OpenBiodiv for Users: Applications and Approaches to Explore a Biodiversity Knowledge Graph

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

‡ Pensoft Publishers, Sofia, Bulgaria
§ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria

Corresponding author: Lyubomir Penev (l.penev@pensoft.net)

Abstract

OpenBiodiv is a biodiversity database —knowledge graph based on Resource Description Framework (RDF)—that contains information extracted from the scientific literature. It provides access to an ecosystem of tools and services, including a Linked Open Dataset, an ontology (OpenBiodiv-O) and a website (Dimitrova et al. 2021).

Using the available data, OpenBiodiv discovers links between various biodiversity data types (e.g., taxon names, treatments, specimens, sequences, people and institutions), to answer a user's questions about specific taxa, scientific articles, materials examined and others.

The full-text XML content is converted into Linked Open Data from journals on the ARPHA Publishing Platform and treatments extracted by Plazi's TreatmentBank (stored in the Biodiversity Literature Repository at Zenodo). The database is updated and indexed daily using a workflow based on the Apache Kafka event-streaming platform. The workflow was developed during the European Union-funded *Biodiversity Community Integrated Knowledge Library* (BiCIKL) project (Penev et al. 2022b). By 1 of August 2023, the graph consisted of 24,939 articles; 167,471 treatments; 130,359 authors; 736,809 taxon names; 129,257 sequences; 1,390 institutions and collections, 117,854 figures; 18,585 tables, and 90,008 materials examined sections.

Each semantic statement (e.g., authors, articles, treatments, taxonomic names, localities) has its own globally unique, persistent and resolvable identifier (GUPRI).

There are four ways a user can explore the data on OpenBiodiv:

General search

The search engine is accessible from the OpenBiodiv homepage. The user needs to type in a key term, (e.g., a taxonomic name, authority or an article title), and the system

retrieves information about it. Errors caused by misspellings are avoided due to the Elasticsearch index. It can also determine the semantic type of the searched entity.

Application Programing Interface (API)

OpenBiodiv can be used through a RESTful API for programmatic access. The documentation of the API is described on Swagger. The API construction and functionalities follow the recommendations elaborated by the Technical Research Infrastructures forum of the BiCIKL project (Addink et al. 2023).

User applications based on a query algorithm

This function can be applied for any data class. The method uses the relationships between an element type (e.g., taxon name) and the type of the section, where it can be found.

An application example is *Literature exploration*, designed to answer the question: *Give me information about X mentioned within article section type Y.* The results show the number of mentions of the entity (e.g., taxon name) in the section(s) of interest (e.g., Title, Abstract, Treatment). A click navigates the user to the place in the article that mentions the item (Fig. 1).

SPARQL queries in a thematic context

OpenBiodiv provides a SPARQL endpoint through the Ontotext GraphDB solution*[1]. Several sample SPARQL queries*[2] are also available on the OpenBiodiv website.

Keywords

biodiversity informatics, knowledge graph, SPARQL, RDF

Presenting author

Teodor Georgiev

Presented at

TDWG 2023

Funding program

The BiCIKL project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Addink W, Kyriakopoulou N, Penev L, Fichtmueller D, Norton B, Shorthouse D (2023) Deliverable D1.3 Best practice manual for findability, re-use and accessibility of infrastructures. ARPHA Preprints https://doi.org/10.3897/arphapreprints.e107169
- Agosti D, Benichou L, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e97374
- Dimitrova M, Senderov V, Georgiev T, Zhelezov G, Penev L (2021) Infrastructure and Population of the OpenBiodiv Biodiversity Knowledge Graph. Biodiversity Data Journal 9 https://doi.org/10.3897/bdj.9.e67671
- Penev L, Dimitrova M, Zhelezov G, Georgiev T (2022a) The OpenBiodiv Knowledge Graph Rebuilt: A semantic hub on top of the ARPHA-published content and the Biodiversity Literature Repository. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91357
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022b) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e81136

Endnotes

*1 http://graph.openbiodiv.net/

*2 https://openbiodiv.net/sample-queries

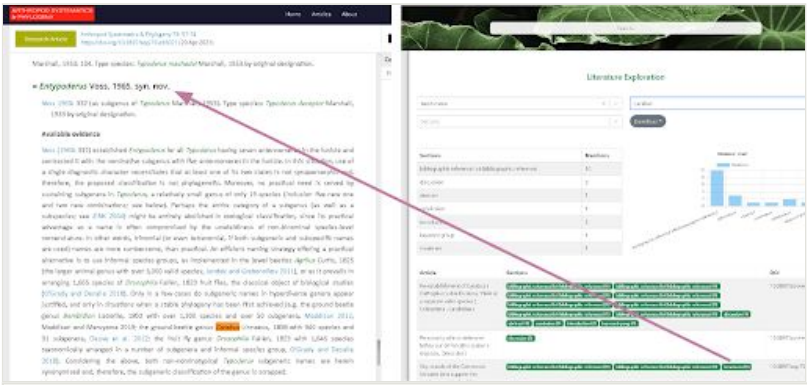

Figure 1.

Back-linking from the OpenBiodiv Literature exploration result page to the respective entity in the original article, provided through the persistent identifiers in the article full-text XML (after Penev et al. (2022a), see also Agosti et al. (2022)).