Guidelines for writing definitions in ontologies

Selja Seppälä
Ph.D., University of Geneva, Switzerland
Postdoctoral Associate, University of Florida, Gainesville, United States.
http://seljaseppala.wordpress.com
E-mail: sseppala@ufl.edu

Alan Ruttenberg
M.S. Massachusetts Institute of Technology
Director of Clinical Data Exchange, School of Dental Medicine,
University at Buffalo, United States.
http://alan.ruttenbergs.com
E-mail: alanruttenberg@gmail.com

Barry Smith
Ph.D., University of Manchester, United Kingdom.
Professor of Philosophy, University at Buffalo, Buffalo, United States.
http://ontology.buffalo.edu/smith/
E-mail: phismith@buffalo.edu

Submetido em: 10/07/2017. Aprovado em: 05/09/2017. Publicado em: 28/12/2017.

ABSTRACT

Ontologies are being used increasingly to promote the reusability of scientific information by allowing heterogeneous data to be integrated under a common, normalized representation. Definitions play a central role in the use of ontologies both by humans and by computers. Textual definitions allow ontologists and data curators to understand the intended meaning of ontology terms and to use these terms in a consistent fashion across contexts. Logical definitions allow machines to check the integrity of ontologies and reason over data annotated with ontology terms to make inferences that promote knowledge discovery. Therefore, it is important not only to include in ontologies multiple types of definitions in both formal and in natural languages, but also to ensure that these definitions meet good quality standards so they are useful. While tools such as Protégé can assist in creating well-formed logical definitions, producing good definitions in a natural language is still to a large extent a matter of human ingenuity supported at best by just a small number of general principles. For lack of more precise guidelines, definition authors are often left to their own personal devices. This paper aims to fill this gap by providing the ontology community with a set of principles and conventions to assist in definition writing, editing, and validation, by drawing on existing definition writing principles and guidelines in lexicography, terminology, and logic.

Keywords: Definitions. Ontology. Textual definitions. Definitions in ontologies. Guidelines. Design patterns. Applied ontology.

Diretrizes para criação de definições em ontologias

RESUMO

Ontologias têm sido usadas cada vez mais para promover a reutilização de dados científicos, permitindo que dados heterogêneos sejam integrados via uma representação normalizada e única. As definições desempenham um papel central no uso de ontologias, seja por humanos seja por computadores. Definições textuais permitem que ontologistas e curadores de dados entendam o significado pretendido dos termos da ontologia, e usem esses termos de forma consistente com o contexto. Definições lógicas permitem que máquinas verifiquem a integridade da ontologia e raciocinem sobre os dados anotados em termos ontológicos, para que as inferências promovam descoberta de conhecimento. Portanto, é importante não apenas incluir em ontologias diversos tipos de definições, tanto em linguagens formais como em linguagens naturais, mas também garantir que essas definições atinjam padrões de boa qualidade e sejam úteis. Embora ferramentas como o Protégé possam auxiliar na criação de definições lógicas bem estruturadas, produzir boas definições em linguagem natural ainda é, em grande medida, uma questão de engenheiros humanos, a qual é apoiada, na melhor das hipóteses, apenas por poucos princípios gerais. Por falta de diretrizes precisas, os autores de definições são muitas vezes deixados à sorte com seus próprios princípios pessoais. Este artigo pretende preencher essa lacuna, fornecendo à comunidade ontológica um conjunto de princípios e convenções para auxiliar na escrita, edição e validação de definições, a partir de princípios e diretrizes existentes em lexicografia, terminologia e lógica.

Palavras-chave*: Definições. Ontologia. Definições textuais. Definições em ontologias.*

Directrices para creación de definiciones en ontologías

RESUMEN

Ontologías han sido usadas cada vez más para promover la reutilización de datos científicos, permitiendo que datos heterogéneos sean integrados vía una representación normalizada es única. Las definiciones desempeña un papel central en el uso de ontologías, sea por humanos sea por computadoras. Definiciones textuales permiten que ontologistas y curadores de datos entiendan el significado pretendido de los términos de la ontología, y usen esos términos de forma consistente con el contexto. Definiciones lógicas permiten que máquinas verifiquen la integridad de la ontología y raciocinen sobre los datos anotados en términos ontológicos, para que las inferencias promuevan descubierta de conocimiento. Por lo tanto, es importante no solamente incluir en ontologías diversos tipos de definiciones, tanto en lenguajes formales como en lenguajes naturales, pero también garantizar que esas definiciones atinjan patrones de buena calidad y sean útiles. A pesar de que herramientas como el Protégé puedan auxiliar a la creación de definiciones lógicas bien estructuradas, producir buenas definiciones en lenguaje natural aún es, en gran medida, una cuestión de ingenieros humanos, la cual es apoyada, en la mejor de las hipótesis, solamente por pocos principios generales. Por falta de directrices precisas, los autores de definiciones son muchas veces dejados a la surte con sus propios principios personales. Este artículo pretende llenar esta laguna, suministrando a la comunidad ontológica un conjunto de principios y convenciones para auxiliar en la escrita, edición y validación de definiciones, a partir de principios y directrices existentes en lexicografía, terminología y lógica.

Palabras-clave*: Definiciones. Ontología. Definiciones textuales. Definiciones en ontologías.*

INTRODUCTION

Ontologies are being used increasingly to promote reusability of scientific and other sorts of information and to address semantic interoperability issues by integrating heterogeneous data under a common, normalized representation. Definitions play a central role in the use of ontologies both by humans and by computers. Textual definitions allow ontologists and data curators to understand the intended meaning of ontology terms and to use these terms in a consistent fashion across multiple heterogeneous contexts. Definitions formalized in a logical language such as the Web Ontology Language (OWL) allow machines to check the integrity of ontologies and reason over data annotated with ontology terms to make inferences that promote knowledge discovery in areas such as biomedicine.

