Ontology for terminology: A passport to the digital world

Christophe Roche

ABSTRACT

Many IT applications rely on a computational representation of the conceptual system of terminologies, such as content management systems and the semantic web, to name a few. In this context, the notion of ontology in the sense of Knowledge Engineering (KE) represents one of the most promising perspectives for the operationalization of terminologies. Unfortunately, the theories of concept in KE and ISO as defined by ISO 1087 and 704 standards are not compatible. Even if the objective of the ISO standards is for communication between humans and not for IT, their principles deserve to be taken into account for building ontologies since they correspond to the way terminologists work. The question is how to translate these principles into a software environment.

0 Introduction

There are many applications that rely on a computational representation of terminologies, i.e., a computational representation of the linguistic dimension and of the conceptual system. Multilingual semantic search engines, content management systems, translation, and specialised dictionaries (Alcina et al., 2019) are some examples. This approach is not new. Figure 1 is taken from a multilingual content management system applied to technical documents written in different languages in the field of renewable energy (Roche, 2010). The documents are classified on the concepts according to their designations which appear in the texts. This approach enables a concept-driven search. The system returns all the documents, regardless of the language in which they are written, indexed by the concepts extracted from the query. Using conceptual relationships improves the relevance of the results.

A computational representation of the conceptual system is an ontology in the sense of Knowledge Engineering[1] (KE). Ontology is one of the most promising ways of operationalising terminologies leading to the notion of ontoterminology, a terminology whose conceptual system is a formal ontology (Roche, 2007;Papadopoulou & Roche, 2018).

[1] Discipline names are capitalized.

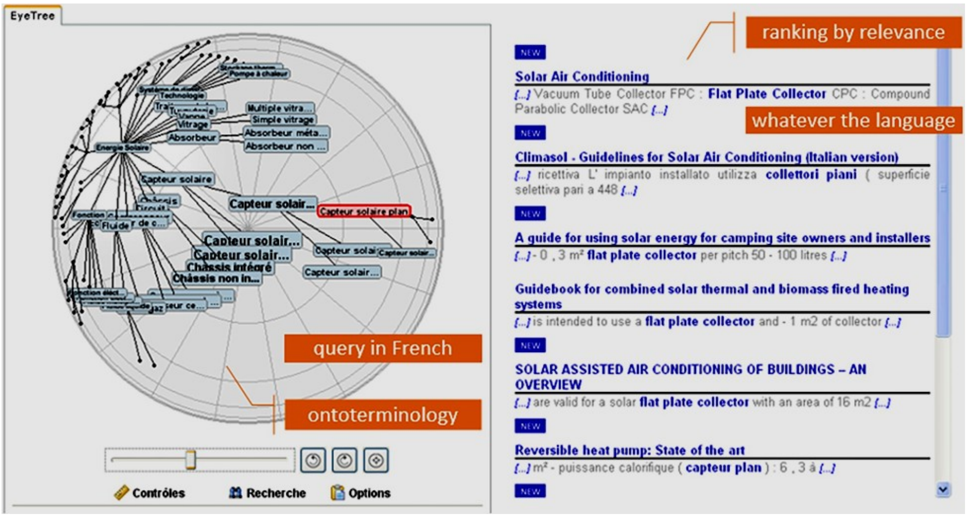

Figure 1. Ontology-oriented Content Management System

More recent applications of the Semantic Web, Linked and Open Data (LOD), and Knowledge Graph (KG), including Internet content management applications such as e-commerce sites or cultural content sites, require a formal definition of concepts. This raises the problem of the compatibility of the theories of concept in KE and Terminology.

The article is structured as follows. The first chapter is devoted to a reminder of the theory of concept in Terminology as defined by the ISO 1087:2019 (International Organization for Standardization [ISO], 2019) and ISO 704:2022 (International Organization for Standardization [ISO], 2022) standards. We will then introduce the theory of concept in ontology in KE, as it is implemented in Protégé, the most widely used ontology building software. The third part will be devoted to the translation of the ISO principles into Protégé. The fourth part is a response to the problems raised by this translation. Indeed, implementing the ISO principles needs some clarification. The TEDI software environment dedicated to ontoterminology is representative of this approach. The article will be illustrated with an example taken from the Krater's terminology of ancient Greek vases.

