list some example mappings in Table 4. BioTop’s classes Atom and SubAtomicParticle map straightforwardly to ours.

Some of them, however, we list as undefined. The reason is that they have is_a descendants which belong to different upper-level types. MonoMolecularEntity and PolyMolecularCompositeEntity, for example, both have some descendants that map to molecule and some that map to molecular part. The same applies to NucleicAcid and Peptide.

Other terms are undefined because their definitions refer to different upper-level types, for example PortionOfHeterogeneousSolid, where the examples given in the
Table 4. Some example mappings from BioTop to current work

<table>
<thead>
<tr>
<th>Current work</th>
<th>BioTop</th>
</tr>
</thead>
<tbody>
<tr>
<td>atom</td>
<td>Atom</td>
</tr>
<tr>
<td>subatomic particle molecule</td>
<td>SubAtomicParticle</td>
</tr>
<tr>
<td>molecule</td>
<td>EntireMolecularEntity</td>
</tr>
<tr>
<td></td>
<td>WaterMolecule</td>
</tr>
<tr>
<td></td>
<td>Lipid</td>
</tr>
<tr>
<td></td>
<td>Eicosanoid</td>
</tr>
<tr>
<td></td>
<td>Steroid</td>
</tr>
<tr>
<td></td>
<td>EntireNucleicAcidMolecule</td>
</tr>
<tr>
<td></td>
<td>FattyAcid</td>
</tr>
<tr>
<td></td>
<td>EntireProteinMolecule</td>
</tr>
<tr>
<td>mixture</td>
<td>PortionOfHeterogeneousLiquid</td>
</tr>
<tr>
<td>molecular part</td>
<td>Monomer, HeterocyclicBase</td>
</tr>
<tr>
<td>undefined</td>
<td>MonoMolecularEntity</td>
</tr>
<tr>
<td></td>
<td>NucleicAcid</td>
</tr>
<tr>
<td></td>
<td>Carbohydrate</td>
</tr>
<tr>
<td></td>
<td>ChainOfCarbohydrateMonomers</td>
</tr>
<tr>
<td></td>
<td>PortionOfHeterogeneousSolid</td>
</tr>
<tr>
<td></td>
<td>MoleculeComplex</td>
</tr>
<tr>
<td></td>
<td>Peptide</td>
</tr>
<tr>
<td></td>
<td>AminoAcidMonomer</td>
</tr>
</tbody>
</table>

BioTop definition are a NaCl crystal, which we identify as a salt, and graphite, which is a pure substance, and AminoAcidMonomer, whose definition is “Amino Acids molecules or residues (residues as in peptide bonds)”. In some cases there is not enough information to decide.

6. Conclusions and future work

We have examined ChEBI in the light of the OBO Relations Ontology (RO), Rector et al.’s notion of granular and collective parts and BFO, and used it to create an upper-level ontology of independent chemical continuants and propose a thorough overhaul of the structure of ChEBI with RO-compatible relationships. We have also looked briefly at how the upper-level types relate to ChEBI, BioTop and the InChI identifier. However, this is still essentially an armchair ontology. We have not attempted to test it with automated reasoning, as has been done as part of BioTop[10] or by Keet and Artale for generic parthood relations.[26] The other thing that is missing is a proper consideration (and alignment to BFO) of the dependent continuants in ChEBI. We have shown that a number of them only make sense if applied to collectives rather than grains, but we have not determined whether this is the case for all of them.
Acknowledgements

I thank Peter Corbett, Chris Mungall and Barry Smith for useful preliminary discussions. I would also like to thank the participants in a workshop on the future of ChEBI held on the 9th of July 2008 at the EBI in Hinxton, especially Kirill Degtyarenko, Janna Hastings, Barry Smith and Alan Ruttenberg.

References


[23] B. Smith, personal communication.