Therefore, it is important not only to include in ontologies definitions of both types, but also to ensure that these definitions can be created in a way that will contribute in a reliable fashion to their being useful.

According to a recent survey we conducted on definition practices in ontologies (SEPPÄLÄ, 2013; SEPPÄLÄ; RUTTENBERG, 2013), ontologists often lack adequate training in definition writing. This may partly explain the disparities in definition coverage observed in the OBO Foundry ontologies. It may also help to explain why so many ontologies lack either or both forms of definition (SCHLEGEL; SEPPÄLÄ; ELKIN, 2016). While tools such as Protégé can assist in creating well-formed logical definitions, producing good natural language definitions is still more a matter of human ingenuity supported at best by just a small number of general definition writing principles.

Our aim here is to fill this gap by providing the ontology community with a set of principles and conventions to assist in definition writing, editing, and validation. They draw from existing proposals from the disciplines of applied ontology, terminology, lexicography, and logic (ARP; SMITH; SPEAR, 2015; 2009; KELLEY, 1998; LANDAU, 2001; NDI-KIMBI, 1994; PAVEL; NOLET, 2001; SMITH, 2013; SVENSÉN, 1993; SWARTZ, 1997; VÉZINA et al., 2009).[1]

SOME PRELIMINARIES ON DEFINITIONS IN GENERAL

To make good use of the guidelines which follow, it is important to specify in which way the term *definition* is used. It is also important to know about the functions of definitions in ontologies. We briefly describe these aspects to provide background for the definition writing guidelines which follow.2

ONTOLOGICAL DEFINITIONS

The purposes of ontological definitions are different from those of definitions in standard dictionaries. In particular, the latter are required to define *all* lexical items, including those – for example 'of', 'because', 'the' – which do not refer but rather serve structural roles in natural language sentences. Ontological definitions, in contrast, are for 'terms', singular noun phrases which are content words that form part of a domain-specific vocabulary used by a group of experts to communicate about entities to which the terms refer.

HOW IS THE WORD 'DEFINITION' USED?

The word definition can be used in a number of ways. Here we focus on two: (i) *intension* and (ii) representational artifact. The former is the content or meaning of the definition and has a counterpart on the side of reality – the set of things, at any given time, to which the definition applies – which is called the *extension* of a definition. The latter is the form of the definition, the natural language text and potentially the accompanying axioms in some logical formalism that express this meaning. Good definition writing hinges on recognizing that these are the two sides of a single coin.

[1] The present guidelines follow the structure and to some extent the wording of the lexicographic definition writing guidelines prepared for the wordnet community (SEPPÄLÄ, Forthcoming).
[2] The preliminaries section is largely based on (SEPPÄLÄ et al., 2016; SEPPÄLÄ; RUTTENBERG; SMITH, 2016; SEPPÄLÄ; SCHREIBER; RUTTENBERG, 2014; SMITH, 2013), which discuss these aspects in more detail.

STRUCTURE OF A TEXTUAL DEFINITION

When used to refer to the natural language text of a definition, the term 'definition' itself can denote different forms: a sentence and a sentence fragment. Broadly, a definition has the canonical form X is a Y that Zs as in example (**1**) adapted from the definition of 'ligament' (synonym of 'skeletal ligament') in the Uberon multi-species anatomy ontology (UBERON).

(1) A **ligament** is a dense regular connective tissue connecting two or more adjacent skeletal elements.

Definitions in this form have a three-part structure:

1. a **definiendum** [*X*], i.e., the defined term;
2. a **definiens** [*a Y that Zs*], i.e., the part that expresses the definition content and that is called a **definition** in dictionaries;
3. a **copula** [*is*] that expresses an equivalence between definiendum and definiens.

In example (**1**), 'ligament' is the definiendum, which is connected to the definiens "a dense regular connective tissue connecting two or more adjacent skeletal elements" with the copula "is". Thus, in the wider meaning, 'definition' signifies a whole consisting of definiendum, copula, and definiens and takes the linguistic form of a sentence, as in (**1**).

More narrowly, it denotes the **definiens** alone and takes the linguistic form of a sentence fragment, as in (**2**):

(2) Dense regular connective tissue connecting two or more adjacent skeletal elements. (UBERON_0008846)

The guidelines presented in this paper are concerned with the formulation of the definiens. In dictionaries, definiendum and definiens appear in distinct entry fields and the copula is usually implicit, as in (**3**):

(3) ligament: *Dense regular connective tissue connecting two or more adjacent skeletal elements.* (UBERON_0008846)

The **definiens** – in italics – is subdivided into at least two parts:

- The ***genus*** **–** *connective tissue* – which constitutes the head of the definition (the *Y* part) and tells us what kind of thing the defined term denotes. When the genus is the immediate superordinate term it is called the ***genus proximus***.
- One or more ***differentia(e)*** (the *Z* part(s) or ***distinguishing features***) – *dense, regular* and *connecting two or more adjacent skeletal elements* – that tell us what characterizes the things referred to by the definiendum and what distinguishes them from the things referred to by the genus term and by all other terms under the same genus.

Example (**3**) thus tells us that "connective tissue" is the genus of the definition of 'ligament' and that a ligament is connective tissue of a certain type. The definition has three differentiae, "dense", "regular" and "connecting two or more adjacent skeletal elements". These together tell us that 'ligament' is more specific than 'connective tissue' and that 'ligament' differs from other subtypes of connective tissue in virtue of the fact that it is dense and regular and that it has the function of connecting two or more adjacent skeletal elements.

