4 Principles of Best Practice II: Terms, Definitions, and Classification

We assume that, following the recommendations advanced in chapter 3, the appropriate scope of the ontology has been determined and the relevant domain information assembled. We assume also that the ontology builder has created a draft list of terms and associated these with a first draft set of definitions and a provisional *is_a* hierarchy. The next step is to use this list of terms to regiment the domain information in a systematic way, while at the same time allowing an improved understanding of the domain to generate improvements in the list of terms. The goal is to develop a representational artifact that is as logically coherent, unambiguous, and true to the facts of reality as possible.

There are three major facets of regimentation for domain ontologies: terminological, definitional, and location in an *is_a* hierarchy. We will treat each of these issues in turn, though the reader should bear in mind that there is a large degree of overlap and interdependency between the three sets of issues.

Principles for Terminology

Gather and Select Terminology

In chapter 3 we suggested that a good starting point for ontology building is to create a set of terms selected from the most commonly used terms in standard textbooks and in relevant domain ontologies. A first and indispensable step in any ontology development project is to perform due diligence in identifying existing ontology content that is relevant to the task in hand, and to evaluate this content for potential reuse, drawing on tools for searching ontologies such as the NCBO Bioportal (http://bioportal.bioontology.org).

The resultant list of words (or better: of common nouns and noun phrases) forms the first draft of what we can think of as a terminology for the domain in question. Such a terminology may have utility already for human beings, for example, in supporting

consistent use of language in exchanging information. For us, however, it has a more ambitious purpose, which is to enable the scientific information with which it is associated to be incorporated into the specific type of computer-based representational artifact that is an ontology, and for this a special sort of terminology will be needed.

The Gene Ontology (GO), by far the most successful ontology to date, was described by its creators as a "controlled vocabulary" to be used for regimenting the ways in which information about gene products in different model organisms is described. The problem it was designed to address is common across the whole of science: where multiple disciplinary groups are involved in the study of some scientific phenomenon of interest each will likely have its own idiosyncratic vocabulary. The problem is that there are *too many terms* for purposes of successful information exchange across disciplines. The GO provided a strategy to solve this problem by disseminating a set of "preferred terms" for use in describing attributes of gene products in a species-neutral way. Preferred terms are then used systematically by literature curators to describe experimental data appearing in published papers. These data then become more easily retrievable and combinable, in ways that overcome the problems caused by multiple conflicting vocabularies.

The success of the GO is due in large part to the fact that the influence of its creators was such that they were able to establish their chosen preferred labels as attractors for a large core of users in each of a variety of multiple interacting disciplines studying a variety of different species of organisms. To replicate this success, ontology builders today need to find a way of selecting terms that are as close as possible to the actual usage of a large fraction of those working in the relevant field without alienating those working in this field whose established terminologies involve the use of different terms. This goal can be achieved, in part, through disseminating the chosen preferred labels by using them in the curation of large bodies of data useful to the wider community, and—again following a practice pioneered by the GO—incorporating community-specific "synonyms" into the ontology alongside the preferred labels. Three principles of terminology construction can then be gleaned initially from the experience of the GO:

1. Include in the terminology terms used by influential groups of scientists for the most important types of entities in the domain to be represented.

2. Strive to ensure maximal consensus with the terminological usage of scientists in the relevant discipline. This may well involve working with domain experts, for instance in negotiating terminological compromises.

3. Identify areas of disciplinary overlap where terminological usage is not consistent. Look for and keep track of synonyms for terms already in the terminology list from these areas.

This strategy alone will work in cases where overlapping disciplines differ merely in their choice of words for the representation of identical entities. Where the terminologies employed by distinct disciplines in such overlapping areas differ in more substantial ways, more complex strategies need to be employed. Two ontologies may, for example, deal with the same phenomena, but at different levels of granularity (for example, molecule and cell); or they may differ in that one ontology deals with objects while another deals with processes; or one may deal with objects, while the other deals with images of objects.

In such cases multiple ontologies must be developed (or multiple branches of a single ontology), and the corresponding terms connected together through relations and through corresponding definitions and axioms. These are workable strategies because we are dealing with areas of *established* science, where we can assume that the disciplines in question will be consistent with each other as concerns their scientific content. Often it will be possible to formulate mapping rules—analogous to, for example, rules for conversion between different systems of scientific units—which allow assertions formulated using the terms from one discipline selected as synonyms in an ontology to be converted into assertions formulated using the terms selected as preferred labels.

What is at all costs to be avoided is the creation of entirely new expressions as preferred labels in ontologies to represent entities with which domain experts are already familiar under established names. Similarly, the ontologist should avoid using familiar terms with new and different meanings. To avoid confusion both in the encoding of information into the ontology and in the interpretation of such information by end users, the terminological choices of domain ontology builders should be as respectful as possible of the current terminology, usage, and practice of contemporary domain experts and of potential users. This leads to a fourth principle for terminology construction, which echoes the principle of reuse from chapter 3.

4. Don't reinvent the wheel. In term selection, stay as close as possible to the usage of actual domain experts. In terminology construction and ontology design, make use of as many existing resources (terminologies and ontologies) as possible.

Formatting Terminology

5. Use singular nouns.

The terms in an ontology should as far as possible have the grammatical form of singular nouns or singular noun phrases.

Two sorts of reasons support adoption of this convention. First (and this will be a common refrain in what follows when we deal with recommendations about syntax

and terminology), it is crucial that some syntactical standard, some rule of the road, be adopted and adhered to by all of those involved in ontology building in order to synchronize the multiple such efforts running in parallel at any given time. To see what happens when this rule is not followed, consider, for example, the case of MeSH,[1] whose hierarchy implies *is_a* relations such as

communism is_a political systems,

political systems is_a social sciences,

social sciences is_a behavioral disciplines and activities,

behavioral disciplines and activities is_a topical descriptor,

and so forth. Mixed use of singular and plural nouns may be perfectly appropriate for purposes such as the construction of library catalogs; it causes problems, however, when compiling information in a form that will be reasoned over.

The singular noun rule has been well tested in practice, and yields a simple and cost-free form of synchrony. There is also a principled reason for insisting that all terms in an ontology should take the form of singular nouns. This is because each such term is intended to refer not to some plural or collective entity, but rather either to a universal or to a defined class. In either case, its reference will be singular. There is only one universal *organism*, even if it has many instances, and there is only one defined class *cause of traffic accident*, even if it has many and diverse members.[2]

6. Use lowercase italic format for common nouns.

Along the same lines as principle 5, we recommend that when preparing ontology content for review by human beings *lowercase italic letters* be used for terms referring to universals or classes (this recommendation being based in part on the fact that initial capital letters are normally used in English to indicate proper names, which are names of instances ("Tom," "Seattle," "Jupiter"). Thus *cat*, not "Cat" or "CAT," and *eukaryotic cell*, not "Eukaryotic Cell" or "EUKARYOTIC CELL."