1 The theory of concept of ISO 1087[2] and ISO 704

The ISO (International Organization for Standardization) Terminology is defined by two standards of the Subcommittee ISO/TC37/SC1 'Principles and methods' of the Technical Committee ISO/TC37 'Language and terminology'. The ISO 1087:2019 standard establishes

[2] The author of the article was Project Leader for the latest version of ISO 1087:2019 standard, and expert for ISO 704:2022. He chaired the Terminology Committee of AFNOR (French Organization for Standardization) from 2014 to 2023.

"*basic terms and definitions for terminology work and terminology science*" when the ISO 704:2022 establishes "*the basic principles and methods for preparing and compiling terminologies both inside and outside the framework of standardization. It describes the links between objects, concepts, definitions and designations*".

ISO 1087 and 704 standards are primarily concerned with human communication, not with IT applications: "*It does not include terms and definitions that are specific to computer applications in terminology work*" (ISO, 2019), "*The goal of terminology work as described in this document is, thus, a clarification and standardization of terminology for communication between humans.*" (ISO, 2022). Although they may, in theory, be applicable to other fields, including information modelling, this is not their primary objective: "*Terminology work can also support knowledge modelling, information modelling, data modelling and classification, but this document does not cover these fields.*" (Ibidem). The ISO theory of the concept as stated by the ISO 1087 and 704 standards cannot be directly operationalized, as some ISO standards rightly point out, which calls for a revision of some notions: "*The preparation of this International Standard brought to light an urgent need to review the family of terminological standards ISO 704, ISO 1087, ISO 17115 and EN 12264 in order to clarify the relations between concept, generic concept, specific concept, object, class, instance, designation and formal representation. This also applies to the forthcoming edition of ISO/TR 24156 (all parts).*" (ISO, 2012).

The ISO standards define a *concept* as a "*unit of knowledge created by a unique combination of characteristics*" (ISO,2019) independently of any language: "*Concepts are not necessarily bound to particular natural languages. They are, however, influenced by the social or cultural background which often leads to different categorizations.*" (Ibidem). A *characteristic* is defined as an "*abstraction of a property*" of an object. The focus is on *essential characteristics* as they are indispensable to understanding concepts. ISO distinguishes two types of concepts according to the number of objects they correspond to. The *individual concept* "*corresponds to a unique object*" (ISO,2019) and the *general concept* "*corresponds to a potentially unlimited number of objects which form a group by reason of shared properties*" (Ibidem).

Concepts are organised into a *concept system* defined as a "*set of concepts structured in one or more related domains according to the concept relations among its concepts*" (Ibidem). All relations are "*relations between concepts*" (Ibidem) divided into *hierarchical relations*, i.e., *generic relations* and *partitive relations*, and *associative relations* including *sequential*, *spatial*, *temporal*, and *causal relations*. To sum up:

- The notion of *characteristic* is central to the ISO theory of concept and, more particularly, the notion of *essential characteristic*.
- The notion of *object* has disappeared in favour of that of *individual concept*. The object is only considered through the concept that corresponds to.
- All relationships are relationships between concepts only.

While this approach is easy to understand, it raises epistemological and logical problems, which are obstacles to computer implementation (Roche, 2012). An *object* is only taken into account through an *individual concept* whose extension is a set containing only the object itself. The object 'Moon'[3] will be represented by the individual concept 'Moon' whose extension is the set containing the single object 'Moon', i.e., using the set notation: { 'Moon' }. Apart from the unnecessary complexity this introduces, it raises the problem of the nature of a concept whose extension is reduced to a single object, for example, the concept of ‘natural satellite of Earth'. Is it an individual concept or a general concept? The latest version of the standard sidesteps the problem by adding the words "potentially unlimited" to the definition of the general concept, which is not satisfactory. This refusal of an explicit representation of objects has an impact on relationships. The instantiation relation (exemplification) between an object and the concept to which it belongs does not exist. It is replaced by the generic relation, which raises the question of the essential and distinctive characteristics of individual concepts. Partitive and associative relations, relations between objects, become conceptual relations in the same way as generic relations.