Each part serves as a condition for determining which things are members of the defined term's extension. In ontologies, definitions include ***necessary conditions*** that apply to all the members of the extension, but which may also apply to members of other term extensions, and, whenever possible, jointly ***sufficient conditions*****,** which allow a user to determine whether a given entity is a member of the extension.

(4) triangle: *A polygon with three edges and three vertices.* (adapted from WIKIPEDIA, 2017)

This definition tells us that every triangle is necessarily a polygon that has three edges and three vertices, and is sufficient for us to know that anything that is a polygon with three edges and three vertices is a triangle. Together, these

conditions form a ***definition by necessary and jointly sufficient conditions***, also called a ***classical definition***.

A definition that contains only necessary conditions that are not jointly sufficient to rule out instances that are not members of the extension is called a ***partial definition***. For example, defining a bird as an animal that lays eggs produces a definition that applies to all instance of birds but also to amphibians.

FUNCTIONS OF DEFINITIONS

Irrespective of the type of resource in which they appear and their context of use, definitions have two primary functions:

- a *cognitive* function: to augment or reconfigure (constrain or correct) our knowledge of the world relating to a given term, thus allowing us to better understand a term's meaning in a specific context of use and create new meaningful semantic connections and interpretations;
- a *linguistic* function: to describe or prescribe what we should understand when one or more speakers use the defined term in a specific context of use.

For example, at the cognitive level, the definition of the term 'ligament' in (**3**) tells us *what a ligament is* (some physical object), *what it does* (it has a connective function), and *what its properties are* (it has the physical qualities of being dense and regular). At the linguistic level, this definition tells us how the UBERON developers intended us to understand and use the term.

To fulfill the mentioned functions, the content and form of definitions have to be adapted to each context of use (for example teaching a child to speak, teaching undergraduates in a classroom, or annotating data with an ontology of microbiology) and to the target audience, which in ontologies is for example human curators of databases or developers of natural language processing software. This explains why ontologies have definition types different from those of dictionaries and other resources. In ontologies, specifying the intended meaning of a term involves clarifying and disambiguating this meaning so that it can be interpreted by a computer. Definitions in ontologies *stipulate* and *disambiguate* the meanings of terms in consistent and non-circular ways.

For ontologies, the primary context of use requires the inclusion of logical definitions since these are the sorts of definitions needed by a reasoner to perform logical operations, for example, with data annotated with ontology classes. Textual definitions are needed to facilitate ontology development and application by human users. In order to ensure consistency in ontology development and use, the textual and logical definitions of a term must convey the same type of content. This need for parallel semantic content along with the constrained logical properties conferred by the use of necessary conditions give logical definitions additional derived functions that are useful in definition writing and checking:

- *consistency checking and inference*: necessary conditions allows us to check whether an instance is consistent with the classes of which it is asserted to be a member, and to infer that the instance has all the properties that are necessary conditions, even if not explicitly stated.
- *instance classification*: sufficient conditions allow inferences to be made to the effect that given entities are instances of the class that is being defined;
- *taxonomic schematization*: the axioms of a class's logical definition can be used as a template for the axioms of its subclasses, as well as for the contents of the associated textual definitions;
- *regularizing expression and interpretation of facts*: the controlled vocabulary used in axioms allows us to check the intended meaning of a natural language expression in the textual definition.
- *avoiding circularity*: for any given ontology, there should be certain terms accepted as primitive (that are not defined in the ontology); other definitions

are then built in a step-by-step manner from these identified primitives (which may be defined in some superordinate ontology).

It is important to note that the differences between ontology and dictionary definitions result in varied sets of definition correctness criteria.

DEFINITION WRITING GUIDELINES

The guidelines we present in this paper are relevant for writing good definitions in ontologies and emphasize textual definitions. The principles are numbered for easy reference and, whenever possible, illustrated with definitions from existing ontologies.[3] The examples include a term (sometimes called a class label), the definition, and the source of the definition – which may be identified through a URI[4]. Our proposed solutions are based on the original definitions, unless otherwise specified.

1. Guideline: CONFORM TO CONVENTIONS

Authors should conform to the usual linguistic and lexicographic conventions of the English language (where relevant extended by discipline specific idiolects).

- 'Definition' in what follows means: 'definiens'. Thus, the definition should not include the term being defined (the definiendum) and it should not include the copula *is*; it should be limited to the definiens, a logically connected sentence fragment – see the definition of 'data set' in example (**5**).
- A definition should avoid punctuation marks other than commas, but should end with a period if the first letter is capitalized. Uses of other sorts of punctuation – e.g., parentheses, colons, slashes (as in 'and/or'), or semi-colons – are to be avoided, as they detract from the requirement that the definition should be a single logically unitary and unambiguous sentence fragment.
- Definitions should be written in a natural language supplemented by the technical terms and symbols used in specific disciplines where necessary.
- Nouns are divided into 'count' and 'mass' according to whether they can be pluralized ('cow' and 'datum' are examples of the former; 'cattle' and 'information' of the latter). Definitions of count nouns should start with an article ('a', 'an', 'the').

1.1. Guideline: HARMONIZE DEFINITIONS

Harmonize the definitions in the ontology in order that they all conform to a single set of typographical conventions and editorial guidelines. Example (**5**) illustrates a case where the needed consistency is lacking. The definitions of 'data set' and 'measurement datum' in the Information Artifact Ontology (IAO) should either all be limited to the ***definiens*** (recommended) or all contain the ***defined term and the 'is' copula***.

(5)

× **data item**: ***A data item is*** *an information content entity that…* (IAO_0000027)

✓ data set: **A data item that is** an aggregate of other data items of the same type that… (IAO_0000100)

× **measurement datum**: ***A measurement datum is*** *an information content entity that…* (IAO_0000109)

Where it is agreed by the ontology editors that a specified limited vocabulary will be used in definitions, the wording of the definitions should also be checked to ensure conformance.

