BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59042
doi: 10.3897/biss.4.59042

OPEN ACCESS

Conference Abstract

The Pensoft Annotator: A new tool for text annotation with ontology terms

Mariya Dimitrova[‡,§], Georgi Zhelezov[‡], Teodor Georgiev[‡], Lyubomir Penev[‡,|]

‡ Pensoft Publishers, Sofia, Bulgaria
§ Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
| Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria

Corresponding author: Mariya Dimitrova (m.dimitrova@pensoft.net)

Received: 28 Sep 2020 | Published: 28 Sep 2020

Citation: Dimitrova M, Zhelezov G, Georgiev T, Penev L (2020) The Pensoft Annotator: A new tool for text annotation with ontology terms. Biodiversity Information Science and Standards 4: e59042. https://doi.org/10.3897/biss.4.59042

Abstract

Introduction

Digitisation of biodiversity knowledge from collections, scholarly literature and various research documents is an ongoing mission of the Biodiversity Information Standards (TDWG) community. Organisations such as the Biodiversity Heritage Library make historical biodiversity literature openly available and develop tools to allow biodiversity data reuse and interoperability. For instance, Plazi transforms free text into machine-readable formats and extracts collection data and feeds it into the Global Biodiversity Information Facility (GBIF) and other aggregators. All of these digitisation workflows require a lot of effort to develop and implement in practice. In essence, what these digitisation activities entail are the mapping of free text to concepts from recognised vocabularies or ontologies in order to make the content understandable to computers.

Aim

We aim to address the problem of mapping free text to ontological terms ("strings to things") with our tool for text-to-ontology mapping: the Pensoft Annotator.

Methods & Implementation

The Annotator is a web application that performs direct text matching to terms from any ontology or vocabulary list given as input to the Annotator. The term 'ontology' is used loosely here and means a collection of terms and their synonyms, where terms are uniquely identified via a Uniform Resource Identifier (URI). The Annotator accepts any of the following ontology formats (e.g. OBO, OWL, RDF/XML, etc.) but does not require the existence of a proper ontology structure (logical statements). We use the ROBOT command line tool to convert any of these formats to JSON. After the upload of a new ontology, the Annotator processes the ontology terms by normalising all exact synonyms and by removing all of the other synonyms (related, narrow and broad synonyms). This is done to limit the number of false positive matches and to preserve the semantic similarity between the matched ontology term and the text.

After matching the words in the input text and the ontology term labels, the Pensoft Annotator returns a table of matched ontology terms including the following fields: the identifier of the ontology term, the ontology term label or the label of the synonym, the starting position of the matched term in the text, the term context (words surrounding the matched term in the text), the type of ontology term (class or property), the ontology from which the matched term originates and the number of times a given term is mentioned in the text. The Pensoft Annotator allows simultaneous annotation with multiple ontologies. To better visualise the exact ontology from which a matching term has been found, the terms are highlighted in different colour depending on the ontology. The Pensoft Annotator is also accessible programmatically via an Application Programming Interface (API), documented at https://annotator.pensoft.net/api.

Discussion & Use Cases

The Pensoft Annotator provides functionalities that will aid the transformation of free text to collections of semantic resources. However, it still requires expert knowledge to use as the ontologies need to be selected carefully. Some false positive matches from the annotation are possible because we do not perform semantic analysis of the texts. False negatives are also possible since there might be different word forms of ontology terms, which are not direct matches to them (e.g. 'wolf' and 'wolves'). For this reason, matched terms can be reviewed and removed from the results within the web interface of the Pensoft Annotator. After removal of terms, they will not be present in the downloaded results.

The Pensoft Annotator can be used to annotate biodiversity and taxonomic literature to help with the extraction of biodiversity knowledge (e.g. species habitat preferences, species interaction data, localities, biogeographic data). The existence of some domain and taxon-specific ontologies, such as the Hymenoptera Anatomy Ontology, provides further opportunities for context-specific annotation. Semantic analysis of unstructured texts could be applied in addition to ontology annotation to improve the accuracy of ontology term matching and to filter out mismatched terms. Annotation of structured or semi-structured text (e.g. tables) can be done with better success. A recent example demonstrates the use of the Annotator to extract biotic interactions from tables (Dimitrova et al. 2020).

The Annotator could also be used for ontology analysis and comparison. Annotation of text can help to discover gaps in ontologies as well as inaccurate synonyms. For instance, a certain word could be recognised as an ontology term match because it is an exact synonym in the ontology but in reality it might be more accurate to mark it as a related synonym. In addition, annotation with multiple ontologies can help to elucidate links between ontologies.

Keywords

text mining, semantics

Presenting author

Mariya Dimitrova

Presented at

TDWG 2020

Funding program

This research has received partial funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840.

References

- Dimitrova M, Poelen J, Zhelezov G, Georgiev T, Penev L (2020) Pensoft – GloBI workflow for FAIR data exchange and indexing of biotic interactions locked within scholarly articles. https://blog.pensoft.net/2020/07/17/pensoft-globi-workflow-for-fair-data-exchange-and-indexing-of-biotic-interactions-locked-within-scholarly-articles/. Accessed on: 2020-8-11.