8

An Overview of OntoClean

Nicola Guarino[1] and Christopher A. Welty[2]

[1]Laboratory for Applied Ontology (ISTC-CNR)
Polo Tecnologico, Via Solteri 38, 38100 Trento, ITALY
guarino@isib.cnr.it

[2]IBM Watson Research Center
19 Skyline Dr., Hawthorne, NY 10532, USA
welty@us.ibm.com

Summary. OntoClean is a methodology for validating the ontological adequacy of taxonomic relationships. It is based on highly general ontological notions drawn from philosophy, like *essence*, *identity*, and *unity*, which are used to characterize relevant aspects of the intended meaning of the properties, classes, and relations that make up an ontology. These aspects are represented by formal metaproperties, which impose several constraints on the taxonomic structure of an ontology. The analysis of these constraints helps in evaluating and validating the choices made. In this chapter we present an informal overview of the philosophical notions involved and their role in OntoClean, review some common ontological pitfalls, and walk through the example that has appeared in pieces in previous papers and has been the basis of numerous tutorials and talks.

8.1 Introduction

The OntoClean methodology was first introduced in a series of conference-length papers in 2000 [Guarino and Welty, 2000a-c; Welty and Guarino, 2001], and received much attention and use in subsequent years. The main contribution of OntoClean was the beginning of a formal foundation for ontological analysis. Alan Rector, a seasoned veteran at ontological analysis in the medical domain, said of OntoClean, "…what you have done is reduce the amount of time I spend arguing with doctors that the way I want to model the world is right…" [Rector, 2002]. A similar comment came from the CYC people attending our AAAI-2000 tutorial, "You showed why the heuristic choices we adopted were right." Most experienced domain modelers can see the correct way to, e.g. structure a taxonomy, but are typically unable to justify themselves to others. OntoClean has provided a logical basis for arguing against the most common modeling pitfalls, and arguing for what we have called "clean ontologies".

In this chapter we present an informal overview of the basic notions *essence, identity,* and *unity,* and their role in OntoClean. We then review the basic ontol-

ogy pitfalls, and walk through the example that has appeared in pieces in previous papers and has been the basis of numerous tutorials and talks beginning with AAAI-2000.

Background

The basic notions in OntoClean were not new, but existed in philosophy for some time. Indeed, the practice of modeling the world for information systems has many parallels in philosophy, whose scholars have been trying to describe the universe in a formal, logical way since the time of Aristotle. Philosophers have struggled with deep problems of existence, such as God, life and death, or whether a statue and the marble from which it is made are the same entity. While these problems may seem irrelevant to the designer of an information system, we found that *the conceptual analysis and the techniques used to address these problems* are not, and form the basis of our methodology.

Properties, classes, and subsumption

Many terms have been borrowed by computer science from mathematics and logic, but unfortunately this borrowing has often resulted in a skewed meaning. In particular, the terms *property* and *class* are used in computer science with often drastically different meanings from the original. The use of the term *property* in RDF is an example of such unfortunate deviation from the usual logical sense.

In this chapter, we shall consider properties as the *meanings* (or *intensions*) of expressions like *being an apple* or *being a table*, which correspond to unary predicates in first-order logic. Given a particular maximal state of affairs (or *possible world*), we can associate with each property a *class* (its *extension*), which is the set of entities that exhibit that property in that particular world. The members of this class will be called *instances* of the property. Classes are therefore sets of entities that share a property in common; they are the extensional counterpart of properties. In the following, we shall refer most of the time to properties rather than classes or predicates, to stress the fact that their ontological nature (characterized by means of *metaproperties*) does not depend on syntactic choices (as it would be for predicates), nor on specific states of affairs (as it would be for classes).

The independence of properties from states of affairs gives us the opportunity to make clear the meaning of the term *subsumption* we shall adopt in this paper. A property p subsumes q if and only if, *for every possible state of affairs*, all instances of q are also instances of p. On the syntactic side, this corresponds to what is usually held for description logics, P subsumes Q if and only if there is no model of $Q \wedge \neg P$.

8.2 The basic notions

Essence and Rigidity

A property of an entity is *essential* to that entity if it *must* be true of it in every possible world, i.e. if it *necessarily holds* for that entity. For example, the property of *having a brain* is essential to human beings. Every human *must* have a brain in every possible world.

A special form of essentiality is rigidity; a property is *rigid* if it is essential to all its possible instances; an instance of a rigid property cannot stop being an instance of that property in a different world. For example, while having a brain may be essential to humans, it is not essential to, say, scarecrows in the *Wizard of Oz*. If we were modeling the world of the *Wizard of Oz*, the property of *having a brain* would not be rigid, though still essential to humans. On the other hand, the property *being a human* is typically rigid, every human is necessarily so.

Note that we use the word "typically" here to stress that the point of OntoClean is *not* to help people decide about the ontological nature of a certain property, but rather to help them explore the logical consequences of making certain choices. Rigidity is the first ingredient of this framework: it is a *metaproperty,* deciding whether it holds or not for the relevant properties in an ontology helps to clarify its *ontological commitment.*

Obviously there are also *non-rigid* properties, which can acquire or lose (some of) their instances depending on the state of affairs at hand. Of these we distinguish between properties that are essential to *some* entities and not essential to others (*semi-rigid*), and properties that are not essential to *all* their instances (*anti-rigid*). For example, the property *being a student* is typically anti-rigid – *every* instance of student can cease to be such in a suitable state of affairs, whereas the property *having a brain* in our *Wizard of Oz* world is semi-rigid, since there are instances that must have a brain as well as others for which a brain is just a (desirable) option.