[3] Note that the sources may have been edited since we accessed them on BioPortal (http://bioportal.bioontology.org, accessed July 14, 2017) and OntoBee (http://www.ontobee.org, accessed July 14, 2017).

[4] The ontologies referenced in the examples are listed at the end of the guidelines. All the CURIEs are prefixed with http://purl.obolibrary.org/obo/ unless otherwise specified in the ontology reference list.

2. Guideline: PRINCIPLES OF GOOD PRACTICE

The following are principles for working with definitions that should be followed to promote good practice.

- Very high-level ontologies will have many terms — such as 'entity' — which are so general that they are not capable of being defined without circularity; these should be marked as 'primitive'; they should be accompanied by elucidations and examples of use.
- Definitions should be unique (i.e., no two terms in a single ontology should share what is, logically speaking, the same definition).[5]
- When referring to other classes, use the controlled vocabulary specified in the ontology (i.e., the terms in the ontology labels) to produce consistently written definitions across the ontology.
- Include a single definition per class. This means that there should be a single 'definition' annotation property (IAO_0000100), which should contain a single definition (see also principle 6.2).
- Cite your sources and do it in a separate annotation property, preferably using the IAO 'definition source' annotation property (IAO_0000119).
- Credit authorship, preferably using the IAO 'term editor' annotation property (IAO_0000117).

3. Guideline: USE THE GENUS-DIFFERENTIA FORM

A definition should have the genus-differentia form, where the genus anchors the defined entity to a known entity at a higher level of generality, and the differentia (or differentiae) picks out those cases within this higher level that fall under the term defined. In example (**6**), only the first definition of mammal is correctly structured. It has a ***genus*** and three *differentiae*; the second definition only has a genus followed by example subtypes; the last definition lacks a genus (including the definiendum instead) and has one differentia.

(6)
- ✓ mammal: ***a vertebrate*** *that has hair, gives live birth, and nurses its young*
- × mammal: ***a vertebrate*** *like, a cat, dog, or whale*
- × mammal: *Mammals nurse their young*.

3.1. Guideline: INCLUDE EXACTLY ONE GENUS

A definition should always have one and only one genus (7). The genus is a superordinate term that tells us with what kind of thing we are dealing. Definitions of things like colors are expressed in the nominal form and start with a genus such as "a quality". Similarly, definitions of processes and other entities that unfold in time, which are often expressed by verbs are, in ontologies, defined with nominal phrases.

In example (7), the definition of 'recombinant vector' in the Ontology for Biomedical Investigations (OBI) does not have a genus. This can be fixed by adding its *genus proximus*, which in OBI is 'processed material'.

(7)
- × recombinant vector: *A recombinant vector* ***is created by*** *a recombinant vector cloning process, …* (OBI_0000731)
- ✓ recombinant vector: ***A processed material*** created by a recombinant vector cloning process, …

3.1.1. Guideline: USE THE *GENUS PROXIMUS*

Use the *genus proximus*, that is, the closest parent term. This ensures that the genus is specific enough and that all the terms at the same level have the same genus.

[5] *OBO Foundry Principles: Textual Definitions* at http://obofoundry.org/principles/fp-006-textual-definitions.html [accessed: July 14, 2017].

As an ontology is developed and new classes interposed between existing levels, the *genus proximus* may also change and the corresponding definitions need to be updated. Example (**8**) shows two definitions of sibling classes: the definition of 'data set' includes the correct *genus proximus* ('data item'), whereas the definition of 'measurement datum' does not, as it includes the parent of its *genus proximus* ('information content').

(8) × **data item**: *A data item is an **information content entity** that…* (IAO_0000027)

data set: *A **data item** that…* (IAO_0000100)
measurement datum: *A measurement datum is an **information content entity** that…* (IAO_0000109)

✓ **data item**: *An **information content entity** that…*

data set: *A **data item** that…*

measurement datum: *A **data item** that…*

3.1.2. Guideline: AVOID PLURALS

The genus of a definition should have the same syntactic properties as the definiendum. According to the naming conventions contained in the OBO Foundry Principles,[6] which are also in line with good terminology practice accepted elsewhere, terms in ontologies should be written "as if writing in plain English". This includes writing them in the singular form. Therefore, the genus of a definition should also be in the singular form. In example (**9**), the definition of 'myeloablative agonist' in the National Cancer Institute Thesaurus (NCIt) has a plural genus, which can be replaced by its singular form. Note that the rest of the definition needs to be adapted to accommodate this change.[7]

(9) × Myeloablative Agonist: ***Agents** that destroy bone marrow activity. They are used to prepare patients for bone marrow or stem cell transplantation.* (NCIt:C1711)

✓ myeloablative agonist: ***An agent** that destroys bone marrow activity and that is used to prepare patients for bone marrow or stem cell transplantation.*

3.13. Guideline: AVOID CONJUNCTIONS AND DISJUNCTIONS

The genus of a definition should be a single word: not a conjunction (**10**) and also not a disjunction (**11**). A conjunction or disjunction in a genus can be avoided by using a more general genus (**10**) and, if applicable, by adding a differentia specifying the things denoted by the conjunction or disjunction, as in (**11**). In example (**10**), we avoid the conjunction by using a more general term and adding the elements of the conjunction as differentia – note that the new genus might have to be added to the ontology if it falls within its scope. In example (**11**), we avoid the disjunction by using a parent term.

