Open Peer Review on Qeios

Creating ontological definitions for use in science

Susan Michie[1], Robert West[1], Janna Hastings[1]

1 University College London, University of London

Abstract

Ontological definitions provide clarity and facilitate communication which accelerate the development of understanding and the accumulation of evidence about the world. It is hard to write good definitions. Too often they are partial, vague, or fail adequately to characterise the entity to which they refer. Ontological definitions are descriptions of entity classes or relationships that represent their essential properties in such a way that the defined entities are uniquely and fully specified. These definitions are then assigned a label to allow them to be used in scientific discourse. This article provides a brief guide to help with writing good ontological definitions. The standard format of such a definition is: A is a B that C, or involves or relates to C in some way, where A is the class being defined, B is a parent class and C describes a set of properties of A that distinguish it from other members of B.

Ontologies and entities

Ontologies are ways of representing knowledge in a form that can be used for searching, aggregation and inference by humans and computers. They increase clarity of conceptualisation and provide a basis for the development of an integrated, cumulative knowledge base (Hastings, 2017).

Although different formalisms can have different features, in the most general sense ontologies represent knowledge in the form of defined ‘entities’ and their properties. These properties can be expressed as relationships with other entities. Relationships link entities together (e.g., ‘is a’ in ‘Craving is a mental process’ and ‘is intended to measure’ in ‘The FTCD is intended to measure cigarette dependence’). Thus ontologies can be represented as collections of ‘triples’ of the form 'entity 1 – relationship – entity 2' forming a network in which every entity and relationship is fully defined, labelled and uniquely identified.

The most successful example of ontologies of this kind to date is the Gene Ontology (Ashburner et al., 2000), which was introduced to unify the description of gene functions to enable cross-species comparisons, and which has gone a long way to unifying the field of molecular biology and speeding up its advance.

Entities are anything in the universe that can be represented. This includes objects (e.g., cars), object parts (e.g., table legs), collections of objects (e.g. populations of people), object or site boundaries (e.g., country borders), sites (e.g., geographical regions), attributes (e.g., the colour red), processes (e.g., movement), process boundaries (e.g., start), and spatio-temporal regions (e.g., decade) (Arp, Smith, & Spear, 2016).

Dictionary definitions versus ontological definitions

At the heart of good ontologies are clear definitions of the entities contained in them. In order to write good ontological definitions it is important to understand the distinction between these and dictionary definitions.

Dictionary definitions are statements of the conventional meaning of words or phrases as used in language. Their purpose is to explicate the meaning(s) of terminology, which may differ from context to context. Thus they start with a word or phrase such as 'science' and they seek to capture its correct usage, e.g., 'the study of the nature and behaviour of natural things and the knowledge that we obtain about them'. Dictionaries can offer multiple definitions. For example, alternative definitions of 'science' might include 'a particular branch of knowledge,' which captures the sense in which we can refer to 'a science' rather than just 'science' as a process. Even within a single dictionary definition there may be multiple meanings embedded, which is the case for the first definition above: both the process of studying, and the knowledge obtained from such study, are referred to, despite the fact that these are different types of thing. Since multiple sorts of things are picked out by these dictionary definitions, corresponding to different senses or contexts in which the word can be used, such definitions can be a source of debate.

Ontological definitions are different, in that they aim to uniquely and unambiguously pick out a specific entity or class of entity (a specific type of thing) regardless of how that entity is usually referred to in language. Ontological definitions are thus more primary than the labels that are associated with them. For example, a class with the definition 'The intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment' can

be created, and indeed has been: this definition is associated with the National Cancer Institute Thesaurus class that has the unique online identifier: http://purl.obolibrary.org/obo/NCIT_C61397. This class has been given the label 'science', but it might equally well also have been given the more specific label 'scientific study activity' which more closely corresponds to its meaning as defined. In ontologies, while labels are assigned to classes, the labels are not the primary identifiers. The labels can be the same as labels used in common parlance, but referring to entities that are precisely defined, or they can be new words or phrases that are created for the purpose.

