The OpenBiodiv Knowledge Graph Rebuilt: A semantic hub on top of the ARPHA-published content and the Biodiversity Literature Repository

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

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

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

Abstract

OpenBiodiv is a complex ecosystem of tools and services for RDF conversion of XML narratives of biodiversity articles including Darwin Core data into Linked Open Data (LOD), running on top of a graph database. OpenBiodiv provides four main types of services:

- Searching named entities (e.g., taxon names, taxon concepts, treatments, specimens, occurrences, gene sequences, bibliographic information, institutions, persons) in context, within and between articles.
- Answering questions based on the presence of certain named entities within specific article sections (e.g., titles, abstracts, introduction or other sections, taxon treatments).
- Identifying article sections for further text processing (NLP) and providing contextual information, stored in MongoDB.
- Federating the SPARQL endpoint with other triple stores to enrich the discovered knowledge.

Conversion of such data into RDF follows a general semantic model expressed in the OpenBiodiv-O ontology, an extension of the Treatment Ontology for knowledge representation of current and legacy biodiversity publications (Senderov et al. 2018) and uses two main sources, the full-text article XML published on the ARPHA Publishing Platform and the taxon treatments extracted by Plazi's TreatmentBank from more than 100 biodiversity journals, stored in the Biodiversity Literature Repository at Zenodo. To ensure efficiency, quality control and fast tracking of all stages of the entire process of extraction, conversion to RDF and indexing of the content has been re-built on the Apache Kafka event streaming platform (Fig. 1). In this new format, OpenBiodiv provides not only a GraphDB SPARQL query endpoint but also indexes the named entities through Elasticsearch and additional provision of data to end users through a RESTful API and a number of user applications.

OpenBiodiv is designed for a wide range of users who are interested in a deep-level bibliographic exploration, an ontology-linked search of various data elements (e.g., specimens, sequences, taxon concepts, persons), or co-existence of named entities (e.g., taxon names with a possible biotic relationships between them, or taxon names and potential habitats of occupation) in pre-defined sections of the articles. The SPARQL endpoint allows complex queries of various kinds (Dimitrova et al. 2021).

Keywords

linked open data, RDF, ontology, biodiversity knowledge graph

Presenting author

Lyubomir Penev, Teodor Georgiev

Presented at

TDWG 2022

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

References

- 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
- 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

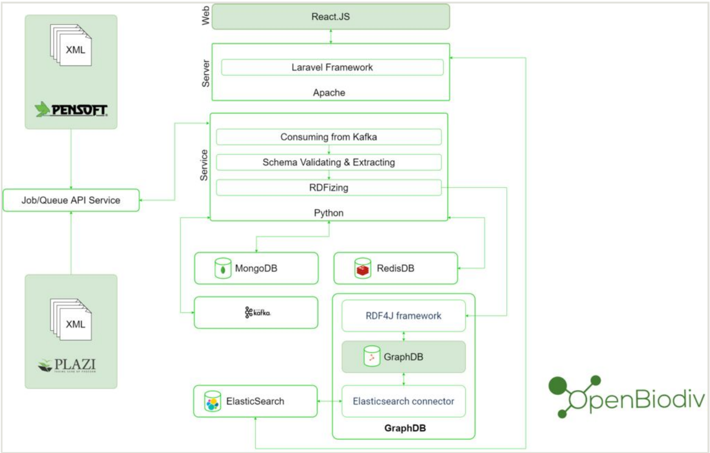

Figure 1.

Data extraction, RDF conversion and indexing workflow of OpenBiodiv.