(10) × Cell Culture System: ***Systems and reagents** for the propagation of cells in tissue culture.* (NCIt:C19147)

✓ cell culture system: ***A system** comprising reagents and other systems which together enable propagation of cells in tissue culture.*

(11) × analyte: *The **sample or material** being subjected to analysis.* (NCIt:C128639)

✓ analyte: *A **spe/cimen** being subjected to analysis.*

3.1.4. Guideline: AVOID CATEGORIZERS

The role of the genus of a definition is to categorize the denoted thing under a more general type. Therefore, it is not necessary to add an extra categorizing expression, such as "type of" and "sort of". These can usually be omitted, as in (**12**).

[6] *OBO Foundry Principles: Naming Conventions* at http://obofoundry.org/principles/fp-012-naming-conventions.html [accessed: June 21, 2017].

[7] In the corrected version of this example, we also edited the rest of the definition to comply with the single fragment principle.

(12) × immunogen: ***A type of antigen*** *capable on its own of inducing an immune response.* (CHEBI_60816)

✓ immunogen: ***An antigen*** *capable, on its own, of inducing an immune response.*

4. Guideline: AVOID USE/MENTION CONFUSION

A definition should tell us about the thing that is denoted by the definiendum, not about the term itself or the class that represents the defined thing. Making this confusion is called the ***use/mention confusion***. It is illustrated in examples (**13**) and (**14**).

(13) × miosis AE: *Miosis AE is an eye AE that* ***is a term with various definitions****, which generally include constriction of the pupil.* (OAE_0001998)

✓ miosis AE: *An eye AE that generally includes constriction of the pupil.*

(14) × mammal*:* ***Representation*** *of an animal that feeds its young with milk.* (ARP; SMITH; SPEAR, 2015)

× mammal*:* ***A term that refers to*** *an animal that feeds its young with milk.*

✓ mammal: ***An animal*** *that feeds its young with milk.*

The first two examples in (**14**) confound the class and the term with the thing that they represent, respectively.

5. Guideline: INCLUDE NECESSARY AND, WHENEVER POSSIBLE, JOINTLY SUFFICIENT CONDITIONS

Definitions should avoid non-defining information. They must consist exclusively of necessary conditions which should, whenever possible, be jointly sufficient.

5.1. Guideline: AVOID ENCYCLOPEDIC INFORMATION

Definitions should avoid including information that goes beyond the minimum useful information for specifying the intended meaning of a term (**15**). They should also avoid including information about the use of a term (**16**). In ontologies, any type of information that goes beyond what is or should be expressed by the axioms may be considered encyclopedic. This kind of information might be useful for understanding a term's intended meaning or its use or is deemed otherwise useful for the target audience but it is not defining information and so should not be included in a definition. Often, an indication that a piece of information is encyclopedic is that it cannot be expressed as a necessary condition. Such information should be included in a separate annotation property, such as 'editor note' (IAO_0000116) or 'example of usage' (IAO_0000112).

(15) × spiracle: *An orifice of the tracheal system by which gases enter and leave the body.* ***Ants have 9 or 10 spiracles on each side of the body. The prothoracic spiracles have been lost, so the first opening occurs on the mesothroax.*** (SIBO_0000502)

✓ spiracle: *An orifice of the tracheal system by which gases enter and leave the body.* **Editor note:** Ants have 9 or 10 spiracles on each side of the body. The prothoracic spiracles have been lost, so the first opening occurs on the mesothroax.

(16) × complex: *Set of interacting molecules that can be copurified.* ***This term and its children should be used only at PARTICIPANT level.*** (MI_0314)

✓ complex: *Set of interacting molecules that can be copurified.*
Editor note: This term and its children should be used only at PARTICIPANT level.

5.2. Guideline: AVOID NEGATIVE TERMS

A definition that tells us what something *is not* is uninformative; it should tell us what something *is*. Therefore, a definition should not include negative features, which should be avoided by stating positive characteristics of the defined entity (**17**). However, there are at least two

exceptions to this principle: (i) when the defined entity is inherently negative (**18**) and (ii) when the negative feature is a distinguishing feature between two sibling classes (**19**).

(17) × mineral: *composed of matter* ***other than plant or animal*** (WordNet 3.1)

✓ mineral: *Composed of solid inorganic matter.*

(18) ✓ non-parametric test: *A statistical hypothesis test* ***that is not based on any parameterized family of probability distributions***. (OBCS_0000236)

✓ bachelor: *A man who* ***is not married***.

(19) ✓ paracentric inversion: *A chromosomal inversion* ***that does not include the centromere***. (SO_1000047)

pericentric inversion: *A chromosomal inversion* ***that includes the centromere***. (SO_1000046)

5.3. Guideline: AVOID DEFINITIONS BY EXTENSION

Definitions by extension (**20**) list instances, i.e., members of the definition's extension, and do not tell us what the thing being defined is, unless we are already familiar with the listed instances. To be maximally informative, use definitions by intension that specify characteristics of the defined thing, as in the corrected version of (**20**).

(20) × planet: *Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune.*

✓ planet: *A celestial body in orbit around the Sun that has a nearly round shape and has cleared the neighborhood around its orbit*[8].

[8] Adapted from the 2006 definition of planet by the International Astronomical Union (IAU).

6. Guideline: ADJUST THE SCOPE

Definitions should have the appropriate scope.

6.1. Guideline: A DEFINITION SHOULD BE NEITHER TOO BROAD NOR TOO NARROW

A definition that is too broad includes things that are not in the extension of the term being defined (**21**). A definition that is too narrow fails to include things that are in the extension of the term being defined (**22**).

(21) × bird: *an animal that lays eggs* (KELLEY, 1998)

✓ bird: *warm-blooded egg-laying vertebrate with feathers and with wings that evolved from forelimbs* (adapted from WordNet 3.1)

In example (**21**), the first definition of bird would be too broad since it includes all oviparous animals, i.e., fish, birds, reptiles and insects. By contrast, the above definition has an adequate scope.