2 Ontology in Knowledge Engineering

An ontology in KE is a “*specification of a conceptualisation*”, i.e., “*a description (like a formal specification of a program) of the concepts and relationships that can exist*” (Gruber et al., 1993). The notion of ontology has given rise to different definitions presented in the article "What Is an Ontology?" (Guarino et al., 2009). Like Terminology, Ontology aims to facilitate communication between agents, including software agents for ontology. It constitutes one of the most promising perspectives for Terminology. For example, it structures terminologies for computer processing in many fields, such as Medicine, Smart City and Digital Humanities (Papadopoulou & Roche, 2018; Piccini, 2015). Nevertheless, an ontology is not a terminology even if some definitions can suggest the opposite: “*an [explicit] ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning (i.e. definitions)*” (Ushold et al., 1996). Indeed, Ontology does

[3] Objects, concepts, properties, characteristics, and relations are indicated by single quotation marks (ISO 1087)

not explicitly take into account the linguistic dimension of Terminology often reduced to labels stuck on concepts.

Figure 2 is an example taken from Gene Ontology, which provides a computational representation of current scientific knowledge about the functions of genes (http://geneontology.org/docs/introduction-to-go).

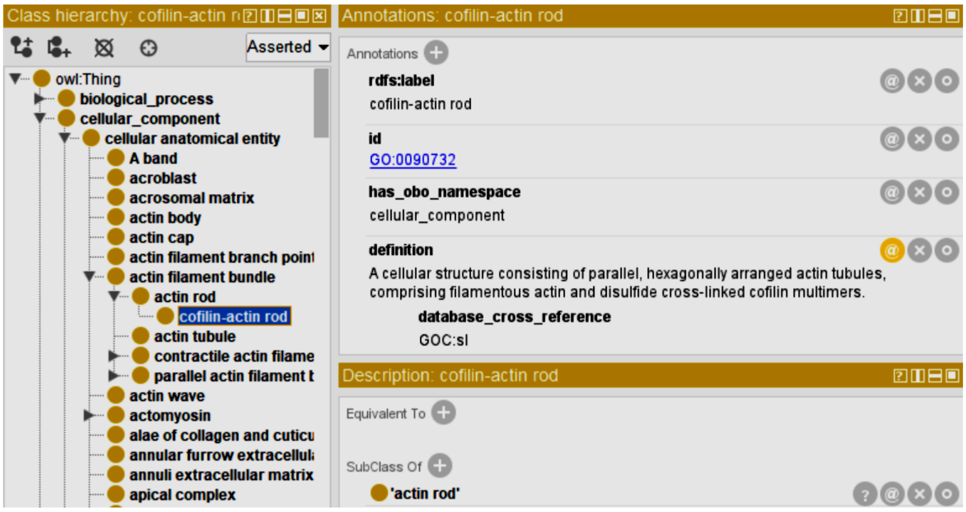

Figure 2. The Gene Ontology uploaded into Protégé

There are different theories of the concept in KE, depending on whether the emphasis is placed on the nature of the objects, their description, structure and composition, their relationships, and more generally their properties. In this article, we focus on the Description Logic (DL) (Baader et al., 2003) that is currently the most widely used for ontology representation. Unlike the ISO theory of concept, objects, also called individuals in the sense they are indivisible, are explicitly represented. Objects are linked by binary relationships, also called properties. The notion of concept is replaced with the notion of class understood as a set of individuals. This approach aims to organise objects into classes according to the relations between objects. The notion of essential characteristic does not exist. It is important to note that a class does not gather objects of the same 'nature', but objects verifying the same property defined in terms of relationships. For example, the class Parisian is the set of individuals, members of the class Person, living in Paris, which will be written in:

$Parisian = \{ x / Person (x) \wedge lives\text{-}In (x,Paris) \}$

Concept and class express a different understanding and organisation of the world. A concept brings together objects of the same nature, for example, the concept of 'Person', whereas a class gathers objects that may be of different nature but which verify the same property. For example, the class of red objects would include my uncle's Ferrari, the apple I ate for lunch and my grandson's ball. These two notions are not mutually exclusive, quite the contrary: they complement each other. The previous example about the notion of "Parisian" will be expressed more "naturally" in the form of a class than a concept.