Some ontology editing programs require the use of underscore (eukaryotic_cell) or single quotation marks ('eukaryotic cell') or camel case (eukaryoticCell) in order to allow the beginnings and endings of noun phrases to be identifiable by the computer. Whichever traffic rule is chosen in this respect, the main goal is to ensure that the convention is consistently adhered to—again for reasons of cross-ontology coordination.

7. Avoid acronyms.

Avoid as far as possible the use of acronyms and abbreviations in formulating ontology terms. The rationale for this is that acronyms and abbreviations are too easy to create

locally—often, for example, by designers of databases for no reason other than to enable all column headings to fit on a single screen. The half-life of acronyms can be very short, and it is not unusual for those who work with databases (even, sometimes, a database's own creator) to forget what their acronyms originally meant. The goal of ontology, in contrast, is to create standard terminologies that can be employed and relied upon by anyone—in the present and in the future—working in a given discipline. Some acronyms and acronym-involving expressions have in some scientific idiolects become part of the language, as, for example, in terms such as "DNA" or "AIDS," or "ATPase"; they have become in this way safe from the possibility of being reused by new groups of researchers with different meanings. Apart from such cases, however, when selecting a primary label for an entry in an ontology a complete noun or noun phrase should in every case be used.

8. Associate each term in the ontology with a unique alphanumeric identifier.

The identifier is associated with the term in a given version of the ontology. Whenever the ontology is revised and published in a new version, then provided the term in question is not changed in this revision, its identifier can be preserved without change. Identifiers are needed for computer purposes—they will, for example, form the basis of the universal resource identifiers with which ontology terms will be identified in web-based systems. Figure 4.1 is a screen shot of a fragment of the Protein Ontology (PRO) that illustrates the approach we recommend.[3]

At the top of the hierarchy in figure 4.1 is the entry for "amino acid chain." Clicking on the entry will take the user to a human-readable definition of the term, along with other information about it. To the left of the term is its alphanumeric identifier PR:000018263, which uniquely identifies the location of this term in the PRO structure for purposes of computer programming and is used also in the creation of cross-links from other ontologies and databases back to the PRO. The identifier will be associated not merely with the term but also with its unique human-readable definition (for purposes of construction, maintenance, and use of the ontology by human beings), and also with the logically formalized version of this definition.

9. Ensure univocity of terms.

Terms should have the same meaning on every occasion of use. In an ontology, "cell" should refer always to the universal *cell*, "cancer" always to the universal *cancer*, and so on. The principle of univocity in ontology terminology development is difficult to maintain because it is so regularly violated both in ordinary and in scientific (and clinical) language. This occurs, first of all, because of ambiguous expressions, including "cell" itself, which has not only a biological meaning but also (related) meanings in

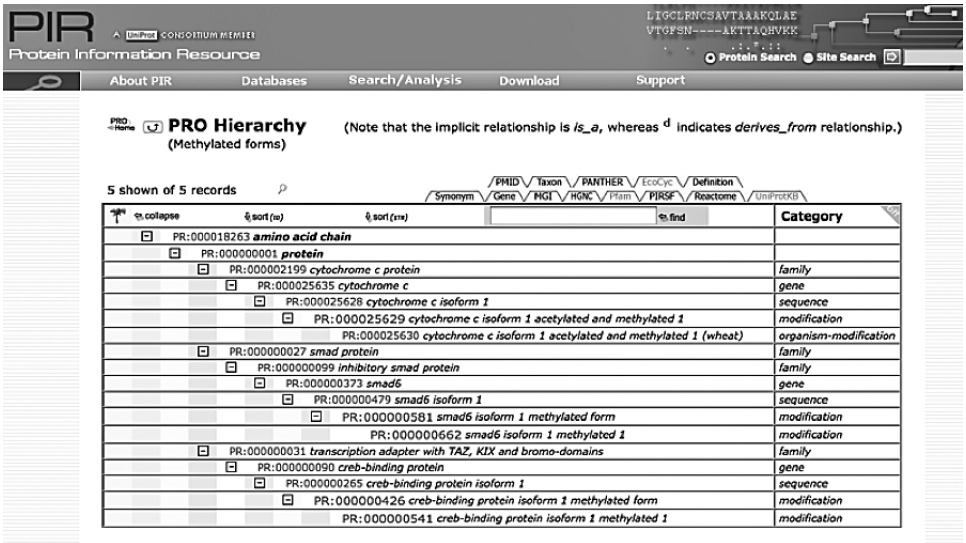


Figure 4.1
Screenshot from the Protein Ontology (PRO) browser containing terms identified by alphanumeric identifiers

relation to, for example, prison cells or cells in a spreadsheet. A more important reason, however, turns on the fact that departures from univocity occur because of the human tendency to use ellipsis in local circumstances (for example, to use "third left hip" to refer to the hip fracture patient in the third bed on the left-hand side of the ward). The reason for insisting upon univocity in the context of ontology design is quite straightforward. If the same term is used in different ways in different contexts, then the humans involved in ontology building are more likely to make errors. Ontologies are of course devised for use primarily by computers, and there the problems of ambiguity are alleviated by the use of unique alphanumeric identifiers for each ontology term. Working hard to avoid departures from univocity is still important, however, since experience shows that such departures are a source of human errors during ontology authoring and maintenance.

It should be noted here that what the principle of univocity specifically says is that every term in an ontology should have exactly one meaning. We do not rule out the presence in an ontology of multiple terms having the same meaning—but this should occur always through the device of declaring one such expression the preferred term, with which synonyms may then be associated according to the terminological needs of the different communities using the ontology.

An example of violation of the principle of univocity is the treatment of the term "disease progression" in the National Cancer Institute [NCI] Thesaurus (version dated August 2, 2004), which offered three different possible interpretations:

(I) Cancer that continues to grow or spread;
(II) Increase in the size of a tumor or spread of cancer in the body;
(III) The worsening of a disease over time. This concept is most often used for chronic and incurable diseases where the stage of the disease is an important determinant of therapy and prognosis.[4]

In definitions (I) and (II) "disease progression" is something that involves only cancer; in definition (III), however, "disease progression" involves the worsening of any disease over time. In the third definition, too, a "disease progression" is identified as a "concept," not as a process. This definition also contains a clause describing how the term is often used. Such information can be included in a comment that is associated with the term in question; for logical reasons, however, it should not be included in the definition itself.[5]

Note that the identified problems still persist in the current (June 30, 2014) version of the NCI Thesaurus, where we have, for example, two terms "cell," defined as meaning "any small compartment" and as "the individual unit that makes up all of the tissues of the body." The former is asserted to be a subtype of "conceptual entity"; the latter of "microanatomic structure."[6]

10. Ensure univocity of relational expressions.

Univocity applies also to the relational expressions used in ontology hierarchies, for example, *is_a* and *part_of*. The early years of ontology development were marked by a phenomenon of "*is_a* overloading" whereby "*is_a*" could mean in different contexts either subclass of, or instance of, or some confused mixture of both.[7] Similarly, "A *part_of* B" was sometimes used to mean that all As are part of some B, all Bs have some A as part, some As have some Bs as part, or again some confusing mixture of all of these.[8] For further details of how these issues are to be resolved, see chapter 7.

