ARPHA Preprints

PENSOFT

Project Report

Author-formatted document posted on 24/02/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e102612

Deliverable D6.4 Applications for interoperable access to OpenBiodiv through semantically enhanced queries

Lyubomir Penev, **Mariya Dimitrova,** **Georgi Zhelezov,** **Teodor Georgiev**

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Applications for interoperable access to OpenBiodiv through semantically enhanced queries

Deliverable D6.4

30 December 2022

Lyubomir Penev[1,2], Mariya Dimitrova[3], Georgi Zhelezov[1], Teodor Georgiev[1]

[1] *Pensoft Publishers, Sofia, Bulgaria*
[2] *Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria*
[3] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria*

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

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

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Applications for interoperable access to OpenBiodiv through semantically enhanced queries
Deliverable n°:	D6.4
Nature of the deliverable:	Report and Product (Beta)
Dissemination level:	Public
WP responsible:	WP6
Lead beneficiary:	Pensoft Publishers
Citation:	Penev, L., Dimitrova, M., Zhelezov, G. & Georgiev, T. (2022). *Applications for interoperable access to OpenBiodiv through semantically enhanced queries*. Deliverable D6.4. EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 20
Actual submission date:	30 December 2022

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	01 Nov 2022	Lyubomir Penev, Pensoft
1.1	Draft revised	14 Nov 2022	Lyubomir Penev, Mariya Dimitrova, Georgi Zhelezov, Teodor Georgiev, Pensoft
1.2	Images added	20 Nov 2022	Lyubomir Penev. Teodor Georgiev, Georgi Zhelezov, Pensoft
1.3	Second revision	30 Nov 2022	Lyubomir Penev, Mariya Dimitrova, Georgi Zhelezov, Teodor Georgiev, Kristina Hristova, Pensoft
1.4	Review	23 Dec 2022	Patrick Ruch
1.5	Review	23 Dec 2022	Debora Paul
1.6	Review	27 Dec 2022	Quentin Groom
1.7	Final text	28 Dec 2022	Lyubomir Penev, Kristina Hristova, Pensoft
1.8	Submission	30 Dec 2022	Kristina Hristova, Pensoft

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Table of contents

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Preface

To the best of our knowledge, OpenBiodiv is the first production-stage semantic system running on top of a reasonably-sized biodiversity knowledge graph. It stores biodiversity data in a semantic interlinked format and offers facilities for working with it (Senderov et Penev 2016, Senderov et al. 2018, Penev et al. 2019, Dimitrova et al. 2021). It is a dynamic system that continuously updates its database as new biodiversity information becomes available by several international biodiversity publishers. It also allows its users to ask complex queries via SPARQL (a query language for semantic graph databases) and a simplified semantic search interface.

OpenBiodiv was created during two EU-funded Marie Sklodowska-Curie PhD projects: BIG4 (Grant Agreement No 642241) and IGNITE (Grant Agreement No 764840). During those projects, the backend Ontology-0, the first versions of RDF converters and the basic website functionalities have been created (see Dimitrova et al. 2021 for overview).

After the start of the BiCIKL project, the entire workflow for processing and RDF conversion of full-text articles in XML and Plazi's treatments in XML has been re-built using up-to-date technological solutions (such as Apache Kafka and Elasticsearch) to fully automatise and speed up the conversion process and to make it trackable and efficient. As a result, the entire graph content has been re-processed and indexed. New user applications described in Milestone MS27 *App specifications* have been discussed and implemented.

The present deliverable describes the newly built workflow and tools for data extraction, conversion and indexing and the user applications, created in the BiCIKL project.

Summary

OpenBiodiv is a complex ecosystem of tools and services for conversion into Linked Open Data (LOD) of XML narratives of biodiversity articles, including Darwin Core data, through Resource Description Framework (RDF) following the OpenBiodiv-O ontology. OpenBiodiv runs on top of a graph database and provides four main types of services:

- Indexing and searching named entities (e.g., taxon names, taxon concepts, treatments, specimens, occurrences, gene sequences, bibliographic information, institutions, people) 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, such as taxon treatments).
- Identifying article sections for further text processing (NLP) and providing contextual information, stored in MongoDB.
- Providing a SPARQL endpoint and RESTful API for machine-readability of the data and to facilitate federated queries with other triple stores to enrich the discovered

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

knowledge. In the future we are considering adding a SPARQL generator, for example SNORQL.

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 message broker. 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.