Relationships between classes (class hierarchy corresponding to set inclusion), between objects and classes (instantiation in the sense of belonging to a set) and between objects (such as partitive and associative relationships) are clearly distinguished.

Protégé (Figure 2), an open-source software developed at Stanford University (https://protege.stanford.edu/), is the most widely used ontology environment based on Description Logic (DL). It offers a wide range of features and benefits from a large community (Musen et al., 2015).

3 Implementation issues of the ISO principles

The ISO theory of concept and the KE theory of class (based on DL) are not compatible. Their theoretical foundations are different. Intensional for the former, extensional for the latter, for which an object is not defined by what it is, but by its relations with other objects. The problem then is to translate the ISO principles into Description Logic. Let us illustrate this process with the example of the Ancient Greek vases of kraters[4]. Kraters are vases for mixing wine and water, whereas amphorae are for storage and transport. Column-kraters, volute-kraters, and calyx-kraters are named after their handles, when bell-krater is named after its body. Column-kraters and volute-kraters are with a neck whereas calyx-krater and bell-krater are without neck. ‘for storage and transport’, ‘for mixing wine and water’, ‘with column-like handles’, ‘with volute-like handles’, ‘with calyx-like handles’, ‘with neck’, and ‘without neck’ are essential and delimiting characteristics. The ‘column-krater’ concept is defined by the unique combination of ‘for mixing wine and water’, ‘with neck’, and ‘with column-like handles’.

The notion of essential characteristic does not exist in Protégé and cannot be directly represented[5]. It has to be translated as property restrictions forcing the expert to change his way of thinking. The object must be split into parts in order to link the whole to its parts with

[4] Beazley archive: http://www.beazley.ox.ac.uk/tools/pottery/shapes/kraters.htm

[5] Essential characteristic corresponds to rigid predicate of higher order logic

the 'has-part'' relationship. The class Colum-krater is the set of vases linked to one neck and to two column-like handles. Translating the absence of part ('without neck') and function ('for mixing wine and water') raises other kinds of problems requiring a background in Logic. Figure 3 illustrates the Krater ontology in Protégé quite far from the ISO approach. The Bell_krater class is defined as the intersection of subclasses.