11. Avoid mass nouns.

Related to the issue of univocity is a basic distinction between *count nouns* and *mass nouns*. Count nouns, such as "cat," "petal," and "cell" refer to universals whose instances can be counted. Thus it is possible to ask *how many* questions (how many cats are there in this building?, how many petals on this flower?, and so on). Terms such as "water," "tissue," "meat," and "chemical substance" are often used as mass nouns. This means that they are used to pick out or refer to a more or less indefinite quantity of stuff. It is possible to ask *how much* water, meat, or chemical substance there is, for

example, in a given container; but not, without further qualification, *how many waters, tissues, meats*. Rather, we ask: "how many *glasses* of water are there?," "how many *pieces* of meat are there?," "how many *liters* of milk are there?," and so on. Now, however, we have replaced the original mass noun with a count noun (more precisely with a count noun phrase) as a means of ensuring that there will indeed be discrete portions of stuff that can be counted.

Certainly there are meaningful sentences involving mass nouns that have not been transformed into count nouns in this way, for instance when a nurse is instructed to store tissue in the freezer or to draw blood from a patient. Reflection reveals, however, that the corresponding transformation is here still being made—even if not explicitly. This is because the relevant amounts or containers are understood. In different contexts, moreover, terms like "blood" may be used to refer not merely to some specific amount of a patient's blood, but to an arbitrary portion or to the maximal portion of blood in a patient's body, and so forth—and "arbitrary portion of blood" and "maximal portion of blood," too, are perfectly acceptable from the point of view of the "avoid mass nouns" principle. A further reason for advancing this principle turns on the ambiguities that arise from the fact that terms like "blood" or "tissue" or "water" or "meat" or "aspirin" are often used to refer to *types*, rather than to particular *portions*, of the substances in question. These ambiguities are of particular importance for ontology builders, since it is precisely the division between types (universals) and instances (particulars), on which ontology is based.

Clearly masses of substances of different types do indeed exist in reality—but on the level of instances they always exist in large or small portions. Thus there is no sugar without there being some determinate portion of sugar; no luggage without there being some determinate number of suitcases and other luggage items. In addition, masses of substances exist on different levels of granularity: thus a mass of body tissue is at one and the same time a collection (mass) of cells.

To summarize: a mass noun such as "tissue" might be used to refer to one or more of the following:

- a portion of stuff within a larger portion of stuff (the tissue within an organ from which a doctor intends to take a sample);
- a discrete (detached) portion of stuff (such as tissue that has been grown independently in order to be placed inside an organ);
- a type of tissue under consideration (lung tissue vs. muscle tissue, healthy tissue vs. cancerous tissue); and
- a maximal or complete quantity of stuff (such as *all* of the tissue comprising the liver).

These different senses of the term "tissue" are involved in quite different theoretical and practical contexts and so it is important to keep them separate for purposes of ontology design. And even if only one such use of a mass noun like "tissue" were selected as the preferred label in an ontology, the mentioned ambiguities would still lead to problems of misuse of this term by human beings. It is for this reason that we recommend that mass nouns be avoided entirely when constructing ontologies. Instead phrases beginning with an appropriate prefix (such as "portion of," "maximal portion of," and so on) should be adopted. This solution has been embraced, for example, by the FMA ontology, which is the leading resource for terms relating (*inter alia*) to different tissue and other body substance types.[9]

To achieve this regimentation, we recommend transforming mass nouns such as "chemical substance" into count nouns by attaching "portion of" or some contextually appropriate equivalent operator to the beginning; thus "portion of chemical substance," "portion of tissue," and so on. Adopting this strategy makes it possible to treat seeming mass nouns as instances of either *fiat parts* or *object aggregates* (see chapter 5). The basic idea though, is that because mass nouns refer to different kinds of entities on different occasions of use, they should be avoided in favor of more ontologically transparent terminology.

12. Distinguish the general from the particular.

Up to this point we have stressed that an ontology is a representation of universals and defined classes. Particular entities—the instances of universals and the members of defined classes—are dealt with, for example, in databases or clinical notes or experimental logbooks. For us, this is a matter of the definition of the word "ontology." Certainly there are some who build ontologies that include an admixture of terms representing individuals—for example, the Standardized Nomenclature for Medicine (SNOMED) includes the term "National Spiritualist Church," which it treats as a subclass of *spiritual or religious belief*.[10] Our reasons for insisting that ontologies should be restricted exclusively to representations of what is general are manifold—but it will be sufficient, for the moment, to mention just one, which is illustrated all too well by the just-mentioned example from SNOMED. Namely, that departure from this principle is often associated with the making of errors: a *church*, however understood (whether as an organization or as a building) is not a special kind of *belief* as SNOMED would have it.[11]

Where an ontology needs to be supplemented by terms representing individuals, then this should be in some separate information artifact—corresponding to the distinction in the Description Logic community between a T-box (for "terminology") and an A-box (for "assertions").[12] The two artifacts can be combined for practical purposes

wherever necessary, forming what some call a "knowledge base." But the result is—again for definitional reasons—not an ontology, any more than the description or illustration of how a scientific theory has been applied in a specific series of experiments is itself a scientific theory.

Terms referring to universals and terms referring to instances should be clearly distinguished. For example, the common noun "teapot" as it occurs in a sentence such as "the teapot is a device for pouring tea" can plausibly be understood as referring to a type or universal *teapot*. The term "teapot" as it occurs in the sentence "John's teapot has been stolen" has to be understood as referring to a single particular teapot.[13]

Principles for Definitions

13. Provide all nonroot terms with definitions.

We have addressed the syntactic issues involved in regimenting the terminology of an ontology, for example by addressing conventions for use of nouns and noun phrases. An ontology is however above all a semantic artifact—it has to do with regimenting terms in such a way that they will be associated with specific *meanings*, and for this purpose the ontology must provide definitions, conceived as statements of the necessary and sufficient conditions which an entity must satisfy if the term, in its intended meaning, is to apply to that entity. To say that being an A is a *necessary condition* for being a B is another way of saying that every B is an A; to say that being an A is a *sufficient condition* for being a B is to say that every A is a B.

A definition, now, is a statement of a set of necessary conditions that are also jointly sufficient, as in the following example:

X is a triangle = def. *X* is a closed figure; *X* has exactly three sides; each of *X*'s sides is straight; *X* lies in a plane

Not every statement of necessary and jointly sufficient conditions is a definition.