Rigidity and its variants are important metaproperties, every property in an ontology should be labeled as rigid, non-rigid, or anti-rigid. In addition to providing more information about what a property is intended to mean, these metaproperties impose constraints on the subsumption relation, which can be used to check the ontological consistency of taxonomic links. One of these constraints is that anti-rigid properties cannot subsume rigid properties. For example, the property *being a student* cannot subsume *being a human* if the former is anti-rigid and the latter is rigid. To see this, consider that, if *p* is an anti-rigid property, all its instances can cease to be such. This is certainly the case for *student*, since any student may cease being a student. However, no instance of *human* can cease to be a human, and if all humans are necessarily students (the meaning of subsumption), then no person could cease to be a student, creating therefore an inconsistency.

Identity and Unity

Although very subtle and difficult to explain without experience, identity and unity are perhaps the most important notions we use in our methodology. These two things are often confused with each other; in general, *identity* refers to the problem of being able to recognize individual entities in the world as being the same (or different), and *unity* refers to being able to recognize all the parts that form an individual entity.

Identity *criteria* are the criteria we use to answer questions like, "is that my dog?" In point of fact, identity criteria are conditions used to *determine* equality (sufficient conditions) and that are *entailed by* equality (necessary conditions).

It is perhaps simplest to think of identity criteria over time (*diachronic* identity criteria), e.g. how do we recognize people we know as the *same* person even though they may have changed? It is also very informative, however, to think of identity criteria at a single point in time (*synchronic* identity criteria). This may, at first glance, seem bizarre. How can you ask, "are these *two* entities the same entity?" If they are the same then there is one entity, it does not even make sense to ask the question.

The answer is not that difficult. One of the most common decisions that must be made in ontological analysis concerns identifying circumstances in which one entity is actually two (or more). Consider the following example, drawn from actual experience: somebody proposed to introduce a property called *time duration* whose instances are things like *one hour* and *two hours*, and a property *time interval* referring to specific intervals of time, such as "1:00 – 2:00 next Tuesday" or "2:00 – 3:00 next Wednesday." The proposal was to make *time duration* subsume *time interval*, since all time intervals are time durations. Seems to make intuitive sense, but how can we evaluate this decision?

In this case, an analysis based on the notion of identity can be informative. According to the identity criteria for time durations, two durations of the same length are the same duration. In other words, all one-hour time durations are identical – they are the *same* duration and therefore there is only one "one hour" time duration. On the other hand, according to the identity criteria for time intervals, two intervals of the same duration occurring at the same time are the same, but two intervals occurring at different times, even if they are the same duration, are different. Therefore the two example intervals above would be different intervals. This creates a contradiction: if all instances of *time interval* are also instances of *time duration* (as implied by the subsumption relationship), how can they be two instances of one property and a single instance of another?

This is one of the most common confusions of natural language when used for describing the world. When we say "all time intervals are time durations" we really mean "all time intervals *have* a time duration" – the duration is a component of an interval, but it is not the interval itself. In this case we cannot model the relationship as subsumption, time intervals have durations (essentially) as *qualities.* More examples of such confusions are provided at the end of this article.

One of the distinctions proposed by OntoClean is between properties that *carry an identity criterion* and properties that do not. The former are labeled with an *ad-*

hoc metaproperty, **+I**. Since criteria of identity are inherited along property subsumption hierarchies, a further distinction is made to mark those properties that *supply* (rather just *carrying*) their "own" identity criteria, which are not inherited from the subsuming properties. These properties are marked with the label **+O** (where **O** stands for "own").

Unfortunately, despite their relevance, recognizing identity criteria may be extremely hard. However, in many cases identity analysis can be limited to detecting the properties that are just *necessary* for keeping the identity of a given entity, i.e. what we have called the *essential properties*. Obviously, if two things do not have the same essential properties they are not identical. Take for instance the classical example of the statue and the clay: is the statue identical to the clay it is made of? Let's consider the essential properties: having (more or less) a certain shape is essential for the statue, but not essential for the clay. Therefore, they are different: we can say they have different identity criteria, even without knowing exactly what these criteria are. In practice, we can say that "sharing the essential property P", where P is essential for all the instances of a property Q different from P, is the weakest form of an identity criterion carried by Q. Such criterion can be used to make conclusions about non-identity, if not about identity.

A second notion that is extremely useful in ontological analysis is *Unity*. Unity refers to the problem of describing the parts and boundaries of objects, such that we know in general what is part of the object, what is not, and under what conditions the object is *whole*.

Unity can tell us a lot about the intended meaning of properties in an ontology. Certain properties pertain to wholes, that is, all their instances are wholes, others do not. For example, *being (an amount of) water* does not have wholes as instances, since each amount can be arbitrarily scattered or confused with other amounts. In other words, knowing an entity is an amount of water does not tell us anything about its parts, nor how to recognize it as a single entity. On the other hand, *being an ocean* is a property that picks up whole objects, as its instances, such as "the Atlantic Ocean," are recognizable as single entities. Of course, one might observe that oceans have vague boundaries, but this is not an issue here: the important difference with respect to the previous example is that in this case we have a criterion to tell, at least, what is *not* part of the Atlantic Ocean, and still part of some other ocean. This is impossible for amounts of water.

In general, in addition to specifying whether or not properties have wholes as instances, it is also useful to analyze the specific conditions that must hold among the parts of a certain entity in order to consider it a whole. We call these conditions *unity criteria* (UC). They are usually expressed in terms of a suitable *unifying relation*, whose ontological nature determines different kinds of wholes. For example, we may distinguish *topological wholes* (a piece of coal), *morphological wholes* (a constellation), *functional wholes* (a hammer, a bikini). As these examples show, nothing prevents a whole from having parts that are themselves wholes (under different unifying relations). Indeed, a *plural whole* can be defined as a whole that is a mereological sum of wholes.

In OntoClean, we distinguish with suitable metaproperties the properties *all* whose instances *must* carry a *common* UC (such as *ocean*) from those that do not.