(22) × antidote: *a substance that counteracts snakebite* (KELLEY, 1998)

✓ antidote: *a remedy that stops or controls the effects of a poison* (WordNet 3.1)

In example (**22**), the first definition of antidote would be too narrow since it excludes other types of poison form its extension. By contrast, the WordNet 3.1 definition of antidote has an adequate scope.

6.2. Guideline: DEFINE ONLY ONE THING WITH A SINGLE TEXTUAL DEFINITION

A definition should define only one thing. Therefore, a class should have a single textual definition. Two cases that should be avoided are often found: (i) nested definitions, where definitions contain definitions of other terms embedded within them (**23**), and (ii) multiple definitions, where a single definition annotation property contains definitions in addition to the primary definitional sentence fragment (**24**). Such cases might indicate that a new class should be added to the ontology. If the additional definition is

of an entity that falls within the scope of the ontology, we recommend either (i) removing the additional definition from the annotation property and adding it as the definition of a new term, then using that in the first definition, as in (**24**), (ii) if the class already exists in the ontology, use the class in the primary definition and delete the additional definition (**23**). If the additional definition is of an entity that is out of the scope of the ontology, we recommend either (i) if the term exists in another ontology, using MIREOT (COURTOT et al., 2011) to import the term and then using it, (ii) adding it into a separate editor note, (iii) deleting the additional definition, or (iv) replacing the term deemed to be unknown to the reader with its definition, as in the second solution in (**23**).

Whenever the term is already in the ontology, a cross-reference is preferable since it preserves the link to the full definition of the potentially problematic entity introduced in the definition, as well as to the term that denotes it, both of which are lost when the term is replaced by its definition. Replacing the term by its definition may cause issues if the corresponding class is deprecated and the definition must be updated. It is more difficult to find an embedded definition than it is to identify a term in a definition.

(23) The WordNet 3.1 definition of *cytolytic*, which contains a nested definition of *cytolysis*, can be edited as follows:

× cytolytic: *of or relating to* ***cytolysis****,* ***the dissolution or destruction of a cell*** (WordNet 3.1)

✓ cytolytic: *of or relating to* ***cytolysis***

✓ cytolytic: *of or relating to* ***the dissolution or destruction of a cell***

(24) × cell measurement: *Any quantification of a morphological or physiological parameter of one or more cells.* ***A cell is a membrane-enclosed protoplasmic mass constituting the smallest structural unit of an organism that is capable of independent functioning.*** (CMO_0000227)

✓ cell measurement: *Any quantification of a morphological or physiological parameter of one or more* ***cells****.*

cell: ***A membrane-enclosed protoplasmic mass constituting the smallest structural unit of an organism that is capable of independent functioning.***

7. Guideline: AVOID CIRCULARITY

For a definition to be informative, it should avoid circularity. Circularity can manifest itself in two ways: within the same definition or within the system of definitions.

Circularity within the same definition occurs when a class is defined in terms of itself using one of the labels attached to the class (**25**) or a synonym thereof (**26**), or some grammatically derived form, as in (**27**) when *fearful* is not separately defined independently of *fear*.

(25) × training objective: *A* ***training objective*** *which is fulfilled by the provision of some training* (OBI_0000962)

✓ ctraining objective: *An* ***objective*** *which is fulfilled by the provision of some training.*

(26) × large: *The attribute of something that is* ***big****.* (KELLEY, 1998)

✓ large: *The attribute something* ***has when it has a measurable quality that is above average****.* (adapted from the definition of *large* in WordNet 3.1)

(27) × fear: *The state of being* ***fearful****.* (LANDAU, 2001)

✓ fear: ***an emotion experienced in anticipation of some specific pain or danger*** (WordNet 3.1)

Circularity within the system of definitions occurs when terms are defined in terms of each other, forming a circular pair (**28**) or a circular chain of definitions.

(28) × **training objective**: *A training objective which is fulfilled by the provision of some* ***training*** (OBI_0000962)

training process: *A process that achieves a* ***training objective*** (OBI_0000962)

In example (**28**), the pair of circular definitions tell us that a training objective is fulfilled by some training (process) and that a training process achieves a training objective. It does not tell us what either of these things really are.

8. Guideline: INCLUDE JOINTLY SATISFIABLE FEATURES

The stated conditions must be jointly satisfiable (ARP; SMITH; SPEAR, 2015), that is

- the entity defined must have instances (**29**),
- there should be no logical contradictions involved (**29**). (ARP; SMITH; SPEAR, 2015)

(29) × round square: ***A geometric figure that is simultaneously round and square shaped.***

9. Guideline: USE APPROPRIATE DEGREE OF GENERALITY

A definition should be general.

9.1. Guideline: AVOID GENERALIZING EXPRESSIONS

Avoid the use of expressions such as "usually" and "generally" since either

- a definition is already a statement of general knowledge about a typical case (e.g., in biology) (**30**); or
- the feature containing the expression is not a defining feature, but rather an encyclopedic piece of information that can be moved to a separate 'editor note' annotation property (**31**).

(30) × increased activity of parathyroid: *increased function of this paired endocrine gland that **normally** produces parathyroid hormone (PTH) that regulates calcium and phosphorous metabolism* (MP_0003432)

✓ increased activity of parathyroid: *Increased function of the paired endocrine gland that produces parathyroid hormone (PTH) that regulates calcium and phosphorous metabolism.*

(31) × Tumor Lysis Syndrome: *A syndrome resulting from cytotoxic therapy, **occurring generally in aggressive, rapidly proliferating lymphoproliferative disorders**. It is characterized by combinations of hyperuricemia, lactic acidosis, hyperkalemia, hyperphosphatemia and hypocalcemia.* (MESH:D015275)

✓ tumor lysis syndrome: *A syndrome resulting from cytotoxic therapy, characterized by combinations of hyperuricemia, lactic acidosis, hyperkalemia, hyperphosphatemia and hypocalcemia.*
Note: Tumor lysis syndrome occurs generally in aggressive, rapidly proliferating lymphoproliferative disorders.