- The stated conditions used to define the term A must themselves use terms which are easier to understand (and logically simpler) than the term A itself. (We return to this issue in our discussion of the "avoid circularity" principle to follow.)
- The stated conditions must be jointly satisfiable. Thus we cannot define, for example, a perpetual motion machine as a prime number that is divisible by 4, even though everything that is a perpetual motion machine is also a prime number divisible by 4, and everything that is a prime number that is divisible for 4 is also a perpetual motion machine (because neither of these things exists or, arguably, could possibly exist).

14. Use Aristotelian definitions.

The recommended best practice for creating definitions along the lines described earlier is to use the Aristotelian form:

S = def. a G that Ds,

where "G" (for: *genus*) is the immediate parent term of "S" (for: *species*) in the ontology for which the definition is being created. "D" stands for *differentia*, which is to say: "D" tells us what it is about certain Gs in virtue of which they are Ss. Ideally, the terms used in formulating the differentia D will themselves be terms taken from some ontology, where they will themselves be defined.

Consider, as a first example, Aristotle's own definition of "human":

human = def. an animal that is rational

Following this Aristotelian definitional structure ensures that the set of definitions in an ontology precisely mirrors the hierarchy of greater and lesser generality among its universals.

Some examples of Aristotelian definitions from the FMA are as follows:

cell = def. an anatomical structure that has as its boundary the external surface of a maximally connected plasma membrane

plasma membrane = def. a cell component that has as its parts a maximal phospholipids bilayer and two or more types of protein embedded in the bilayer

heart = def. an organ with cavitated organ parts, which is continuous with the systemic and pulmonary arterial and venous trees

liver = def. a lobular organ that has as its parts lobules connected to the biliary tree

lobular organ = def. a parenchymatous organ the stroma of which subdivides the parenchyma into lobes, segments, lobules, and acini[14]

Note that these and other definitions in the FMA ontology contain technical terms such as "cavitated organ part" and "venous tree" that are themselves defined at the appropriate places elsewhere in the ontology.

The Aristotelian definitional structure represents a basic format for the formulation of definitions that can be used regardless of ontological domain, and that is inherently directed at representing the position of each defined term within the relevant *is_a* hierarchy. Addressing the task of formulating definitions can thereby provide an extra check on the correctness of the ontology's *is_a* hierarchy, while creating the *is_a* hierarchy provides an easy first step in the formulation of each definition. These are advantages of the Aristotelian definitional structure, and reasons why it should be

adhered to as closely as possible when constructing domain ontologies. Other advantages include:

- Every definition, when unpacked (see below for an explanation of this term), takes us back to the root node of the ontology to which it belongs.
- Circularity is avoided automatically.
- The definition author always knows where to start when formulating a definition.
- It is easier to coordinate the work of multiple definition authors.

Aristotelian definitions work well for common nouns (and thus for the names of universals by which ontologies are principally populated). They do not work at all for those common nouns that are in the root position in an ontology, for here there is no parent term (no genus) to serve as starting point for definition. Root nodes in an ontology must therefore either be defined using as genus some more general term taken from some higher-level ontology, or they must be declared as primitive. Primitive terms cannot be defined, but their meanings can be elucidated (by means of illustrative examples, statements of recommended usage, and axioms—discussed in chapter 7).

15. Use essential features in defining terms.

The definition of a term captures what we can think of as the essential features of the entities that are instances of the designated type. The essential features of a thing are those features without which the thing would not be the type of thing that it is. They are also what we can think of as the constant elements in the structure of the entities in question—those elements that all instances of the relevant universal must possess.

Essential features of a triangle include being a plane figure and having three sides. Inessential features include the lengths of the lines making up the three sides.

For natural objects such as those studied by chemistry, biology, and physics, the essential features of a thing are typically features that play a prominent role in scientific explanation of its existence and behavior. Thus a definition of "portion of water" in terms of essential features would be as follows:

portion of water = def. portion of molecular substance each constituent molecule of which consists of two hydrogen atoms and one oxygen atom connected by covalent bonds

For artifacts, objects deliberately created (or, in some cases, selected) by human beings to be used to achieve certain goals, the essential features have to do with the purpose or use for which the artifact was created. Thus a knife is a tool for cutting things. Its essential features will thus include: having a blade made of a sufficiently hard

substance; the blade's having a sharp edge; its having a handle made of some hard substance; its being small and light enough to be manipulated by a single person, and so forth.

What sorts of features will be essential to the objects in a given scientific domain is specified in the relevant scientific literature, and what literature is relevant is determined in turn through specification of the ontology's scope. As can be seen from the definitions provided in our preceding list of examples, essential features of anatomical entities in the FMA include their location in the body, the sorts of anatomical entities that they have as parts, their spatial and physical relationships to other anatomical entities, and so on.

A useful way to go about determining the essential features of entities instantiating a given universal is the following. In the light of available scientific knowledge, attempt to imagine the subtraction and variation of the features of a typical entity falling under the corresponding term, checking at each stage to see whether the considered variation or subtraction would bring it about that the entity in question would no longer be an instance of the universal in question. If such variation or removal of a feature of the entity in imagination has this consequence, then it is highly likely that that feature is one of the essential features of entities of the given kind. Thus it is possible to imagine chairs with different sizes, shapes, materials, and colors; however, the second that one imagines a thing on which it is impossible for a normal human being to sit—for example, that it is made of a soft Jell-O-like material—then it is clear that, whatever the entity being imagined is, it is not a chair, and so the feature of being a thing upon which a normal human being can sit is at least a necessary condition for something's being a chair. Solidity is in this sense a constant (essential) feature of a chair; color a variable feature.[15]

A final point is that, with regard to defined classes, the essential features are just the features mentioned in the definition. Thus the features essential to being a member of the class of people in the United States who have cancer are just to be a person, to have cancer, and to be in the United States.

To see what goes wrong when definition authors fail to utilize essential features of the things being defined, consider again an example from HL7:

person = def. a living subject representing a single human being who is uniquely identifiable through one or more legal documents[16]

16. Start with the most general terms in your domain.

Another recommendation is to define the terms in an ontology from the top down. Thus, in accordance with our Aristotelian template for definitions, we start the process

of definition by defining the most general universals first and then working downward through the *is_a* hierarchy toward progressively more specific terms. Beginning with the root, terms on the next level down can be defined by determining relevant differentia in each case. This procedure can be reiterated as many times, and at as many different levels, as are necessary to address identified needs, but allowing the ontology author to start out from the most general level helps to keep things simple at the beginning, and provides a robust perspective from which to address the task of creating a comprehensive ontology at successively deeper levels.