Among the latter, we further distinguish properties all of whose instances must be wholes, although with different UCs, from properties all of whose instances are not necessarily wholes. An example of the former kind may be *legal agent*, if we include both people and companies (with different UCs) among its instances. *Amount of water* is usually an example of the latter kind, since none of its instances *must* be wholes. We say that *ocean* carries unity (**+U**), *legal agent* carries no unity (**-U**), and *amount of water* carries anti-unity (**~U**).

The difference between unity and anti-unity leads us again to interesting problems with subsumption. It may make sense to say that "Ocean" is a subclass of "Water", since all oceans are water. However, if we claim that instances of the latter must not be wholes, and instances of the former always are, then we have a contradiction. Problems like this again stem from the ambiguity of natural language, oceans are not "kinds of" water, they are *composed* of water.

Constraints and Assumptions

A first observation descending immediately from our definitions regards some *subsumption constraints.* Given two properties, *p* and *q,* when *q* subsumes *p* the following constraints hold:

1. If *q* is anti-rigid, then *p* must be anti-rigid
2. If *q* carries an identity criterion, then *p* must carry the same criterion
3. If *q* carries a unity criterion, then *p* must carry the same criterion
4. If *q* has anti-unity, then *p* must also have anti-unity

Finally, we make the following assumptions regarding identity (adapted from Lowe [Lowe, 1989]):

- *Sortal Individuation*. Every domain element must instantiate some property carrying an IC (+I). In this way we satisfy Quine's dicto "No entity without identity" [Quine, 1969].
- *Sortal Expandability*. If something is an instance of different properties (for instance related to different times), then it must be also instance of a more general property carrying a criterion for its identity.

Together, the two assumptions imply that every entity must instantiate a *unique* most general property carrying a criterion for its identity.

8.3 An extended example

In this section we provide a walk-through of the way the OntoClean analysis can be used. This example is based on those presented at various tutorials and invited talks.

We begin with a set of classes arranged in a taxonomy, as shown in Figure 1. The taxonomy we have chosen makes intuitive sense *prima facie*, and in most cases the taxonomic pairs were taken from existing ontologies such as Wordnet[1], Pangloss[2], and the 1993 version of CYC[3].

We have chosen, following our previous papers, to use a shorthand notation for indicating metaproperty choices on classes. Rigidity is indicated by **R**, identity by **I**, unity by **U**, and dependence by **D**. Each letter is preceded by **+**, **-** or **~**, to indicate the positive, negative, or *anti* metaproperty, e.g., being rigid (**+R**), carrying an identity criterion (**+I**), carrying a common unity criterion (**+U**); not rigid (**-R**), not carrying an identity criterion (**-I**), not carrying a common unity criterion (**-U**); being anti-rigid (**~R**) and having anti-unity (**~U**). We also used (**+O**) to indicate when a property carries its *own* identity criterion, as opposed to inheriting one from a more general property.

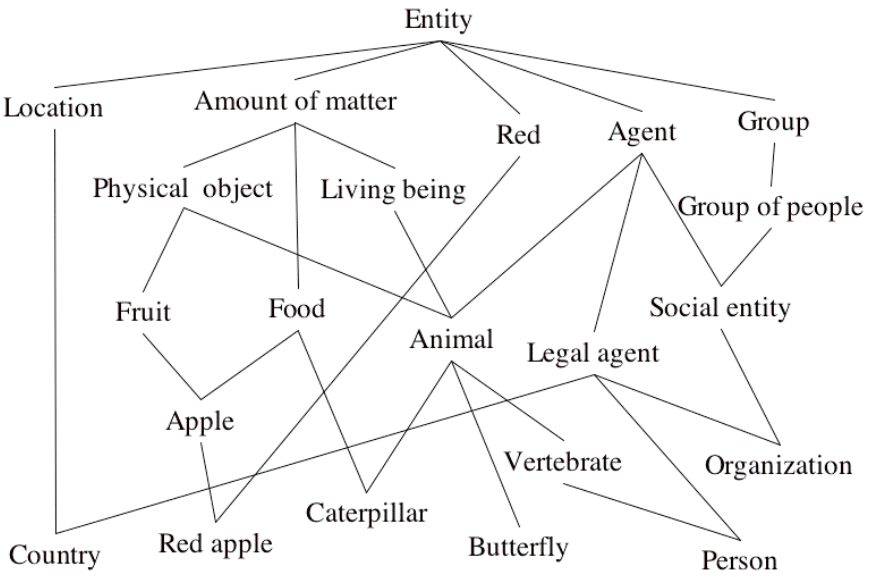


Figure 1. An uncleaned taxonomy.

[1] http://www.cogsci.princeton.edu/~wn/

[2] http://www.lti.cs.cmu.edu/Research/Pangloss/

[3] The current version of Cyc no longer contains these errors: http://www.cyc.com

Assigning Metaproperties

The next step is to assign the metaproperties discussed above to each property in the taxonomy. When designing a new ontology, this step may occur first, before arranging the properties in a taxonomy. Note that the assignments discussed here are not meant to be definitive at all: rather, these represent *prima facie* decisions reflecting our intuitions about the meaning ascribed to the terms used. The point of this exercise is not so much to discuss the ontological nature of these properties, but rather *to explore and demonstrate the logical consequences of making these choices*. As we shall see, in some cases they will be contradictory with respect to the formal semantics of our metaproperties, although intuitive at a first sight. In our opinion, this proves the utility of a formal approach to ontology analysis and evaluation.

Entity
Everything is necessarily an entity. Our metaproperties assignment is -I-U+R. This is the most abstract property, indeed it is not necessary to have an explicit predicate for it.