9.2. Guideline: AVOID EXAMPLES AND LISTS

Avoid using examples and expressions that specify or enumerate examples (or counterexamples) of things within the definition, such as "etc.", "for example", and "such as", whether it applies (i) to members of the definition's extension (**32**) or (ii) to various types of things falling under the extension of a distinguishing feature, as in (**33**). In the first case, the example(s) should be included in a separate annotation property, such as 'example' (IAO_0000112). In the second case, when possible, replace the examples in the definition with a more general term, the extension of which includes all the examples, as in (**33**).

(32) × cellular_organism: *An organism of microscopic or submicroscopic size, **especially** **a bacterium or protozoan*** (NCRO_0000483)

✓ cellular organism: *An organism of microscopic or submicroscopic size.* **Examples: a bacterium; a protozoan**

In example (**32**), bacterium and protozoan are examples of cellular organisms introduced by the expression "especially".

(33) × patient questionnaire: *A questionnaire that comprises a set of **questions about a patient, such as height, weight, race, biological sex, clinical history, etc.**, which will be filled by the human subject.* (OBIB_0000020)

✓ patient questionnaire: *A questionnaire that comprises a set of **demographic and medical questions**, which will be filled by the human subject.*

In example (**33**), "etc." marks the presence of examples listed to illustrate the types of questions that are part of a questionnaire (i.e., the extension of a differentia). These examples can be replaced by more general terms.

9.3. Guideline: AVOID INDEXICAL AND DEICTIC TERMS

Avoid indexical and deictic terms, such as 'today', 'here', and 'this' when they refer to (the context of) the author of the definition or the resource itself. Such expressions often indicate the presence of a non-defining feature or a case of use/mention confusion. Most of the times, the definition can be edited and rephrased in a more general way (**34**).

(34) × hypersecretion of corticotropin-releasing hormone: *excessive release of **this** factor, which normally stimulates the pituitary to release adrenocorticotropic hormone, from the hypothalamus* (MP_0001753)

✓ hypersecretion of corticotropin-releasing hormone: *Excessive release of corticotropin-releasing hormone from the hypothalamus.* Editor note: corticotropin-releasing hormone normally stimulates the pituitary to release adrenocorticotropic hormone

9.4. Guideline: AVOID SUBJECTIVE AND EVALUATIVE STATEMENTS

Definitions should avoid any kind of subjective and evaluative language, as in (**35**) where the words 'delicious' and 'beautiful' are removed from the definition without any change in the informative content.

(35) × cranberry bean: *Also called shell bean or shellout, and known as borlotti bean in Italy, the cranberry bean has a large, knobby beige pod splotched with red. The beans inside are cream-colored with red streaks and have a **delicious** nutlike flavor. Cranberry beans must be shelled before cooking. Heat diminishes their **beautiful** red color. They're available fresh in the summer and dried throughout the year.* (FOODON_03411186)

✓ cranberry bean: *A bean that has a large, knobby beige pod splotched with red, that is cream-colored with red streaks, and has a nutlike flavor.*

Synonyms: shell bean; shellout; borlotti bean
Editor note: Cranberry beans must be shelled before cooking. Heat diminishes their beautiful red color. They're available fresh in the summer and dried throughout the year.

10. Guideline: DEFINE ABBREVIATIONS AND ACRONYMS

Abbreviations and acronyms should be defined not explicated (**36**). In most cases, the full form of the term should be added as the preferred term using the 'editor preferred term' (IAO_0000111) annotation property, and the abbreviation or acronym added as a synonym label using 'alternative term' (IAO_0000118). Note that in (**36**) 'laser' can be kept as the preferred term since the acronym has become lexicalized.

(36) × laser: *A **laser (acronym for light amplification by the stimulated emission of radiation)** is a light source that…* (OBI_040064)

✓ laser: *A light source that…*
Alternative term: light amplification by the stimulated emission of radiation

Editor note: "laser" is an acronym

11. Guideline: Match Textual and Logical Definitions

Since textual and logical definitions are meant to specify the intended meaning of an ontology's terms, they should *in principle* include the same type of information, although in some cases the logical definition might have additional axioms that help reasoning but which otherwise might be considered inessential. To ensure consistency in ontology development and use, and to promote cohesion across definitions, it is recommended that the parts of a given textual definition match the parts of the corresponding logical definition as in (**37**). However, the parts need not appear in the same order.

(37) ✓ bacteremia: *An infection that has as part bacteria located in the blood.* (IDO_0000506)

bacteremia
EquivalentTo
infection **and**
(has_part some
(infectious agent and Bacteria and
(located_in some blood)))

11.1. Guideline: Proofreading Definitions

Always check the spelling and grammar using appropriate tools. In example (**38**), the definition of 'anesthesiology residency program', 'specialty' is misspelled and 'administration' is spelled incorrectly in two different ways (see words in bold). This could be easily rectified by running a spellchecker.

(38) ✓ anesthesiology residency program: *A medical residency in the medical **speciality** that focuses on the **administeration** of medication for the temporary general or local suppression of sensory or motor nerve function during some health care encounter or on making decisions regarding the **adminstration** of such medication.* (OOST_00000232)

FINAL REMARKS

The objective of these guidelines was to provide the applied ontology community with the relevant principles and conventions for good definition practices in ontologies. These guidelines were also meant to give basic foundational background knowledge on definitions to help readers understand the rationale behind these principles. Finally, these definition writing guidelines have the added benefit of serving as a verification tool for existing ontology definitions.