Thus, while dictionary definitions can be wrong, if they fail to capture an accurate meaning of the label they are defining, e.g., saying that the meaning of 'science' is 'the manufacture of articles of clothing', ontological definitions cannot be wrong in the same way because they specify a class regardless of what (if anything) that class is usually called. Similarly, while one can engage in debate about what the most accurate meaning of the word 'science' is, for example in terms of how far it should include the methods used for it or the body of knowledge arising from it, with ontological definitions one can allow particular labels to be used to refer to somewhat different entities (e.g., within different communities of practice) as long as it is clear what entity it is referring to based on the ontological definition.

The distinction between labels and definitions, with the latter being primary, is vital in areas of science where strong preferences exist for conceptualising the subject matter in different ways. It can be fruitless to debate which dictionary meaning should best be associated with which word in such cases as, if the community is divided between several options with no good reason to prefer one or the other, there will be no realistic prospect of sequestering a particular word for just one of those conceptualisations in a way that would satisfy all interested parties.

Having said this, there clearly is merit in limiting the use of the same labels to refer to different entities where possible because it is impracticable to keep looking up definitions to see what entities they refer to. Moreover, science requires a high degree of common conceptualisation in order to build, advance and use models, theories and evidence and apply accepted methodologies. Therefore, labels attaching to ontological definitions should aim as far as possible to adopt terminology that users can readily understand and fits with their usage.

Writing good ontological definitions

Where the entities in an ontology are physical objects that can be uniquely identified and objectively characterised, definitions are relatively straightforward. For example, the CHEBI ontology covers chemical entities that can be defined in terms of a number of properties including their chemical structure (https://www.ebi.ac.uk/chebi/). In many such cases (but by no means all, c.f., Akhondi, Muresan, Williams, & Kors, 2015) there is no ambiguity as to what the entity is, and a label can be given to it that is unique and unambiguous.

In other areas of science, writing good ontological definitions is much harder. This is particularly the case in the behavioural, human and social sciences, as these disciplines harness many terms that are in everyday use, but need to refer to entities that are more precisely defined. For example, the common term 'nudge' has been given a specific meaning in the context of behavioural science, as the use of implicit means to change behaviour. In another example, the common term 'effect' has a specific meaning in a technical, statistical context. On occasions, new terms are created to refer to such entities, but even in those cases there can be different usages and formulations that can cause confusion. For example, a core element of 'nudge theory' is the relationship between the environment and behaviour, which is called by the newly coined phrase 'choice architecture' in that context, but is called 'contingencies' in more traditional behavioural analysis (Simon & Tagliabue, 2018).

Ontologies relate entities to other entities, and definitions of entities need to reflect this. One of the most important such relationships is the taxonomic subsumption relationship, such that all members of one class are also members of another, which can be referred to as 'subclass', 'is-a' or 'type-of'. The reason this is important is because when X is a subclass of Y it inherits all the properties of Y and so can be used to improve economy of expression as well as being a powerful tool for inference. For example, we can create a class defined as 'A class of warm-blooded vertebrate animal having skin more or less covered with hair with young born alive and nourished with milk, except for the subclass of monotremes'; and we can label this class 'mammal'. We can then define a subclass, labelled 'dog' as 'A carnivorous mammal that typically has a long snout, an acute sense of smell, non-retractable claws, and a barking, howling, or whining voice'.

For this reason ontological definitions should take the form 'A [parent class] that [specification of characteristics that distinguish it from other members of the parent class, signified using terms such as 'that …', 'involving …', or 'relating to …']', or equivalent phrasing.

Guidelines for writing good ontological definitions

The guidelines in this article are taken from Seppälä, Ruttenberg and Smith's 'Guidelines for writing definitions in ontologies', rephrased in an attempt to make it easier for people who are new to ontologies to understand and use (Seppälä, Ruttenberg, & Smith, 2017).