Location
A location is considered here as a generalized region of space. Our assignment is +O~U+R. We assume the property to be rigid since instances of locations cannot change being locations. Identity is given by the fact that two locations are the same if and only if they have the same parts. This kind of criterion is fairly common, and is known as *mereological extensionality*. It applies to all entities that are trivially defined to be the sum of their parts. It is important to realize that this criterion implies that a location or region cannot "expand" – if so then the identity criteria would have to be different. So, extending a location makes it a different one. So we see that identity criteria are critical in specifying precisely what a property is intended to mean.

Amount of Matter
We conceptualize an amount of matter as a clump of unstructured or scattered "stuff" such as a liter of water or a kilogram of clay. Amounts of matter should not be confused with *substances*, such as water or clay; an amount of matter is a particular amount of the substance. Therefore, amounts of matter are mereologically extensional, so we assign **+O** to this property. As discussed above, they are not necessarily wholes, so our assignment is **~U**. Finally, every amount of matter is necessarily so, therefore the property is **+R**.

Red
What we have in mind here is the property of being a red *thing*, not the property of being a particular shade or color. We see in this case that it is useful to ask ourselves *what* the instances of a certain property are: Do we have apples and peppers in the extension of this property, or just their colors? In this case, we do include the apples and peppers, and not the colors. Red entities share no common identity criteria, so our assignment is **–I**. A common confusion here regarding identity criteria concerns the fact that all instances of *red* are colored red, therefore we have a clear *membership criterion*. Membership criteria are not identity criteria, as the latter gives us information about how to distinguish entities from each other. Having a color red is common to all instances of this property, and thus is not informative at all for identity.

A red amount of matter would be an instance of this property, which is not a whole, as would a red ball, which is a whole. Therefore we must choose **–U**, indicating that there is no common unity criterion for all instances.

Finally, we choose **–R** since some instances of *Red* may be necessarily so, and most will not. This weak and unspecific combination of metaproperties indicates that this property is of minimal utility in an ontology, we call them *attributions* [Welty and Guarino 2001]. We discuss this point further below.

Agent
We intend here an entity that plays a causal part in some event. Just about anything can be an agent, a person, the wind, a bomb, etc. Thus there is no common identity nor unity criterion for all instances, and we choose **-I-U**. No instance of *agent* is necessarily an agent, thus the property is **~R**. Clearly this assignment of metaproperties selects a particular meaning of *agent* among the many possible ones. See for example [Gangemi *et al*. 2003] for a discussion on the meaning of *causal agent* in WordNet.

Group
We see a group as an *unstructured* finite collection of wholes. Instances of *group* are mereologically extensional as they are defined by their members, thus **+O**. Since, given a group, we have no way to isolate it from other groups, no group is *per se* a whole, thus **~U**. In any case, like many general terms, *Group* is fairly ambiguous, and once again this choice of identity criteria and anti-unity exposes the choice we have made. Finally, it seems plausible to assume that every instance of group is necessarily so, thus **+R.**

Physical Object
We think of physical objects as isolated material entities, i.e. something that can be "picked up and thrown" (at a suitable scale, since a planet would be considered an instance of a physical object as well). Under this vision, what characterizes physical objects is that they are *topological wholes* – so we assign **+U** to the corresponding property.

For the sake of simplicity, we assume here that no two instances of this property can exist in the same spatial location at the same time. This is an identity criterion, so we assign **+O** to this property. Note that this is a *synchronic* identity criterion (see identity and unity, above) – we do not assume a common diachronic identity criterion for all physical objects.

Physical object is a rigid property, so we have **+R**. To see this, consider the alternative: there must be some instance of the property that can, possibly, *stop* being a physical object, yet still exist and retain its identity. By assigning rigidity to this property, we assert that there is no such instance, and that every instance of *Physical Object* ceases to exist if it ceases to be a physical object.

Living Being

Instances of *living being* must be wholes according to some common biological unity criterion. We don't need to specify it to assign **+U** to this property.

For identity, it is difficult to assume a single criterion that holds for all instances of living being. The way we, e.g. distinguish people may be different from the way we distinguish dogs. However, a plausible diachronic criterion could be *having the same DNA* (although only-necessary, since it does not help in the case of clones). Moreover, we can easily think of essential properties that characterize living beings (e.g., the need for taking nutrients from the environment), and this is enough for assigning them **+O**.

We assume *living being* to be a rigid property (**+R**), so if an entity ceases to be living then it ceases to exist. Notice that this is a precise choice that goes a long way to reveal our intended meaning: nothing would exclude considering life as a *contingent* (non-rigid) property; by considering it as rigid, we are indeed *constructing* a new kind of entity, justified by the fact that this property is very relevant for us.

Food

Nothing is necessarily food, and just about anything is possibly food. In a linguistic sense, 'food' is a role an entity may play in an eating event. Considering that anything that is food can also possibly *not* be food, we assign ~R to this property. We also assume that any quantity of food is an amount of matter and inherits its extensional identity criterion, thus +I and ~U.

Animal

Like *living being*, the identity criteria for *animal* may be difficult to characterize precisely, but we can devise numerous essential properties that apply only to them, or only-sufficient conditions that act as heuristics especially for diachronic identity criteria. Humans, in particular, are quite good at recognizing most individual animals, typically based on clues present in their material bodies. The undeniable fact is that we do recognize "the same" animal over time, so there must be some way that is accomplished. Therefore, we assign **+O**.

The property is clearly rigid (+R); moreover, being subsumed by *living being*, it clearly carries unity (+U).

Legal Agent
This is an agent that is recognized by law. It exists only because of a legal recognition. Legal agents are entities belonging to the so-called *social reality*, insofar as their existence is the result of social interaction. All legal systems assign well-defined identity criteria to legal agents, based on, for example, an id number. Therefore, it seems plausible to assign **+O**. Concerning unity, if we include companies (as well as persons) among legal agents, then probably there is no unity criteria shared by all of them, so we assign **-U**. Finally, since nothing is necessarily a legal agent, we assign **~R**. For instance, we may assume that a typical legal entity, such as a person, becomes such only after a certain age.