A more general consideration in favor of the top-down approach derives from the proposition that an ontology should have a well-defined and clearly delimited domain, one that is determined, as far as possible, by some preexisting scientific discipline or unified practical field. Beginning with the most general types of entities determined by the specific target domain and working downward from there helps to rule out from the beginning the inclusion in the ontology of content that is not relevant to the chosen domain.

17. Avoid circularity in defining terms.

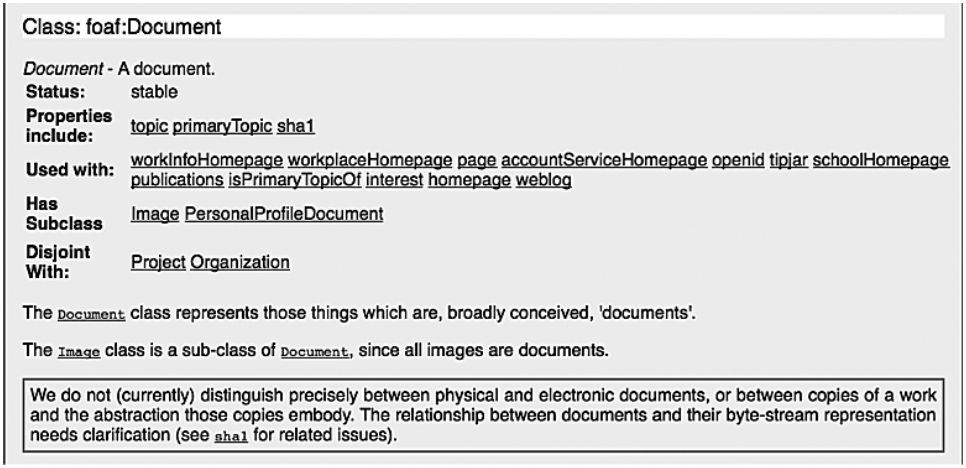
A definition is circular if the term to be defined, or a near synonym of that term, occurs in the definition itself, as for example in the following:

hydrogen = def. anything having the same atomic composition as hydrogen

poodle = def. anything having the biological structure and physical appearance characteristic of poodles

These definitions are circular because they provide no more information about the nature of the things the terms refer to than do these terms themselves. Since definitions are intended to explain the meaning of a term to someone who does not already understand it, using the term itself or some very similar expression in its own definition defeats the purpose of providing a definition in the first place. Figure 4.2 is a screen shot of the FOAF (Friend of a Friend) Vocabulary Specification 0.99, whose definition of "document" clearly exhibits circularity.

Avoiding circularity is important also for reasons having to do with the correct structuring of an ontology. If we think of a well-built ontology as having the structure of a graph, with a central backbone formed by the taxonomy of the ontology, then for reasons we shall present further on under the heading of "asserted single inheritance," it is important that every node in the graph is linked to the root by means of a unique chain of *is_a* relations. Avoiding circular definitions is a discipline that helps to ensure that this condition is satisfied.



Class: foaf:Document

Document - A document.

Status:	stable
Properties include:	topic primaryTopic sha1
Used with:	workInfoHomepage workplaceHomepage page accountServiceHomepage openid tipjar schoolHomepage publications isPrimaryTopicOf interest homepage weblog
Has Subclass	Image PersonalProfileDocument
Disjoint With:	Project Organization

The `Document` class represents those things which are, broadly conceived, 'documents'.

The `Image` class is a sub-class of `Document`, since all images are documents.

We do not (currently) distinguish precisely between physical and electronic documents, or between copies of a work and the abstraction those copies embody. The relationship between documents and their byte-stream representation needs clarification (see `sha1` for related issues).

Figure 4.2
Circularity in the FOAF Vocabulary Specification 0.99

18. To ensure the intelligibility of definitions, use simpler terms than the term you are defining.

The terms used in a definition should be more intelligible—for example, by being more scientifically, logically, or ontologically basic—than the term that is being defined. This is to promote the definitions' utility to human beings, for example, to support collaboration across disciplines, by allowing experts in one discipline to use the ontologies prepared for other disciplines as an initial means of orientation. Definitions are used in such cases in order to explain to people who do not know the meaning of a term what that meaning is. Someone who does not know the meaning of a term, especially a technical term, will not be helped by a definition that fails to satisfy the principle of intelligibility.

The following examples, again from HL7, will suffice to illustrate the problem we have in mind:

stopping a medication = def. change of state in the record of a Substance Administration Act from Active to Aborted

health chart entity = def. a health chart included to serve as a document receiving entity in the management of medical records[17]

In scientific contexts—and in the sorts of complex administrative contexts with which HL7 is concerned—it is inevitable that definitions will involve a certain degree

of specialized terminology. However if this terminology is to be managed in an effective way, then it is indispensable that for each new step in the direction of greater complexity and of specialization, the terms required are defined using terms already defined in earlier steps, potentially by means of definitional resources imported from other, external ontologies.

19. Do not create terms for universals through logical combination.

From the ontological realist's perspective, that a specific universal exists is never a matter of what can be inferred by logical means alone; it is always only something that must be discovered, through observation and application of the scientific method. Ontology is not analogous to set theory. It embraces what philosophers have referred to as a "sparse" theory of universals, which does not accept that the realm of universals is subject to any rule that allows arbitrary (for example) Boolean combinations.[18]

Thus from the fact that "*u*" names a universal, we cannot infer that non-*u* is a universal also, where "non-*u*" is defined in terms of logical negation as follows:

(*) *x* instantiates non-*u* = def. it is not the case that (*x* instantiates *u*)

Similarly from the fact that "*u*" and "*v*" name universals, we could not infer that "*u* and *v*" or "*u* or *v*" name universals, defined, respectively, by

(**) *x* instantiates *u* and *v* = def. *x* instantiates *u* and *x* instantiates *v*

(***) *x* instantiates *u* or *v* = def. either *x* instantiates *u* or *x* instantiates *v*

We recommend in particular that when building ontologies negative terms should be avoided entirely. That is, the ontology builder should assume that the universals are in every case positive, and so terms such as "nonrabbit" or "nonheart"—defined in accordance with (*)—should not be used, since there are no corresponding negative universals.

This principle applies not merely to terms representing universals, but also to terms representing defined classes. In relation to defined classes, we can formulate the rule as follows:

Avoid postulating complements of classes as entries in an ontology.

The complement of a class is the class containing all of the entities that do not belong in that class. Thus the complement of the class denoted by "dog" is the class denoted by "nondog," a class that includes not merely all cats, all fish, all rabbits, and so forth, but also all cardinal numbers, all musical instruments, all planetary bodies, and everything else that is not a dog.

