Creating ontological definitions for use in science

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

1 University College London, University of London

Abstract

Ontologies aim to represent the world in terms of uniquely defined classes and their properties which are expressed as relationships with other classes. They are becoming widely used in science to improve clarity, searching, inference and the ability to link data from different sources. Ontological definitions are descriptions that represent the essential properties of classes of entities (which include objects, object attributes and processes) that distinguish them from other classes. Classes have unique IRIs (Internationalised Resource Identifiers) that can be used for searching; they also have labels which are words of phrases that people can use to identify the class in passages of text or tables. This article provides a brief guide to help with writing good ontological definitions. The standard format of such a definition of a class, A, is: 'a B that C' or its semantic equivalent, 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.

Definitions

Addiction
Defined by Addiction Ontology

Ontologies and entities

Ontologies are ways of representing the world in a form that can be used for searching, data aggregation and inference by people and computers[1]. They increase clarity of conceptualisation and provide a basis for the development of an integrated, cumulative knowledge base[2].

Although different formalisms of ontologies can have different features, in the most general sense they represent the world as classes of entities and their properties. The properties of a class are expressed as dyadic relationships of that class with other classes. Relationships link classes together (e.g., 'is a' in 'craving *is a* mental process' and 'causes condition' in 'tobacco smoking *causes condition* lung cancer'). Thus ontologies can be thought of as collections of relationships between classes forming a network in which every class and relationship is fully defined, labelled and uniquely identified.

A successful example of ontologies of this kind is the Gene Ontology[3], which was introduced to unify the description of

gene functions to enable cross-species comparisons, and which has gone a long way to bringing together the field of molecular biology.

Classes represent entities, which encompass anything in the universe. This includes objects (e.g., planets), object parts (e.g., cell nuclei, atomic nuclei), collections of objects (e.g., human populations), object or site boundaries (e.g., country borders), sites (e.g., laboratory facilities, educational facilities), attributes (e.g., height, mass), processes (e.g., movements, patterns of neural activity), process boundaries (e.g., start), and spatio-temporal regions (e.g., decade)[4]. Entities also include immaterial things that contain information (e.g., data items, documents).

Dictionary definitions versus ontological definitions

At the heart of good ontologies are clear definitions of the classes of 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 meanings of words or phrases as used in language. Thus they start with a word or phrase such as 'science' and they seek to capture its conventional 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 in which the word can be used, such definitions can be a source of confusion and disagreement.

Ontological definitions are different in that they aim to uniquely and unambiguously delineate a single class of entity that is then given a unique identifier. It is also given a label that we can use to refer to in discourse. The label is a word or phrase that is unique within a given ontology. For example, we can delineate a class defined as 'A human activity that involves the systematic study of the structures and processes in the world through observation and the development of models and theories'. We can then give this class the label 'scientific activity' to avoid confusion with the more general term 'science' in common usage which can include both the activities and topics of study.

The primacy of the definition over the label in ontologies is particularly important in areas of science where strong preferences exist for conceptualising the subject matter in different ways. Thus different researchers may want to use labels such as 'anxiety' to mean somewhat different things, e.g., as a mental process versus as a disposition to experience a mental process, but it is crucial to be clear as to what it is that one is referring to. Ontological definitions do this, and labels can be framed to help with this, for example using the label 'anxiety feeling' for the mental process and 'anxiety disposition' for the disposition.

The same word or phrase may be attached to different ontological definitions by different people or teams if they are using

different ontologies. Thus no one ontology has a hegemony of the usage of a term. The different usages are made clear by the fact that the classes will have different IRIs consisting of a designation of the ontology, known as as the 'namespace', and the identifier within that ontology. Thus 'cell nucleus' in the Gene Ontology is 'GO:0005634' with the namespace GO while the same term in the Foundational Model of Anatomy Ontology has the identifier 'FMA:63840'. This avoids fruitless arguments about the 'true' meaning of labels and allows different perspectives to co-exist while ensuring that they are clearly expressed.

The publishing platform, Qeios, is particularly well suited for writing articles that use ontological definitions because it has the facility to tag terms used in articles with their definitions that are published in Qeios. The Addiction Ontology has made use of this facility by setting up, with Qeios, an automated process whereby classes that have reached a certain stability within the ontology are sent to Qeios to be published as definitions (e.g., addiction). An advantage of this approach is that authors can use terms that are stylistically appropriate within the text but ambiguous and tag them with a definition to make it easy for readers to identify the precise class being referred to.