Group of People
A special kind of *Group* all of whose members are instances of *Person*. Identity and unity criteria are the same as *Group*, and thus we have **+I~U**. Finally, we consider *Group of People* to be rigid, since any entity which is a group of people must necessarily be such. Note here that having the same identity criteria does not imply having the same *membership* criteria, nor indeed anything at all about it, as the membership criteria for this property is clearly more refined than for *Group*.

Social Entity
A group of people together for social reasons, such as the "Bridge Club" (i.e. people who play cards together). We can't imagine a common identity criteria for this property, however we assume it is rigid and carries unity. **-I+U+R**.

Organization
Instances of this property are intended to be things like companies, departments, governments, etc. They are made up of people who play specific roles according to some structure. Like people, organizations seem to carry their own identity criterion, and are wholes with a functional notion of unity, so we assign +O+U+R.

Fruit
We are thinking here of individual fruits, such as oranges or bananas. We assume they have their own essential properties, and can clearly be isolated from each other. Therefore, +O+U+R seems to be an obvious assignment.

Apple
This likely adds its own essential properties to those of fruits, so we assign it +O+U+R.

Red-Apple
Red apples don't have essential metaproperties in addition to apples. Moreover, no red apple is necessarily red, therefore we assign +I+U~R.

Vertebrate
This property is actually intended to be vertebrate-*animal*. This is a biological classification that adds new *membership* criteria to *Animal* (has-backbone), but apparently no new identity criteria: +I+U+R.

Person
Like *Living Entity* and *Animal*, the *Person* property is +U+R. It seems clear that specializing from *Vertebrate* to *Person* we add some further essential properties, thus we assume that *Person* has its own identity criteria, and we assign +O.

Butterfly and Caterpillar
Like *Animal*, *Butterfly* and *Caterpillar* have +I+U. However, every instance of *Caterpillar* can possibly become a non-caterpillar (namely a butterfly), and every instance of *Butterfly* can possibly be (indeed, must have been) a non-butterfly (namely a caterpillar), thus we assign ~R to each.

Country
Intuitively, a country is a place recognized by convention as having a certain political status. Identity may be difficult to characterize precisely, but some essential properties seem to be clearly there, so +O. Countries are certainly wholes, so +U. Interestingly, it seems clear that some countries, like Prussia, still exist but are no longer countries, so we must assign ~R.

Analyzing Rigid Properties

The backbone taxonomy

We now focus our analysis on what we have called the *backbone taxonomy*, that is, the rigid properties in the ontology, organized according to their subsumption relationships. These properties are the most important to analyze first, since they represent the invariant aspects of the domain. Our sortal expandability and individuation principles guarantee that no element of the domain is "lost" due to this restriction, since *every element must instantiate at least one of the backbone properties*, that supplies an identity criterion for it.

The backbone taxonomy based on the initial ontology is shown in **Figure 2**.

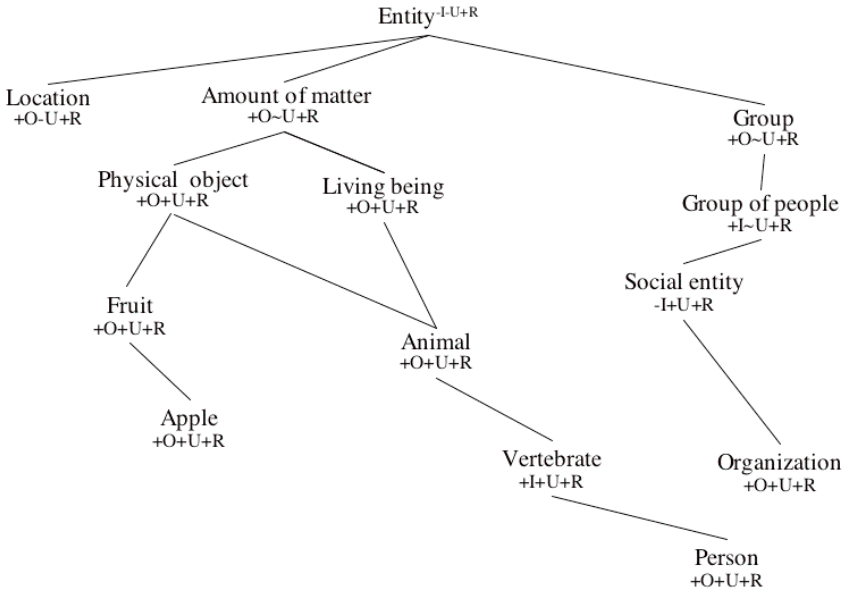


Figure 2. The initial backbone taxonomy with metaproperties

Backbone Constraint Violations

After making the initial decisions regarding metaproperties and arranging the properties in a taxonomy, we are then in a position to verify whether any constraints imposed by the metaproperties are violated in the backbone. These violations have proven to be excellent indicators of misunderstandings and improperly constructed taxonomies. When a violation is encountered, we must reconsider the assigned metaproperties and/or the taxonomic link, and take some corrective action.

Living beings are not amounts of matter. The first problem we encounter is between Amount of Matter and Living Being. The problem is that a ~U property can't subsume one with +U. While it certainly seems to make sense to say that all living beings are amounts of matter, based on the meaning we have assigned there is an inconsistency: every amount of matter can be arbitrarily scattered, but this is certainly not the case for living beings. A further reason against this subsumption link is in the identity criteria: amounts of matter have an extensional identity, that is, they are different if any of their parts is substituted or annihilated – if you remove some clay from a lump of clay, it is a different amount. Living beings, on the other hand, can change parts and still remain the same – when you cut your fingernails off you do not become a different person.

This is one of the most common modeling problems we have seen. Living beings are *constituted* of amounts of matter, they are not themselves the matter.