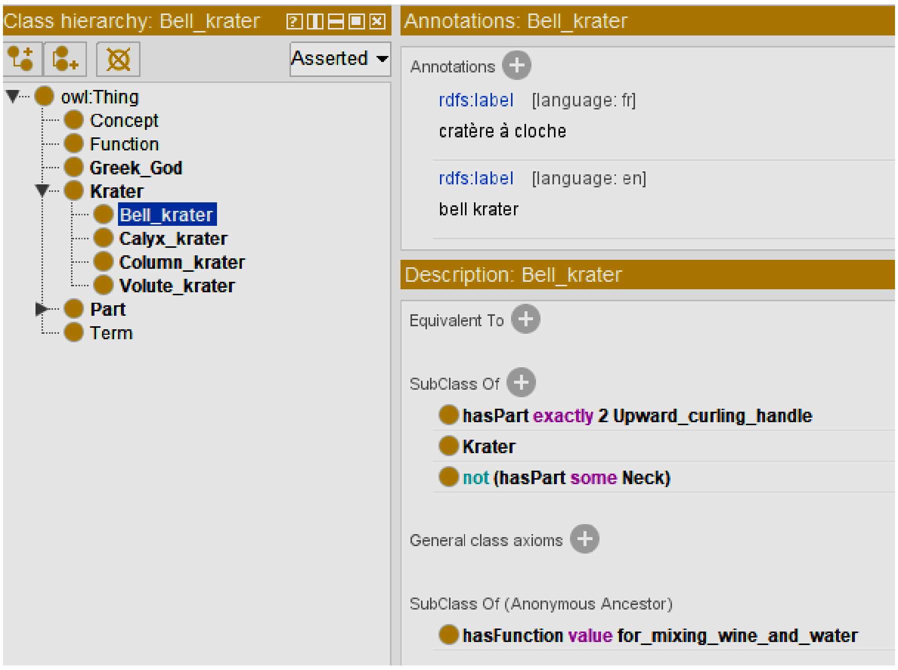

Figure 3. Definition of the Bell-krater class in Protégé

Protégé is a powerful tool for building ontologies. It is dedicated to the organisation of individuals (objects) into classes based on the relationships between objects. The notions of concept and essential characteristic do not exist. They must be translated, which forces terminologists and experts to change their way of thinking (Roche & Papadopoulou, 2019), and adapt to a tool that is difficult to master (Horridge et al., 2013).

4 Ontoterminology

The alternative approach is to directly implement the ISO principles. While essential characteristics and concepts can be explicitly represented, some notions need to be clarified. The result is a system where:

- objects are explicitly represented; the individual concept is no longer necessary. The object 'Moon' is no longer confused with the concept 'natural satellite of the Earth' whose extension is the set reduced to the only object 'Moon'

- descriptive characteristics, such as colour and weight, complete the description of objects
- the relations are clearly distinguished according to the type of entities linked by the relations:
    - the 'instantiation' relation links an object to the concept of which it is an instance (exemplification). The object 'Canada' is an instance of the concept 'country'.
    - the 'generic relation' links a specific concept to a generic concept. The concept 'chair' is a subordinate concept of 'seat'.
    - 'partitive' and 'associative' relations link objects. For example, the 'hasPart' relation links objects of type 'chair' to objects of type 'back' and 'foot'.

