BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35089
doi: 10.3897/biss.3.35089

OPEN ACCESS

Conference Abstract

OpenBiodiv: Linking Type Materials, Institutions, Locations and Taxonomic Names Extracted From Scholarly Literature

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

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

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

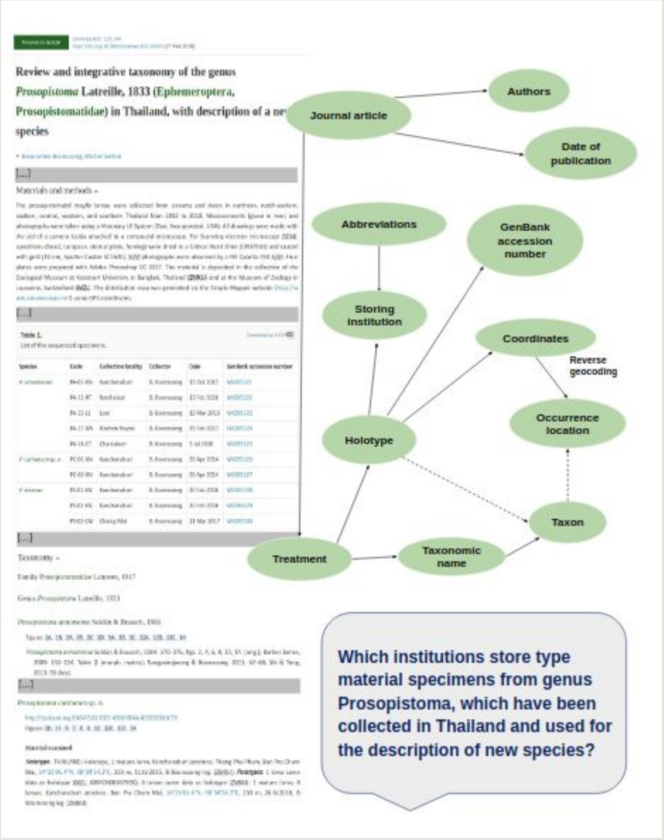
Received: 02 Apr 2019 | Published: 18 Jun 2019

Citation: Dimitrova M, Senderov V, Georgiev T, Zhelezov G, Penev L (2019) OpenBiodiv: Linking Type Materials, Institutions, Locations and Taxonomic Names Extracted From Scholarly Literature. Biodiversity Information Science and Standards 3: e35089. https://doi.org/10.3897/biss.3.35089

Abstract

OpenBiodiv is a knowledge management system containing biodiversity knowledge extracted from scholarly literature: both recently published articles in Pensoft's journals and legacy (taxon treatments extracted by Plazi) (Senderov et al. 2017). OpenBiodiv advances our understanding of the use of scientific names, collection codes and institutions within published literature by using semantic technologies, such as the conversion of XML-encoded text to RDF triples, linked via the OpenBiodiv-O onthology (Senderov et al. 2018). In this poster, we show how OpenBiodiv, currently containing more than 729 million statements, can be used to address a specific use case: finding institutions storing type material specimens of the genus *Prosopistoma* from various literature sources (Fig. 1). This use case is important for various groups of users: institutions, taxonomists, and curators. Answering this complex question is made possible through the application of semantic technologies within OpenBiodiv. Data extraction from taxonomic articles and treatments is enabled the utilisation of common schemas and standards into the extraction process, whereas the conversion of XML-encoded scholarly literature into Resource Description Framework (RDF) is facilitated by OpenBiodiv-O. The

code base for information extraction and data transformation is wrapped in the R packages rdf4r and ropenbio.



Which institutions store type material specimens from genus Prosopistoma, which have been collected in Thailand and used for the description of new species?

Figure 1.

The OpenBiodiv workflow: from information extraction to answering complex biodiversity questions.

The ontology allows to model the structure of research articles and treatments, as well as their corresponding metadata. Thus, OpenBiodiv-O is used to represent not only the sections of treatments but also the various entities within them, for instance geographic coordinates and institution codes within the “Type materials” section of a treatment. Institution codes marked up within articles using the Darwin Core standard (Wieczorek et al. 2012) are mapped to GRBio's institution records. Institutions which are not present in GRBio can often be extracted from the “Abbreviations” section of a given article, thus utilising the power of semantic publishing workflows to discover information hidden within scholarly literature (Penev et al. 2011, Agosti and Egloff 2009). Institutional codes (abbreviations) are then mapped to the narrative section, containing the type materials information. The extraction of coordinates in the taxonomic treatment section allows to establish the location of the collection event through reverse geocoding and enables the

selection of treatments linked to a specific geographic region. Modelling of the “Nomenclature” section within OpenBiodiv-O helps to link taxonomic names, mapped to GBIF’s taxonomic backbone, to their type materials, thus facilitating the discovery of materials corresponding to species from a certain higher-rank taxon.

Keywords

Biodiversity Knowledge Graph, Semantic Technologies, Use Case, Ontology, Information Extraction

Presenting author

Mariya Dimitrova

References

- Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2 (1): 53-53. https://doi.org/10.1186/1756-0500-2-53
- Penev L, Lyal CC, Weitzman A, Morse D, King D, Sautter G, Georgiev TA, Morris R, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. ZooKeys 150: 89-116. https://doi.org/10.3897/zookeys.150.2213
- Senderov V, Georgiev TA, Agosti D, Catapano T, Sautter G, Tuama ÉÓ, Franz N, Simov K, Stoev P, Penev L (2017) OpenBiodiv: an Implementaion of a Semantic System Running on top of the Biodiversity Knowledge Graph. Biodiversity Information Science and Standards 1: e20084-e20084. https://doi.org/10.3897/tdwgproceedings.1.20084
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris R, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (1). https://doi.org/10.1186/s13326-017-0174-5
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7 (1): e29715-e29715. https://doi.org/10.1371/journal.pone.0029715