There are, however, some cases where classes involving a negative element in their definition will properly be included in an ontology.[19] Thus, for example, prokaryotic cells are distinguished from eukaryotic and all other cells precisely by the fact that they lack a cell nucleus. This is, in effect, negative information used to define a class. However, the class of prokaryotic cells is not a complement class. If it were, then it would be equivalent to the class of noneukaryotic *things*, and would thus include, again, all musical instruments, all cardinal numbers, all planets, and so forth. Rather, "prokaryotic cell" is the name of a distinct class of cells that can be clearly defined. It is just that the definition of these cells itself includes some negative information (that they are cells that do not have a nucleus). Only the former formulation ("noneukaryotic thing") is a case of logical or "external" negation. In the latter ("prokaryotic cell"), we have rather internal negation, which from the realist point of view is perfectly acceptable for use in definitions.

The recommendation to avoid negative terms thus needs to be applied with care, since clinical research involves multiple sorts of defined classes referred to by terms in which prefixes like "non-" are used, but which are not defined along the lines of (*). An example is "nonsmoker," which is found in influential health-related terminologies such as SNOMED-CT and MedDRA, and used in scientific assertions such as "nonsmokers are less susceptible to cardiopulmonary diseases than are smokers."

The term "nonsmoker" is perfectly admissible provided it is defined, for example, along the lines of

nonsmoker = def. a human being who does not smoke,

which has exactly the Aristotelian form we recommended earlier.[20] Other putatively negative terms—such as "odorless," "colorless," "invisible," "unfriendly"—are similarly admissible, since they, too, can be defined in a positive way in terms of "lacks".

Another class of negative terms that should be avoided involves the use of negation operators that modify the associated phrases not logically but rather in some other way. Examples are

- canceled oophorectomy
- absent nipple
- unlocalized ligand

In each of these cases we are dealing not with special kinds of entity—there are, strictly speaking, no such things as canceled oophorectomies—but rather with special kinds of knowledge. When we use a term like "canceled oophorectomy" we are talking

in an abbreviated way about the fact that "oophorectomy" had earlier been entered into a surgeon's schedule and later removed. Ontologically misleading abbreviations of this sort should not be used when formulating terms in an ontology.

20. Definitions should be unpackable.

Definitions should be substitutable for their defined terms without a change in meaning. If we define an A as "a B that Cs," then we should be able to replace every occurrence of "an A" in a sentence with "a B that Cs," and the result will have the same meaning (and thus also the same truth value) as the sentence with which we began. This process of substitution is called "unpacking."

Note that the unpackability criterion holds only in what are called "extensional" contexts, which means contexts not governed by, for example, expressions such as "John believes," or "In the dictionary it is stated that." In all extensional contexts a defined term should be intersubstitutable with its definition in such a way that the result is grammatically correct and preserves both meaning and truth. The basic idea behind this principle is that, whatever a term refers to (in our case always a universal or defined class), the definition of the term should successfully refer to the very same thing. Thus in the FMA the reference of "heart" should be identical with the reference of the expression "organ with cavitated organ parts, which is continuous with the systemic and pulmonary arterial and venous trees."

The requirement that term and definition be intersubstitutable without affecting meaning is important not only for preserving truth across inference in automated reasoning contexts but also for ensuring intelligibility for human users and maintainers of ontologies. But interchangeability without effect on grammatical correctness is important for human beings also. If replacing a term with its definition results in a grammatically incorrect expression, this will impede the degree to which humans will successfully be able to use the ontology, and it will increase the likelihood of errors.

Principles for Taxonomies

Having set forth principles governing the formulation of definitions, we now move on to principles relating to the role of taxonomies within ontologies.

21. Structure every ontology around a backbone *is_a* hierarchy.

Each ontology should incorporate an *is_a* hierarchy having the structure of a directed acyclical graph with a single root. The terms in the ontology form the nodes of the graph, and the edges represent the *is_a* relation connecting each child to its immediate parent. (In mathematical terms the graph is a directed rooted tree.) The leaf nodes

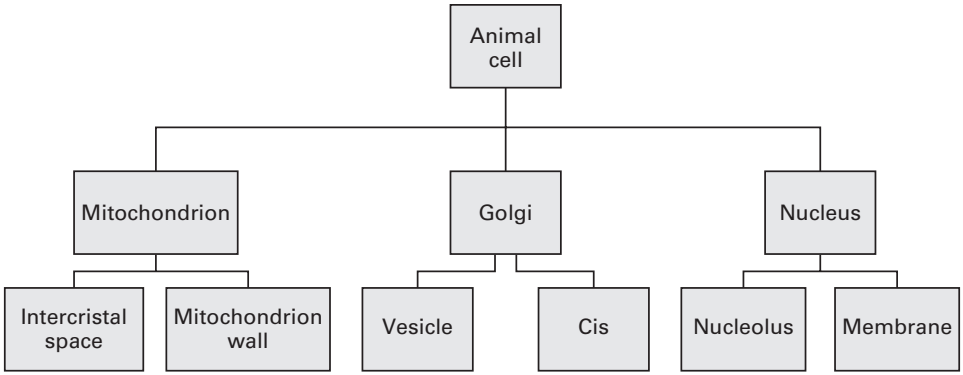


Figure 4.3
Fragment of a partonomy for *animal cell*

(lowest nodes of the ontology) represent the most specific universals or defined classes dealt with by the ontology in its current version. Leaf nodes play no special role in the ontology—since what is a leaf node today may no longer be a leaf node tomorrow because further subtypes have been incorporated.

In addition to those edges representing *is_a* relations, further edges in the graph represent other relations, for example, the *part_of* relationship, which generates what we might call a *partonomy* (as in figure 4.3). (We will discuss the *part_of* relation in more detail in chapter 7.)

Similarly, the relationship *derives_from* can be used to generate hierarchical structures among biological species, as in the simple phylogenetic tree illustrated in figure 4.4.

22. Ensure *is_a* completeness.

The ontology builder should ensure that every term in the ontology is included in its backbone *is_a* hierarchy, and that the ontology is *is_a* complete in the sense that every term in the hierarchy is joined to the root of the tree by a path constituted by successive edges in the graph. Thus if terms are added to the ontology to represent the component parts of the entities for which terms have already been included, then it should be checked that the ontology contains the parent terms needed also for these parts. We note that this principle stands in a mutually supportive relation with the requirement that all terms have definitions constructed using the Aristotelian template (see principle 14), for if this requirement is satisfied then *is_a* completeness will be guaranteed. On the other hand if *is_a* completeness is satisfied, then the creation of Aristotelian definitions is itself more straightforward.