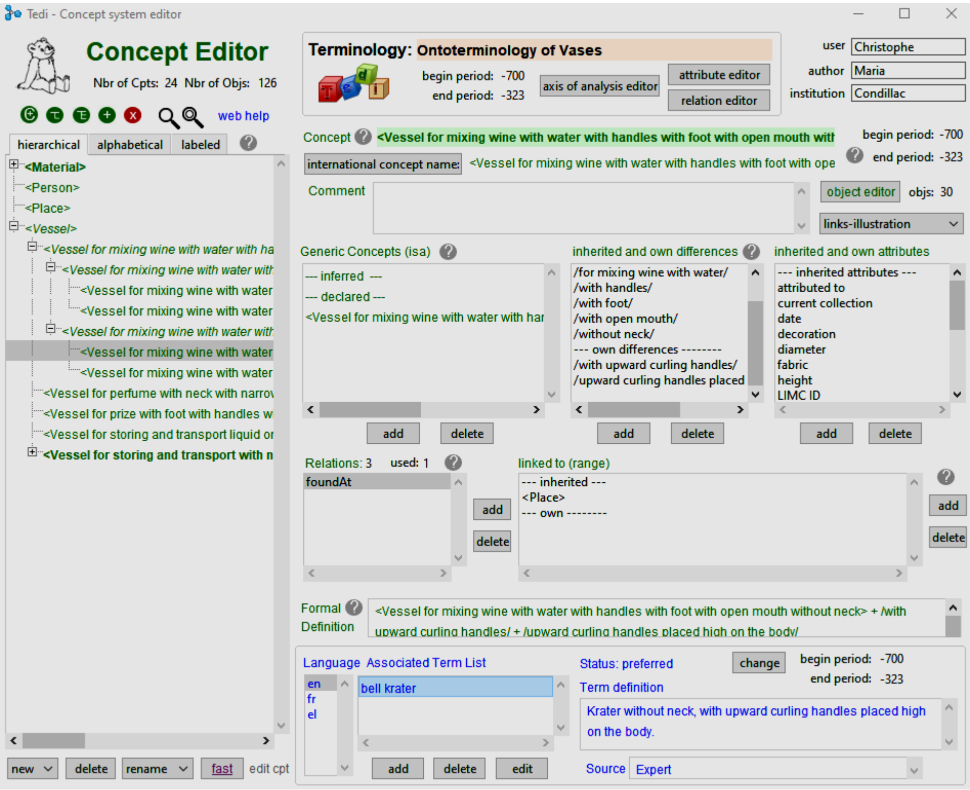

Figure 4. Concept Editor

These principles were implemented in the TEDI software environment (http://ontoterminology.com/tedi). TEDI (for ontoTerminology EDItor) was designed to take into account as much as possible the ISO Terminology principles and the way experts and terminologists think. The two dimensions of terminology, linguistic and conceptual, are explicitly represented leading to the notion of ontoterminology, terminology whose conceptual system is a formal ontology.

Figure 4 illustrates the TEDI Concept editor (essential characteristics are named differences, and descriptive characteristics are called attributes). Concept names and terms are not confused since they belong to different semiotic systems. Concept names are automatically built in such a way that by reading them one understands the nature of the subsumed objects whereas terms are given from texts. Concepts are formally defined in LOK (Language for Ontoterminological Knowledge), a formal language dedicated to defining ontoterminologies.

Figure 5 illustrates the TEDI term editor. Equivalent terms, terminological hypernyms and synonyms are automatically calculated. A template for the definition of terms in natural language is generated from the formal definition of the concepts designated by the terms.