ACKNOWLEDGEMENTS

This work was supported in part by the Swiss National Science Foundation (SNSF), by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064, and by the University at Buffalo UL1 TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the SNSF, the National Institutes of Health, or the NCTE. Many thanks also to Amanda Hicks, Mark Jensen, Daniel R. Schlegel, and Patrick Ray for their useful discussions and comments and their contributions with examples.

REFERENCES

ARP, R.; SMITH, B.; SPEAR, A. D. *Building ontologies with basic formal ontology*. Cambridge, MA: MIT Press, 2015.

COURTOT, M. et al. Mireot: the minimum information to reference an external ontology term. *Applied Ontology*, v. 6, n. 1, p. 23-33, 2011.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION - ISO. *Terminology work:* principles and methods (iso 704:2009). Geneva: ISO, 2009.

KELLEY, D. *The art of reasoning.*Third edition. New York, London: W.W. Norton & Company, 1998.

LANDAU, S. I. *Dictionaries:* the art and craft of lexicography. 2nd editon. Cambridge: Cambridge University Press, 2001.

NDI-KIMBI, A. Guidelines for terminological definitions: the adherence to and deviation from existing rules in bs/iso 2382: data processing and information technology vocabulary. *Terminology*, v. 1, n. 2, 1994. p. 327-350.

PAVEL, S.; NOLET, D. *Handbook of terminology*. Canada: Public Works and Government Services - Translations Bureau, 2001.

SCHLEGEL, D. R.; SEPPÄLÄ, S.; ELKIN, P. L. Definition coverage in the obo foundry ontologies: the big picture. In: *Proceedings of the International Workshop on Definitions in Ontologies* (ICBO BioCreative 2016), August 1-4, Corvallis (ORG), USA, 2016.

SEPPÄLÄ, S. Survey on defining practices in ontologies: Report summary. In: *Proceedings of the International Workshop on Definitions in Ontologies* (DO 2013), 2013, Montreal, Canada. 2013.

_____. Definition writing guidelines for cili. *Linguistic Issues in Language Technology (LiLT)*, Special Issue on on Linking, Integrating and Extending Wordnets. Forthcoming.

_____; RUTTENBERG, A. *Survey on defining practices in ontologies:* report, 2013. Available at: <http://definitionsinontologies.weebly.com/survey-report.html>. Accessed on: 30 nov. 2017.

_____.; _____.; SCHREIBER, Y.. Definitions in ontologies. *Cahiers de lexicologie*, n. 109, p. 173-206, 2016.

_____; SMITH, B. The functions of definitions in ontologies. In: *Formal ontology in information systems*: Proceedings of the 9th international conference (FOIS 2016), v. 283. FERRARIO R.; KUHN, W. (ED). Annecy, France: IOS Press, 2016. p. 37-50.

_____; SCHREIBER, Y.; RUTTENBERG, A. Textual and logical definitions in ontologies. In: *Proceedings of The First International Workshop on Drug Interaction Knowledge Management* (DIKR 2014), 1.,*The Second International Workshop on Definitions in Ontologies* (IWOOD 2014), and *The Starting an OBI-based Biobank Ontology Workshop* (OBIB 2014), October 6-7, Houston, TX, USA. 2014.

SMITH, B. Introduction to the logic of definitions. In: *Proceedings of the International Workshop on Definitions in Ontologies* (DO 2013),) in *Proceedings of the 4th International Conference on Biomedical Ontology Workshops* (ICBO 2013), v. 1061, July 7, Montreal, Canada. 2013.

SVENSÉN, B. *Practical lexicography:* principles and methods of dictionary-making. Oxford England; New York: Oxford University Press, 1993.

SWARTZ, N. *Definitions, dictionaries and meanings.* 1997. Available at: <http://www.sfu.ca/~swartz/definitions.htm>. Accessed on: 14 jul. 2017.

VÉZINA, R. et al. *La rédaction de définitions terminologiques*. Montréal: Office québecois de la langue française, 2009.

WIKIPEDIA. *Triangle:* wikipedia, the free encyclopedia, 2017. Available at: <https://en.wikipedia.org/w/index.php?title=Triangle&oldid=785338335>. Accessed on: 31 jul. 2017.

ANNEX

ONTOLOGIES REFERENCED IN THE EXAMPLES

Unless otherwise specified below, all the cited identifiers of terms from these ontologies are CURIEs,[9] with the prefix http://purl.obolibrary.org/obo/.

- CHEBI: Chemical Entities of Biological Interest
- CMO: Clinical Measurement Ontology
- DDANAT: Dictyostelium discoideum anatomy
- EFO: Experimental Factor Ontology
  (prefix http://www.ebi.ac.uk/efo/)
- FOODON: FoodOntology (FoodOn)
- IAO: Information Artifact Ontology
- MESH: Medical Subject Headings (MeSH)
  (prefix MESH=http://purl.bioontology.org/ontology/MESH)
- MI: Molecular Interactions
- MP: Mammalian Phenotype Ontology
- NCIt: National Cancer Institute Thesaurus
  (prefix NCIt=http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#)
- NCRO: Non-Coding RNA Ontology
- OAE: Ontology of Adverse Events
- OBCS: Ontology of Biological and Clinical Statistics
- OBI: Ontology for Biomedical Investigations
- OBIB: Ontology for Biobanking
- OOST: Ontology of Organizational Structures of Trauma centers and Trauma systems
- SIBO: Social Insect Behavior Ontology
- SO: Sequence Types and Features Ontology
- UBERON: Uberon multi-species anatomy ontology

[9] CURIE Syntax 1.0: A syntax for expressing Compact URIs, W3C Working Group Note 16 December 2010 (https://www.w3.org/TR/curie/, accessed July 14, 2017)