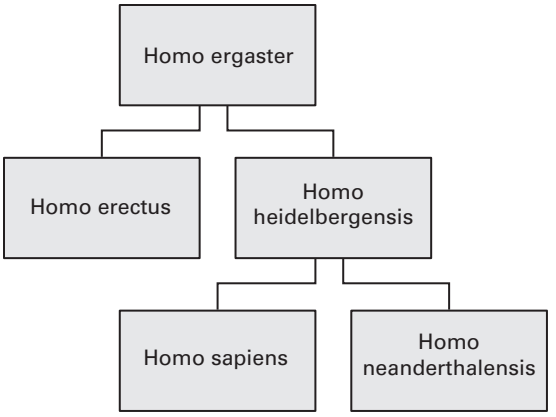


Figure 4.4
A phylogenetic hierarchy

Part of the process of ensuring *is_a* completeness is one of ensuring ontological agreement between terms and their parents. This is achieved by testing the validity of each assertion "A *is_a* B" in a given ontology by checking that, in the relevant domain, every instance of A is an instance of B.This check is needed whether "A" and "B" refer to universals or to defined classes. Bad practice in terminologies often involves the mixing of ontological categories across *is_a* relations, as, for example, in cases such as the following:

nonsmoker is_a finding of tobacco-smoking behavior (from SNOMED-CT);[21]

and, from the still-current version of the Gramene plant environment ontology:

virus is_a plant environment ontology,

unknown environment is_a plant environment ontology,

study type is_a plant environment ontology,

and so forth.[22] Such bad practice would be avoided by application of this simple rule.

23. Ensure asserted single inheritance.

We can think of assertions of the form "A R B," where "A" and "B" are nodes in an ontology and R is a relation that holds between them, as the ontology's *axioms* (for examples, see chapter 7). The ontology builder asserts these axioms during the construction of the ontology. When all the axioms have been asserted, however, then an ontology reasoner such as RACER or FACT (see chapter 8) may *infer* certain further statements. This allows us to distinguish between two different sorts of releases of ontologies: *asserted* and *inferred*.

Our *principle of asserted single inheritance* requires that the central backbone taxonomy of the ontology should be built as an asserted monohierarchy, which means: a hierarchy in which each term has at most one parent.

To speak of "inheritance," here, is to assert that everything that holds of a universal or defined class in an *is_a* hierarchy holds also—*is inherited by*—everything that appears below it in the ontology's *is_a* hierarchy. Because *cat* stands below (is a subclass of) *mammal*, it follows that cats are *vertebrate*, *air-breathing* animals whose females are characterized by the possession of mammary glands. In a similar way, everything that holds of *cell* holds also of *eukaryotic cell*, and everything that holds of *eukaryotic cell* holds also of *plant cell*.

There are a number of reasons for requiring single inheritance in our asserted *is_a* hierarchy. First, adherence to this principle brings certain computational performance benefits.[23] Second, because it ensures that all terms are connected by exactly one chain of parent-child relations to the corresponding root node of the asserted ontology, it provides an easy recipe for the creation of the sorts of definitions we will need in order to apply the Aristotelian template when defining our terms. Indeed, single inheritance is indispensable if the Aristotelian rule is to be applied successfully, since this rule works only if each (nonroot) term in the ontology has exactly one parent.

Third, adherence to single inheritance allows the total ontology structure to be managed more effectively, because it forces the ontology builder to think about each term before positioning it into the ontology, in order to ensure that it is being classified in conformity with the way its neighboring terms are classified. Our own experience with domain experts who are not ontologists and are building ontologies in a variety of different contexts has taught us repeatedly that, when scientists find it difficult to select between multiple parents for a term needing to be included in an ontology, the discipline imposed by the single inheritance principle is welcomed because it repeatedly leads to greater clarity of thinking on the part of those involved.

Fourth, adherence to the principle will make it easier to combine ontologies into larger structures—especially where ontologies need to be combined together automatically.

And finally, and most important, any benefits from multiple inheritance ontologies deriving from their easier surveyability (so that it is easier for human beings to find the terms they need by tracing through multiple parent paths) can be gained in any case by formulating the official (or "asserted") version of an ontology as an asserted monohierarchy and allowing the development of inferred polyhierarchies for specific groups of users. The application ontologies that then result are thus not required to satisfy the principle of asserted single inheritance.

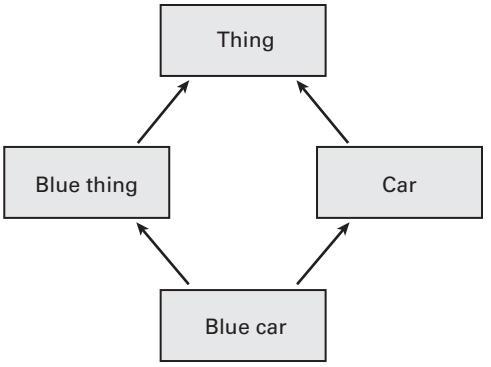


Figure 4.5
A simple illustration of multiple inheritance

Consider figure 4.5, which reflects the attempt to classify things using two different principles of classification (via color and via type of vehicle being classified). The figure does not satisfy the rule according to which (apart from the root) *every node within a hierarchy has exactly one parent*.

Here the principle of single inheritance is violated because two quite distinct *is_a* hierarchies have been run together—a hierarchy of things on the one hand, and a hierarchy of colors on the other. If we need to create such a combined hierarchy for application purposes (for instance to meet the needs of an automobile paint shop), then this can be achieved simply by combining the two mentioned hierarchies together and, using a reasoner, creating an *inferred* hierarchy that will depart from single inheritance but meet our application needs.

To see how resolving apparent multiple inheritance in a classification can be instructive and result in clarification of the correct definitions of the entities involved, consider an assertion such as "some human beings are mothers." What we mean by this assertion is that some human beings at some stage or stages in their lives play the role of mother in relation to other human beings. It may, now, be tempting to classify *mother role* as an example of multiple inheritance, as in figure 4.6.

A treatment along these lines seems to have the advantage that it would allow the variety of different sorts of mother role to be captured in the ontology. At the same time, however, it would gloss over some important distinctions. Recall what the *is_a* relation (represented here by arrows pointing upward) really means:

mother role *is_a* social role $\equiv$ every instance of mother role is an instance of social role

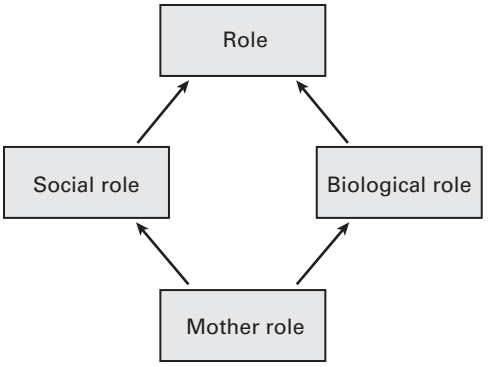


Figure 4.6
The mother role as a putative example of multiple inheritance

mother role *is_a* biological role ≡ every instance of mother role is an instance of biological role

But neither of these assertions is correct, unfortunately. Though often identified by the single word "mother," the social and biological senses of this term are not as closely connected as we might think. It is possible to be a biological mother without playing the social role of being a mother (for example, if one gives one's child up for adoption) and it is possible to play the social role of mother without being a biological mother (if one has adopted a child). Either of the two classifications in figure 4.7 comes closer to a correct way of thinking about the universal *mother* and how it should be classified; and, importantly, neither involves multiple inheritance.