Natural language convention fails to capture this subtle distinction, but it is a violation of the intended meaning to claim that all living beings are mereologically extensional.

The solution here is to remove the subsumption link between these two properties, and represent the relationship as one of constitution.

Physical objects are not amounts of matter. Again, we see a violation since a ~U property can't subsume one with +U. This is yet another example of constitution being confused with subsumption. Physical objects are not themselves amounts of matter, they are constituted of matter. The solution is to make *Physical Object* subsumed directly by *Entity*.

Social entities are not groups of people. Another ~U/+U violation, as well as a violation of identity criteria. Social entities are constituted of people, but, as with other examples here, they are not merely groups of people, they are more than that. A group of people does not require a unifying relation, as we assume these people can be however scattered in space, time, or motivations. On the contrary, a social entity *must* be somehow unified. Moreover, although both properties supply their own identity criteria, these criteria are mutally inconsistent. Take for instance two typical examples of social entities, such as a bridge club and a poker club. These are clearly two separate entities, even though precisely the same people may participate in both. Thus we would have a state of affairs where, if the social entity was the group of people, the two clubs would be the same under the identity criteria of the group, and different under the identity criteria of the social entity. Note also that if a club changes its members it is still the same club, but a different group of people. The solution to the puzzle is that this is, once again, a constitution relationship: a club is constituted of a group of people.

Animals are not physical objects. Although no constraints involving metaproperties are violated in this subsumption link, a closer look at the identity criteria of the two properties involved reveals that the link is inconsistent. Animals, by our account, cease to exist at death, since *being alive* is an essential property for them. However their physical bodies remain for a time after: *being alive* is not essential to them. Indeed, under our assumption *no* physical object has *being alive* as an essential property. Now, if an animal is a physical object, as implied by subsumption, how could it be that it is at the same time necessarily alive and not necessarily alive? The answer is that there must be two entities, related by a form of constitution, and the subsumption link should be removed.
In this example, it is not the metaproperties, but the methodology requiring identity criteria in terms of essential properties that reveals the error.

Analyzing Non-Rigid Properties

Let us now turn our attention to the *non-rigid* properties, which – so to speak – "flesh out" the backbone taxonomy. In [Welty and Guarino 2001] we have dis-

cussed a taxonomy of property kinds based on an analysis of their metaproperties, which distinguishes three main cases of non-rigid properties: *phased sortals, roles,* and *attributions.* All these cases appear in our example, and are discussed below.

Among other things, the differences among these property kinds are based on a metaproperty not discussed here, the notion of *dependence.* Dependence is rather difficult to formalize, however a formalization not essential for an introductory understanding of the OntoClean methodology, so we shall rely on intuitive examples only.

Phased Sortals

The notion of a phased sortal was originally introduced by Wiggins [Wiggins, 1980]. A phased sortal is a property whose instances are allowed to change certain of their identity criteria during their existence, while remaining the same entity. The canonical example is a caterpillar. The intuition here is that when the caterpillar changes into a butterfly, something fundamental about the way it may be recognized and distinguished has changed, even though it is still the same entity. Phased sortals are recognized in our methodology by the fact that they are independent, anti-rigid, and supply identity criteria.

In the typical case, phased sortals come in clusters of at least two properties – an instance of a phased sortal (e.g., *Caterpillar*) should be able to "phase" into another one (*e.g., Butterfly*), and these clusters should have a common subsuming property providing an identity criterion across phases, according to the sortal individuation principle.

Caterpillars and butterflies. Consider now our example. *Caterpillar* and *Butterfly* appear in our initial taxonomy, but there is no single property that subsumes *only* the phases of the same entity. Our formal analysis shows that there *must* be such a property. After some thinking, we find what we need: it is the property *Lepidopteran*, which is +O+U+R. This is what supplies the identity criteria needed to recognize the same entity across phases.

Countries. The property *Country* does not, prima facie, appear to be a phased sortal, yet it meets our definition (+O~R). This is an example where reasoning on the metaproperties assignments and their consequences helps us to push our ontological analysis further: what are we talking of, here? Is it a region that occasionally becomes a country, and in this case acquires some extra (yet temporary) identity criteria? What happens when something is not a country any more? Does it *cease to exist*, or does it just undergo the change of a property, like changing from being sunny and being shady? While answering these questions, we realize we are facing a common problem in building ontologies, that of lumping together multiple meanings of a term into a single property. It seems there are two different interpretations of "country", one as a geographical region, and another as a geopolitical entity. It is the latter that ceases to exist when the property does not hold any more.

So there are two entities: the *Country* Prussia and the *Geographical Region* Prussia. These two entities are related to each other (e.g. countries occupy regions), but are not the same, and therefore we must break the current property into two.

We assign +O+U+R to Country, and +I-U+R to *Geographical Region*. The intuition is that countries have their own identity criteria, while geographical regions inherit the identity of locations. Countries clearly have unity, while this is not the case for arbitrary geographical regions. Both properties are now rigid. Interestingly enough, we replaced an anti-rigid property with two rigid properties.