Ontological definitions written according to these guidelines should meet the requirements of most ontologies, but there is no guarantee that a definition will perfectly capture what is intended. Definitions can be updated to improve them and there can be supplementary clarifications annotated alongside the definition, such as examples of how the entities are referred to in propositions. Review by members of a community is essential to obtain constructive feedback on a definition and to ensure that the definition is as clear as it can be.

In the following guidelines, sources for definitions are indicated, either intact or adapted; where not indicated, these were constructed for the purposes of illustration or sourced from terminologies that were inputs to the preliminary versions of the Behaviour Change Intervention Ontology during its development (see www.humanbehaviourchange.org).

Matters of substance

1. Definitions should take the form 'A [parent class] that [specification of characteristics that distinguish the entity from other members of the parent class]' or semantically equivalent phrasing. The parent class should be the next highest class in the ontology hierarchy, allowing the maximum information to be communicated by virtue of that class membership.

Example

Label: Perception

Good definition: A mental process of an animal that involves generation of a representation of part of the animal or its environment as neural activity.

Less good definition: The act or faculty of perceiving, or apprehending by means of the senses or of the mind; cognition; understanding. [dictionary.com]

2. The parent class should be a single class and not a combination of classes, so this part of the definition should not use 'and' or 'or'.

Example

Label: Beta-lactam

Good Definition: An organonitrogen heterocyclic antibiotic that contains a β-lactam ring. [CHEBI:27933]

Less good definition: A natural or semisynthetic antibiotic with a lactam ring [adapted from Merriam-Webster dictionary]

3. Definitions should uniquely identify all members of the defined class and exclude all entities not in that class.

Example

Label: Person

Good definition: An individual that is a member of the species homo sapiens.

Less good definition: A human that is member of a group or organization. [adapted from http://purl.allotrope.org/ontologies/material#AFM_0001083]

4. Definitions should avoid use of negations (saying what the class is not) unless required for linguistic clarity or when the class is inherently negative.

Example

Label: Infant

Good definition: A person between one month and 2 years of age

Less good definition: A person who is not a child or adult.

5. Definitions should not include other definitions nested within them. If there is a term being used in the definition that itself needs defining, another entry for that entity should be created in the ontology.

Example

Label: Immigrant

Good definition: A person who is currently a resident of a country having previously been resident of a different country.

Less good definition: A person with immigrant status: immigrant status being defined as having previously been a resident of a different country.

6. Definitions should avoid just using a term that has the same meaning as the label, or reference to another label that refers back to it in a circular fashion.

Example

Label: Addiction

Good definition: A chronic mental disorder that is realised as repeated occurrence of strong motivation to enact a behaviour and is acquired through experience, and results

in actual, or risk of, significant harm. [adapted from https://www.qeios.com/read/definition/551]
Less good definition: Being dependent on something.

7. Where possible, definitions should avoid use of expressions such as 'usually' or 'typically' unless these help to clarify what is included or excluded, or where the defined class is a fuzzy set (has ill-defined boundaries).
Example
Label: Epoch
Good definition: An extended period of time that has distinctive features or encompasses distinctive events.
Less good definition: An extended period of time usually characterised by distinctive features or events.

8. Where possible, definitions should avoid relying on special cases or lists. Ontological definitions should be intensional in the sense of stating the characteristics of the entities being defined rather than extensional in the sense of being lists of included instances or classes.
Example
Label: Intervention delivery through printed material
Good definition: A mode of delivery of an intervention that involves presentation of information, instructions or imagery by means of printed materials.
Less good definition: A mode of delivery of an intervention that involves leaflets, brochures, books, newspapers, newsletters, booklets, magazines, manuals or worksheets.

9. Definitions should avoid subjective or evaluative phrases or words.
Example
Label: Antisocial behaviour
Good definition: Behaviour that is judged by a defined population or group to contravene its moral precepts.
Less good definition: Behaviour that is undesirable or bad.

