Collection: SYM09 - Technical standards implications in data liberation and semantic publishing for biodiversity

Print

Biodiversity Information Science and Standards : Conference Abstract

Semantic Publishing Enables Text Mining of Biotic Interactions

Mariya Dimitrova[‡], Jorrit Poelen[§], Georgi Zhelezov[|], Teodor Georgiev[‡], Donat Agosti[¶], Lyubomir Penev[|,‡]

‡ Pensoft Publishers, Sofia, Bulgaria

§ Ronin Institute for Independent Scholarship, Montclair, NJ, USA, Montclair, United States of America

| Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria

¶ Plazi, Bern, Switzerland

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

Citation:

Abstract

Introduction

Scholarly literature is the primary source for biodiversity knowledge based on observations, field work, analysis and taxonomic classification. Publishing such literature in semantic formats (e.g. XML) helps to make this knowledge easily accessible and available both to humans and computers. A recent collaboration between Pensoft Publishers and Global Biotic Interactions (GloBI) (Poelen et al. 2014) demonstrates how semantically published literature can be used to extract species interactions from tables published in the article narratives (Dimitrova et al. 2020) (Fig. 1).

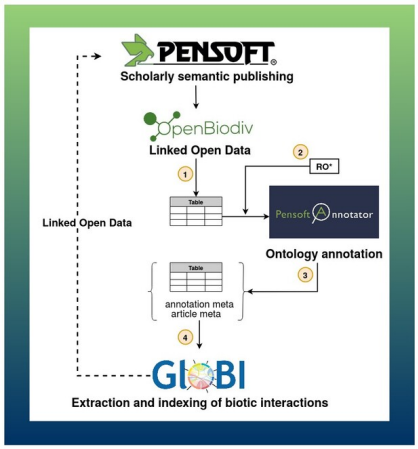

Figure 1.
Collaborative workflow for extraction of biotic interactions from tables in scholarly articles.

Methods

Biotic interactions were extracted from scholarly literature tables published in several Pensoft biodiversity journals. Semantically enhanced publications were processed to extract the tables within them. There were 6993 tables from 21 different journals. Using the Pensoft Annotator [LINK WILL BE ADDED LATER], a text-to-ontology mapping tool, we were able to detect tables which could contain biotic interactions. The Pensoft Annotator was used together with a modified subset of the OBO Relation Ontology (RO), concentrating on the term labeled 'biotically interacts with' and all its children. The contents and captions of all tables were run through the Annotator, which returned the matching ontology terms and their position in the text.

The resulting subset of tables was then processed by GloBI, which parsed the tables to extract the taxonomic names participating in each interaction. The GloBI workflow also generated table citations by SPARQL queries to the OpenBiodiv triple store where all table and article metadata are stored (Penev et al. 2019). OpenBiodiv was also used as a taxon name knowledge base to expand the taxon hierarchy in the tables and to guide the merging of overlapping taxon hierarchy in a single row (e.g. host plant family + host plant species -> host plant species). Taxon name resolution of species interactions was done using the assumption that two non-overlapping taxa are found in a single column. The exact interaction types between the species were not determined, instead the general term labelled "interacts with" was used.

Results

Annotation of biotic interactions via the Pensoft Annotator helped to identify 233 tables possibly containing biotic interactions out of the 6993 tables which were processed. Semantic annotation of taxonomic names within tables allowed GloBI to index the species including their complete taxonomic hierarchies. Currently, GloBI has indexed 2378 interactions, extracted from a subset of 46 of the 233 tables. Interactions extracted via this workflow are available on a special webpage on GloBI's website. Records of the communication behind this collaborative work between GloBI and Pensoft are publically available.

Discussion & Conclusion

One of the limitations of the workflow was the inability to detect the directionality of the interactions. In other words, the tables do not contain information about the subject and object of a given interaction. For instance, in a host-parasite interaction, we can not automatically detect which species is the host and which is the parasite. We plan to address this issue by performing semantic analysis (e.g. part of speech tagging) of the table captions to determine the exact subjects and objects in the interactions. In addition, complicated table structures impeded both the processing of tables by the Pensoft Annotator and their parsing by GloBI's algorithms. We recognise the importance of adopting common formats for sharing interaction data, a practice which would greatly improve the post-publication indexing of tables by GloBI. An example of a standardised table structure is the standard appendix template for primary biodiversity data, introduced by Pensoft (Penev et al. 2020). The template helps authors create semantically enhanced tables, which in turn enables direct harvesting and conversion to interlinked FAIR data. Indexing of biotic interactions by GloBI and Pensoft demonstrates the advantages of storing semantically enhanced data in tables. The adoption of the standard appendix table for primary biodiversity data would improve our ability to extract biotic interactions and to transform scholarly narrative into fully interoperable Linked Open Data.

Keywords

species interactions, semantic publishing, FAIR data, GloBI, text mining

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 Sklodowska-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.
- Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K (2019) OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. Publications 7 (2). https://doi.org/10.3390/publications7020038
- Penev L, Dimitrova M, Kostadinova I, Georgiev T, Agosti D, Poelen J (2020) How to get data from research articles back into the research cycle at no additional costs? https://blog.pensoft.net/2020/04/24/how-to-get-data-from-research-articles-back-into-the-research-cycle-%d0%b0t-no-additional-costs/. Accessed on: 2020-8-11.
- Poelen J, Simons J, Mungall C (2014) Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics https://doi.org/https://doi.org/10.1016/j.ecoinf.2014.08.005