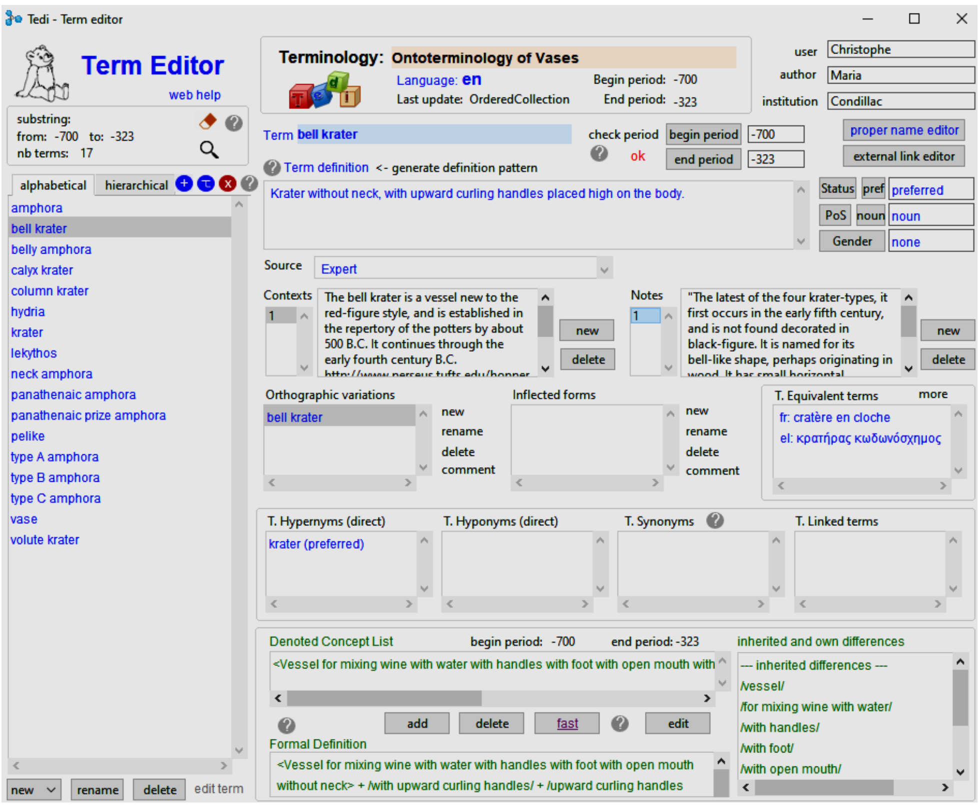

Figure 5. Term Editor

5 Conclusion

Ontology in KE (based on DL as implemented in Protégé) represents one of the most promising perspectives for the operationalization of terminologies. Unfortunately, the KE and ISO theories of concept are not compatible. ISO and KE are two different approaches, intensional for the former and extensional for the latter. Ontology focuses on individuals grouped into classes according to their relationships whereas ISO focuses on concepts defined as combinations of essential characteristics. Translating the ISO principles in a Description Logic is always possible, but rarely in a transparent way. This approach is too distant from the way terminologists think.

The solution is to implement the ISO principles. Unfortunately, some ISO principles are inconsistent from a logical point of view and need to be updated. Objects have to be explicitly represented, replacing the individual concept. Relations are no longer confused, but clearly distinguished according to the types of objects linked by the relations.

Ontology has strongly impacted Terminology in its principles and methods. By clearly separating and expressing the linguistic dimension and the conceptual dimension, terms and concept names are not confused, since they belong to two different dimensions, as well as the definitions of terms in natural language are not confused with the definitions of concepts in formal language.

Bibliography

Alcina A., Costa R., & Roche C. (2019). Terminology and e-dictionaries [Special issue]. *Terminology*, 25(2).

Baader F., Calvanese D., McGuinness D., Nardi D., & Peter Patel-Schneider P. (2003). *The Description Logic Handbook*. Cambridge University Press.

Gruber T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition,* 5(2), 199-220.

Guarino N., Oberle D., & Staab S. (2009). What Is an Ontology?. In *Handbook on Ontologies* (pp. 2-17). Springer.

Horridge, M., Tudorache, T., Vendetti, J., Nyulas, C.I., Musen, M.A., & Noy, N.F. (2013). Simplified OWL Ontology Editing for the Web: Is WebProtégé Enough?. In *The Semantic Web – ISWC 2013. ISWC 2013. Lecture Notes in Computer Science*, vol 8218. Springer. https://doi.org/10.1007/978-3-642-41335-3_13

International Organization for Standardization. (2012). *Health informatics — Categorial structure for terminological systems of surgical procedures* (ISO 1828:2012).

International Organization for Standardization. (2019). *Terminology work and terminology science — Vocabulary* (ISO 1087:2019).

International Organization for Standardization. (2022). *Terminology work — Principles and methods* (ISO 704:2022).

Musen, M.A., & Team Protégé. (2015). The Protégé project: A look back and a look forward. *AI Matters.* Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June. https://doi.org/10.1145/2557001.25757003.

Papadopoulou M., & Roche C. (2018). Ontologization of Terminology. A worked example from the domain of ancient Greek dress. *AIDAinformazioni Journal*, XXXVI(1-2), 89-107.

Piccini, S. (2015). PLOTITERM: modélisation de la terminologie de Plotin en OWL. In *TOTh 2015, Terminology & Ontology: Theories and applications* (pp 313-343).

Roche C. (2007). Le terme et le concept: fondements d’une ontoterminologie. *Terminologie & Ontologie:* In *Théories et applications - TOTh 2007, Annecy* (pp. 1-22).

Roche C. (2010). Ontology-oriented Content Management System: The ASTECH Project. In *Heat-SET 2010. Opatija, CROATIA, October 18-22.*

Roche, C. (2012). Should Terminology Principles be re-examined?. *TKE 2012, 10th Terminology and Knowledge Engineering Conference, Madrid (Spain), 19-22 June* (pp.17-32).

Roche C., & Papadopoulou M. (2019). Mind the Gap: Ontology Authoring for Humanists. In *1st International Workshop for Digital Humanities and their Social Analysis (WODHSA), Graz (Austria), September 23-25*.

Uschold M., & Gruninger M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2), 93-136.

Prof Christophe Roche

ERA Chair Holder, University of Crete (Greece)

Université Savoie Mont Blanc (France)

e-mail: christophe.roche@uoc.gr

Website: http://christophe-roche.fr/