Matters of style

10. Definitions should not include abbreviations or alternate terms. These should go in a separate 'synonym' field.

Example

Label: Sudden infant death syndrome

Good definition: A syndrome that is characterized by the sudden death of an infant that is not predicted by medical history and remains unexplained after a thorough forensic autopsy and detailed death scene investigation.
[http://purl.obolibrary.org/obo/DOID_9007]

Less good definition: A syndrome (SIDS) that is characterized by the sudden death of an infant that is not predicted by medical history and remains unexplained after a thorough forensic autopsy and detailed death scene investigation. [adapted from http://purl.obolibrary.org/obo/DOID_9007]

11. Definitions should not include the words 'a type of' or similar at the beginning because that can be taken as read.

Example

Label: Outcome expectation

Good definition: An expectation that is about the consequences of an action.

Less good definition: A type of expectation that is about the consequences of an action.

12. Definitions should not include the label for the entity.

Example

Label: Outcome expectation

Good definition: An expectation that is about the consequences of an action.

Less good definition: An outcome expectation is an expectation that is about the consequences of an action.

13. Definitions should describe the entity that is being defined, not the label itself or the class that represents the defined thing. This is called 'the use-mention confusion'.

Example

Label: Person

Good definition: An individual that is a member of the species homo sapiens.

Less good definition: The most general classification of a person
[http://xmlns.com/foaf/0.1/Person]

14. Definitions should not include more information than is required to specify the class fully. Definitions are not theories or encyclopaedia entries.

Example

Label: Achieved short-cycle tertiary education

Good definition: The highest level of education that an individual has achieved that is below the level of a Bachelor's programme or equivalent.
Less good definition: The highest level of education that an individual has achieved that is below the level of a Bachelor's programme or equivalent. Entry into short-cycle tertiary education (ISCED level 5) programmes requires the successful completion of ISCED level 3 or 4 with access to tertiary education. Programmes at ISCED level 5, or short-cycle tertiary education, are often designed to provide participants with professional knowledge, skills and competencies. Typically, they are practically-based, occupationally-specific and prepare students to enter the labour market. However, these programmes may also provide a pathway to other tertiary education programmes. Academic tertiary education programmes below the level of a Bachelor's programme or equivalent are also classified as ISCED level 5. [adapted from International Standard Classification of Education]

15. Definitions should start with a capital letter and end with a full stop.
Example
Label: Need for competence
Good definition: A psychological need to believe oneself to be capable and effective at performing valued activities.
Less good definition: a psychological need to believe oneself to be capable and effective at performing valued activities

References

Akhondi, S. A., Muresan, S., Williams, A. J., & Kors, J. A. (2015). Ambiguity of non-systematic chemical identifiers within and between small-molecule databases. Journal of Cheminformatics. https://doi.org/10.1186/s13321-015-0102-6

Arp, R., Smith, B., & Spear, A. D. (2016). Building Ontologies with Basic Formal Ontology. MIT Press. https://doi.org/10.7551/mitpress/9780262527811.001.0001

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. Nature Genetics. https://doi.org/10.1038/75556

Degtyarenko, K., De matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., ... Ashburner, M. (2008). ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Research. https://doi.org/10.1093/nar/gkm791

Hastings J. (2017) Primer on Ontologies. In: Dessimoz C., Škunca N. (eds) The Gene Ontology Handbook. Methods in Molecular Biology, vol 1446. Humana Press, New York, NY: https://link.springer.com/protocol/10.1007/978-1-4939-3743-1_1

Seppälä, S., Ruttenberg, A., & Smith, B. (2017). Guidelines for writing definitions in ontologies. Ciencia Da Informacao. https://doi.org/10.18225/ci.inf.v46i1.4015

Simon, C., & Tagliabue, M. (2018). Feeding the behavioral revolution: Contributions of behavior analysis to nudging and vice versa. Journal of Behavioral Economics for Policy, 2(1), 91–97.