World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:13, No:04, 2019

A Lexicographic Approach to Obstacles Identified in the Ontological Representation of the Tree of Life

Authors : Sandra Young

Abstract : The biodiversity literature is vast and heterogeneous. In today's data age, numbers of data integration and standardisation initiatives aim to facilitate simultaneous access to all the literature across biodiversity domains for research and forecasting purposes. Ontologies are being used increasingly to organise this information, but the rationalisation intrinsic to ontologies can hit obstacles when faced with the intrinsic fluidity and inconsistency found in the domains comprising biodiversity. Essentially the problem is a conceptual one: biological taxonomies are formed on the basis of specific, physical specimens yet nomenclatural rules are used to provide labels to describe these physical objects. These labels are ambiguous representations of the physical specimen. An example of this is with the genus Melpomene, the scientific nomenclatural representation of a genus of ferns, but also for a genus of spiders. The physical specimens for each of these are vastly different, but they have been assigned the same nomenclatural reference. While there is much research into the conceptual stability of the taxonomic concept versus the nomenclature used, to the best of our knowledge as yet no research has looked empirically at the literature to see the conceptual plurality or singularity of the use of these species' names, the linguistic representation of a physical entity. Language itself uses words as symbols to represent real world concepts, whether physical entities or otherwise, and as such lexicography has a well-founded history in the conceptual mapping of words in context for dictionary making. This makes it an ideal candidate to explore this problem. The lexicographic approach uses corpus-based analysis to look at word use in context, with a specific focus on collocated word frequencies (the frequencies of words used in specific grammatical and collocational contexts). It allows for inconsistencies and contradictions in the source data and in fact includes these in the word characterisation so that 100% of the available evidence is counted. Corpus analysis is indeed suggested as one of the ways to identify concepts for ontology building, because of its ability to look empirically at data and show patterns in language usage, which can indicate conceptual ideas which go beyond words themselves. In this sense it could potentially be used to identify if the hierarchical structures present within the empirical body of literature match those which have been identified in ontologies created to represent them. The first stages of this research have revealed a hierarchical structure that becomes apparent in the biodiversity literature when annotating scientific species' names, common names and more general names as classes, which will be the focus of this paper. The next step in the research is focusing on a larger corpus in which specific words can be analysed and then compared with existing ontological structures looking at the same material, to evaluate the methods by means of an alternative perspective. This research aims to provide evidence as to the validity of the current methods in knowledge representation for biological entities, and also shed light on the way that scientific nomenclature is used within the literature.

Keywords : ontology, biodiversity, lexicography, knowledge representation, corpus linguistics
Conference Title : ICKRR 2019 : International Conference on Knowledge Representation and Reasoning
Conference Location : Athens, Greece
Conference Dates : April 08-09, 2019