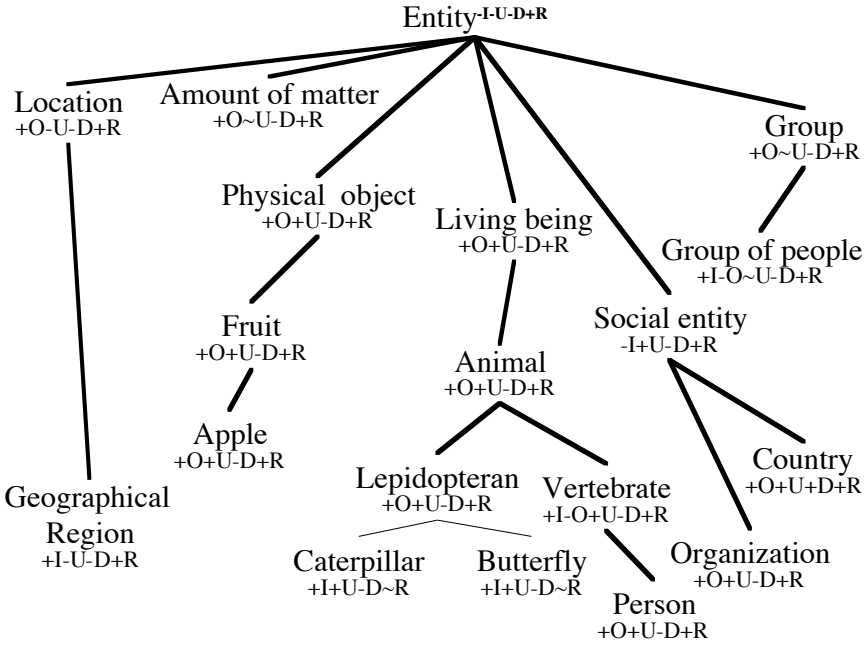
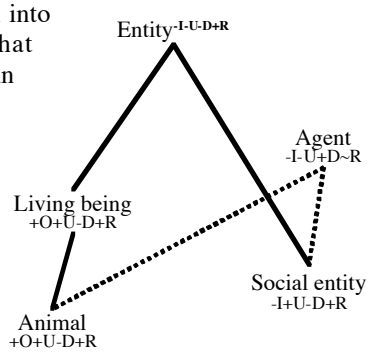


Figure 3. The taxonomy after backbone and phased sortals.

Roles

After analyzing phased sortals, we end up with the taxonomy shown in , and we are now ready to consider adding *roles* back into the taxonomy. Roles are properties that characterize the way something participates in a *contingent* event or state of affairs. It is because of such contingency that these properties are anti-rigid. Differently from phase sortals, roles do not supply identity criteria.

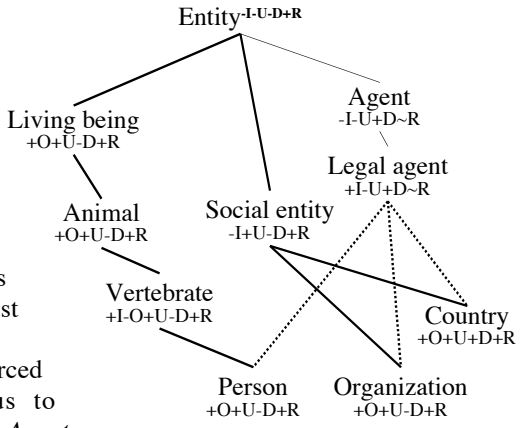


Agent. The analysis of roles often ex-

exposes subsumption violations concerning rigidity, in particular that a property with ~R cannot subsume a property with +R. Indeed, when we add the *Agent* property back to the backbone we see that it originally subsumed two classes, *Animal* and *Social Entity*. These subsumption links (shown on the previous page as dotted lines) should be removed, as they are incorrect.

This is a different kind of problem in which subsumption is being used to represent a type restriction. The modeler intends to mean, not that all animals are agents, but that animals *can be* agents. This is a very common misuse of subsumption, often employed by object-oriented programmers. The correct way to represent this kind of relationship is with a covering, i.e. all agents are either animals or social entities. Clearly this is a different notion than subsumption. The solution is to remove the subsumption links and represent this information elsewhere.

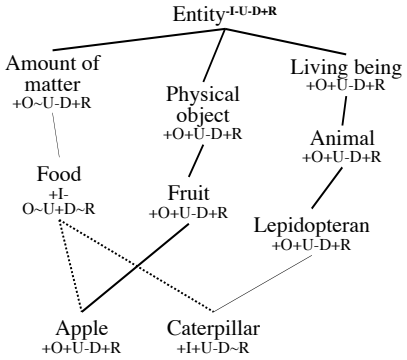
Legal Agent. The next problem we encounter is when the role *Legal Agent* is added below *Agent*, with its subsuming links to *Person*, *Organization*, and *Country*. Again, as with the previous example, we have a contradiction, an anti-rigid property cannot subsume a rigid one, so these subsumption links (shown as dotted lines at right) must be removed.



As with the *Agent* role, being forced to remove these links forces us to reconsider the meaning of the *Legal Agent* property. A legal agent is simply an entity recognized by law as an agent in some transaction or contract. Again, as with the *Agent* example, this is not a true subsumption link, but rather another type restriction. The links should be removed and replaced with a covering axiom.

Food. We chose to model the notion of food as a role, that is a property of things that may or can be food *in some state of affairs*. So nothing is essentially food – even a stuffed turkey during a holiday feast or an enormous bowl of pasta with pesto sauce may avoid being eaten and end up not being food (it's *possible*, however unlikely).

While our notion of what an apple means may seem to be violated by removing the subsumption link to *food*, the point is



that we have chosen to represent the property in a particular way, as a role, and this link is inconsistent with that meaning and should be removed. In this case, the links are probably being used to represent *purpose* (see, e.g., [Fan, et al, 2001]), not subsumption.

Attributions
The final category of properties we consider are *attributions*. We have one such property in our example, *Red*, whose instances are intended to be red things. We think that in general it is not useful to represent attributions explicitly in a taxonomy, and that the proper way to model attributions is with a simple attribute, like color, and a value, such as red. This quickly brings us to the notion of *qualities*, discussed in the related chapter of this handbook on Dolce, and we avoid that discussion here.