Clearly, it is often possible to classify given individuals in more than one way. For example, pediatric surgeons might be classified both according to their patients and according to the procedures they perform. However, in such cases, the answer is not to create a single taxonomy with multiple inheritance. Rather, one begins by constructing separate "normalized" classifications each using only single inheritance and each built by downward population from a shared upper-level ontology. On this basis, one can then use the definitions of the terms that appear in these two ontologies to spell out the relations between the terms appearing in each. Computer reasoners can then use these definitions to create a compound ontology, in which single inheritance no longer holds, to address specific application purposes.[24]

24. Both developers and users of an ontology should respect the open-world assumption.

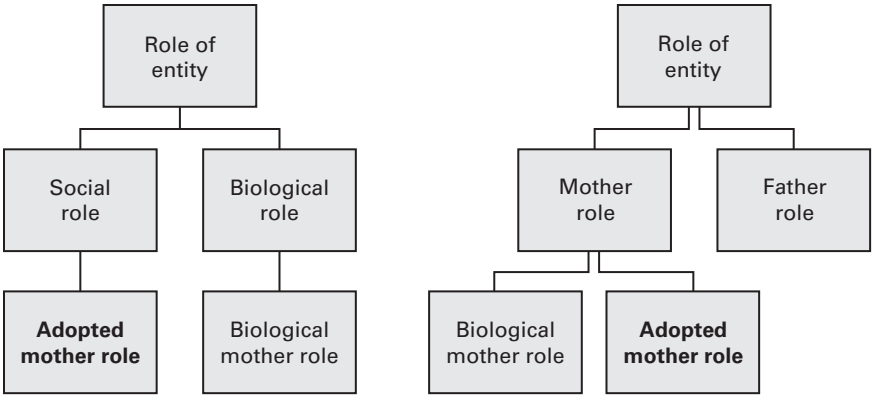


Figure 4.7
Mother classified without multiple inheritance

On the one hand, an ontology should be as complete as possible, given the specific purpose for which it is created. This means: representations of universals should be included in the ontology whenever they are relevant to the purposes of the ontology and fall within its scope. An ideal classification, of the sort that ontologies created for support of scientific research might seek to achieve, would include all existing domain universals at each level in the ontology's *is_a* hierarchy. Thus it would include, for example, all the universals that are discussed in current literature pertaining to the relevant domain. In nontrivial domains such as biology and medicine, of course, this ideal will never be achieved. This is because new scientific information will need to be accommodated with every scientific advance. In such domains an ontology will never be complete, and its authors should make this clear to potential users. Ontologies in general are built on the basis of the *open-world assumption*, or OWA, which means (here) that each ontology is built in a flexible manner to allow extension and correction, and is never put forward as providing a complete assay of the domain in question. Where a new phenomenon is encountered that is relevant to the ontology and falls within its scope, but for which no appropriate term is provided, the proper strategy is to identify the most-specific term in the ontology under which the phenomenon falls, and then to propose an appropriate child term to be added to the ontology hierarchy. The open-world assumption implies that no logical consequences follow from the fact that a given term is *not* included in an ontology.

25. Adhere to the rule of objectivity, which means: describe what exists in reality, not what is known about what exists in reality.

Ontologies created using our recommended methodology are representations of what exists in reality. What exists is not a function of our current state of biological knowledge. The universals treated by natural science in any given domain are discovered, not invented or created.

This is why it is necessary when building an ontology to take into account the best available scientific information about the reality in the salient domain. The goal is to systematically organize the terminological content of that information, paying attention to the essential characteristics of each type of entity. The current state of scientific knowledge is thus crucial for the building of ontologies. Yet the terms included in the ontology *do not refer to our current state of knowledge*. Rather they refer to the corresponding entities in reality. Thus an ontology should not contain classes like: "known allergy," "empirically confirmed boson," or "unclassified influenza." The ontology should not confuse *disease* with *diagnosis* and it should not confuse *result of measurement* with *magnitude that is being measured*.

Conclusion

The process of regimenting the domain information that an ontology contains involves the following steps. First, select the terms that are to be included in the ontology based on domain information that has already been gathered, and distinguish preferred labels and synonyms. Second, provide clear, scientifically accurate and logically coherent definitions for each of these terms. Third, explicitly recognize the location of each of these terms in a hierarchical classification of the domain information. These steps are to be carried out in accordance with the principles listed throughout the chapter. In the next chapter, we will show how these principles are applied in the context of Basic Formal Ontology.

Further Readings on Definitions and Categorization

Ceusters, Werner, and Barry Smith. "A Realism-Based Approach to the Evolution of Biomedical Ontologies." In *Proceedings of the AMIA Symposium*, 121–125. Washington, DC: AMIA, 2006.

Köhler, Jacob, Katherine Munn, Alexander Ruegg, Andre Skusa, and Barry Smith. "Quality Control for Terms and Definitions in Ontologies and Taxonomies." *BMC Bioinformatics* 7 (2006): 212.

Smith, Barry. "New Desiderata for Biomedical Ontologies." In *Applied Ontology: An Introduction*, ed. Katherine Munn and Barry Smith, 84–107. Frankfurt: Ontos Verlag, 2008.

Smith, Barry, and Werner Ceusters. "HL7 Rim: An Incoherent Standard." *Studies in Health Technology and Informatics* 124 (2006): 133–138.

Smith, Barry, Waclaw Kusnierczyk, Daniel Schober, and Werner Ceusters. "Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain." In *Proceedings of the 2nd International Workshop on Formal Biomedical Knowledge Representation* (KR-MED 2006), vol. 222, ed. Olivier Bodenreider, 57–66. Baltimore, MD: KR-MED Publications, 2006. Accessed December 17, 2014. http://www.informatik.uni-trier.de/~ley/db/conf/krmed/krmed2006.html.

Examples of Critical Reviews of Ontologies

Bodenreider, Olivier. "Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention." *Proceedings of the American Medical Informatics Association Symposium* 23 (2001): 57–61.

Ceusters, Werner, Barry Smith, and Louis Goldberg. "A Terminological and Ontological Analysis of the NCI Thesaurus." *Methods of Information in Medicine* 44 (2005): 498–507.

Ceusters, Werner, Barry Smith, Anand Kumar, and C. Dhaen. "Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected?" *Studies in Health Technology and Informatics* 102 (2004): 145–164.

Kumar, Anand, and Barry Smith. "The Unified Medical Language System and the Gene Ontology: Some Critical Reflections." *KI 2003: Advances in Artificial Intelligence* 2821 (2003): 135–148.