Writing good ontological definitions

Writing good ontological definitions is difficult, as is evidenced by the fact that a large proportion of the ones found on the Ontology Lookup Service website (https://www.ebi.ac.uk/ols/index) have important limitations. To help with writing good ontological definitions we have taken an existing set of guidelines and extended them and attempted to make them clearer for non-specialist users[5]. We list nine matters of substance, and six of style.

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. Ideally parent classes should be able to trace ancestry all the way up to the most general type of entity in a unifying upper level ontology such as 'process' in Basic Formal Ontology[6].

Example

Label: Perceptual process

Good definition: A mental process which is a) produced by a causal process involving a part of the environment of the organism, and b) is experienced by the organism as being so caused, and c) in which the relevant part of the environment is thereby represented to the organism. http://purl.obolibrary.org/obo/MF_0000019 (Mental process is the parent class and as such carries a lot of the meaning. Its definition in this case is 'A bodily process that occurs in the brain, and that can of itself be conscious, or can give rise to a process that can of itself be conscious or can give rise to behaviour.' http://purl.obolibrary.org/obo/MF_0000020)

Less good definition: To become aware of, know, or identify by means of the senses.

[https://www.dictionary.com/browse/perceive] (No declaration of the parent class.)

2. The parent class should be a single class and not a combination of classes.

Example

Label: Beta-lactam

Good Definition: An organonitrogen heterocyclic antibiotic that contains a β-lactam ring.

http://purl.obolibrary.org/obo/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: Tobacco-containing product

Good definition: A product that is made of tobacco or has tobacco as a part and is used by people to ingest some tobacco constituent.

Less good definition: A product made or derived from tobacco that is intended for human consumption, including any component, part, or accessory of a tobacco product. (As worded, this definition, used by the US Food and Drug Administration, includes pharmaceutical nicotine products even though the intention is to exclude them.)

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: Human infant

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

Less good definition: A human being 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 human being who is currently a resident of a country having previously been resident of a different

country.

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

6. Definitions should avoid merely 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 mental disposition towards repeated episodes of abnormally high levels of motivation to engage in a behaviour, acquired as a result of engaging in the behaviour, where the behaviour results in risk or occurrence of serious net harm.

Less good definition: Being addicted to something.

7. Definitions should avoid use of expressions such as 'usually' or 'typically'.

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. Definitions should avoid relying on special cases or lists.

Example

Label: Behaviour change intervention delivery through printed material

Good definition: A mode of delivery of a behaviour change intervention that involves presentation of information, instructions or imagery by means of printed materials. (Although 'printed materials' is repeated from the label, added meaning is given by describing what those materials are being used for so the label is not merely being repeated.)

Less good definition: A mode of delivery of a behaviour change 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 population or group to contravene its moral precepts. (Subjectivity is avoided by making reference to the judgement of a population rather than asserting a value judgement oneself.)

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: Human being

Good definition: An extended organism that is a member of the species Homo sapiens.
http://purl.obolibrary.org/obo/MF_0000016

Less good definition: The term that is used to describe a member of the species Homo sapiens.

14. Definitions should not include more information than is required to specify the class fully. Definitions are not theories or encyclopaedia entries. Additional clarification should be included in a 'Comment' field for that class in the ontology.

Example

Label: Achieved short-cycle tertiary education

Good definition: A level of educational achievement that is below the level of a Bachelor's programme or equivalent. (The Comment field is provided to allow ontology developers to provide explanations and elaborations to help users understand how to use the classes.)

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

Conclusions

Ontological definitions that meet all of these guidelines are rare but as ontologies become more widely used, we believe it will be important for standards to be set and widely disseminated to make best use of them. Comments are welcome and this article will be updated as required in the light of these.

References

1. ^*Robert Arp, Barry Smith, Andrew D. Spear. (2015).Building Ontologies with Basic Formal Ontology. doi:10.7551/mitpress/9780262527811.001.0001.*

2. ^*Janna Hastings. (2016). Primer on Ontologies. doi:10.1007/978-1-4939-3743-1_1.*

3. ^*Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, et al. (2000).Gene Ontology: tool for the unification of biology. Nat Genet, vol. 25 (1), 25-29. doi:10.1038/75556.*

4. ^*Robert Arp, Barry Smith, Andrew D. Spear. (2015).Building Ontologies with Basic Formal Ontology. doi:10.7551/mitpress/9780262527811.001.0001.*

5. ^*Seppälä, S., Ruttenberg, A., & Smith, B.. (2017). Guidelines for writing definitions in ontologies.Information Science.*

6. ^*(2015). Basic Formal Ontology at Work. doi:10.7551/mitpress/8743.003.0011.*