Attributions do, however, come in handy on occasions. Their practical utility is often found in cases where there are a large number of entities that need to be partitioned according to the value of some attribute. We may have apples and pears, for example, and decide we need to partition them into red and green ones. Ontologically, however, the notion of red-thing does not have much significance, since there is nothing we can necessarily say of red-things, besides their color. This seems to us a very good reason eliminate attributions from the backbone. The backbone taxonomy helps in focusing on the more important classes for understanding the invariant aspects of domain structure, whereas attributions may help in organizing the instances on an ad-hoc, temporary basis.

8.4 Conclusion

The final, cleaned, taxonomy is shown in . The heavier lines indicate subsumption relationships between members of the backbone taxonomy. Although it is not always the case, the cleaned taxonomy has far fewer "multiple inheritance" links than the original. The main reason for this is that subsumption is often used to represent things other than subsumption, that can be described in language using "is a". We may quite naturally say, for example, that an animal is a physical object, however we have shown in this chapter that this kind of linguistic use of "is a" is not logically consistent with the subsumption relationship. This results in many subsumption relationships being removed after analysis.

Acknowledgements

Many people have made useful comments on OntoClean, and have participated in its refinement. We would like to thank in particular Mariano Fernandez Lopez, Aldo Gangemi, Giancarlo Guizzardi, Claudio Masolo, Alessandro Oltramari, Bill Andersen and Mike Uschold. This work has been partially supported by the IST

Project 2001-33052 WonderWeb (Ontology Infrastructure for the Semantic Web) and the National project TICCA (Cognitive Technologies for Communication and Cooperation with Artificial Agents).

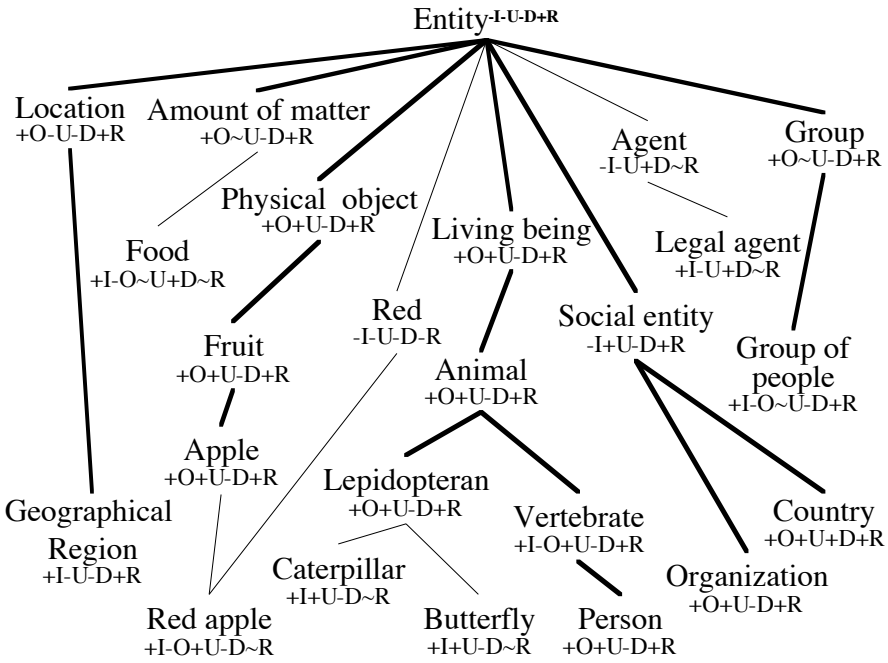


Figure 6. The final cleaned ontology.

References

Fan, James, Ken Barker, Bruce Porter, and Peter Clark. 2001. Representing Roles and Purpose. In *Proceedings of the 1st International Conference on Knowledge Capture (K-Cap'01)*. Vancouver: ACM Press.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. 2003. Restructuring Wordnet's Top-level. To appear on *AI Magazine*.

Guarino, N. 1998. Formal Ontology in Information Systems. In N. Guarino (ed.) *Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*. IOS Press, Amsterdam: 3-15.

Guarino, Nicola and Chris Welty. 2000a. Identity, Unity, and Individuality: Towards a formal toolkit for ontological analysis. In, Horn, W. ed., *Proceedings of ECAI-2000: The European Conference on Artificial Intelligence*. Pp. 219-223. Berlin: IOS Press. August, 2000.

Guarino, Nicola and Chris Welty. 2000b. A Formal Ontology of Properties. In, Dieng, R., and Corby, O., eds, *Proceedings of EKAW-2000: The 12th International Conference*

on Knowledge Engineering and Knowledge Management. Spring-Verlag LNCS Vol. 1937:97-112. October, 2000.

Guarino, Nicola and Chris Welty. 2000c. Ontological Analysis of Taxonomic Relationships. In, Veda Storey and Alberto Laender, eds., *Proceedings of ER-2000: The 19th International Conference on Conceptual Modeling.* Springer-Verlag LNCS Vol. 1920:210-224. October, 2000.

Guarino, Nicola and Chris Welty. 2002. Identity and Subsumption. In Rebecca Green, Carol Bean, and Sung Hyon Myaeng, eds., *The Semantics of Relationships: An Interdisciplinary Perspective*. Pp 111-125. Dordrecht:Kluwer.

Lowe, E. Jonathan. 1989. *Kinds of Being: A Study of Individuation, Identity, and the Logic of Sortal Terms*. Oxford:Basil Blackwell.

Quine, Willard. 1969. *Ontological Relativity and Other Essays*. New York:Columbia University Press.

Rector, Allan. 2002. *Are top-level ontologies worth the effort?* Panel at KR-2002. Toulouse, April, 2002.

Simons, Peter. 1987. *Parts: A study in ontology*. Oxford: Clarendon Press.

Welty, C. and Guarino, N. 2001. Supporting Ontological Analysis of Taxonomic Relationships. *Data and Knowledge Engineering*, **39**(1): 51-74.

Wiggins, David. 1980. *Sameness and Substance*. Oxford: Blackwell.