List of abbreviations

API	Application Programming Interface
Darwin-SW	An ontology using Darwin Core terms to make it possible to describe biodiversity resources in the Semantic Web
DOI	Digital Object Identifier
DoCO	Document Components Ontology
DwC	Darwin Core standard
FaBiO	FRBR-aligned Bibliographic Ontology
FRBR	Functional Requirements for Bibliographic Records
GUPRI	Globally unique persistent and resolvable identifiers
LOD	Linked Open Data
NOMEN	A nomenclatural ontology for biological names
OBKMS	Open Biodiversity Knowledge Management System
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
XML	Extensible Markup Language

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

1. Data, processes and use

1.1. Overview of the system

What is OpenBiodiv?

OpenBiodiv is a biodiversity database containing knowledge extracted from scientific literature, built as an Open Biodiversity Knowledgement Management System (OBKMS). OpenBiodiv consists of a knowledge graph, a Linked Open Dataset, an ontology (OpenBiodiv-O) and a website (Figure 1). The knowledge graph contains semantic statements about authors, articles, treatments, taxonomic names, examined materials, institutions, genomic sequences, habitats, localities, and more extracted from the literature (Figure 2). Each entity in the Linked Open Dataset has its globally unique, persistent and resolvable identifiers (GUPRI).

Data is modelled according to the OpenBiodiv-O ontology (Senderov et al. 2018) integrating semantic resource types from recognised biodiversity and publishing ontologies with biodiversity-specific resource types not modelled before (Figures 3,4).

The aim of OpenBiodiv is to make biodiversity knowledge easily FAIR (findable, accessible, interoperable and re-usable) both by humans and machines. OpenBiodiv has several user-oriented applications, a RESTful API and a SPARQL endpoint where experienced users can write complex queries.

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

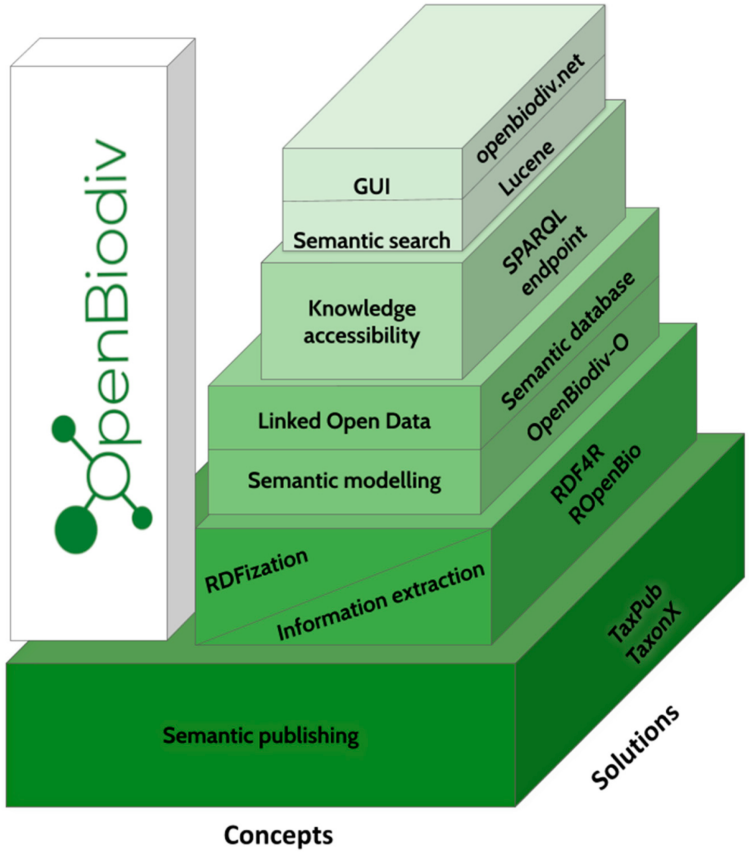

Figure 1: *The main components of the OpenBiodiv Knowledge Graph (after Penev et al. 2019).*

What data is in OpenBiodiv?

OpenBiodiv gathers knowledge extracted from semantically enhanced biodiversity-related articles published in Pensoft's journals (e.g. ZooKeys, PhytoKeys, MycoKeys, Biodiversity Data Journal, etc.) and taxonomic treatments harvested and semantically annotated by Plazi from journals of other publishers (e.g. Zootaxa, European Journal of Taxonomy, etc.) and exposes the links between and within articles (Figure 2).

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

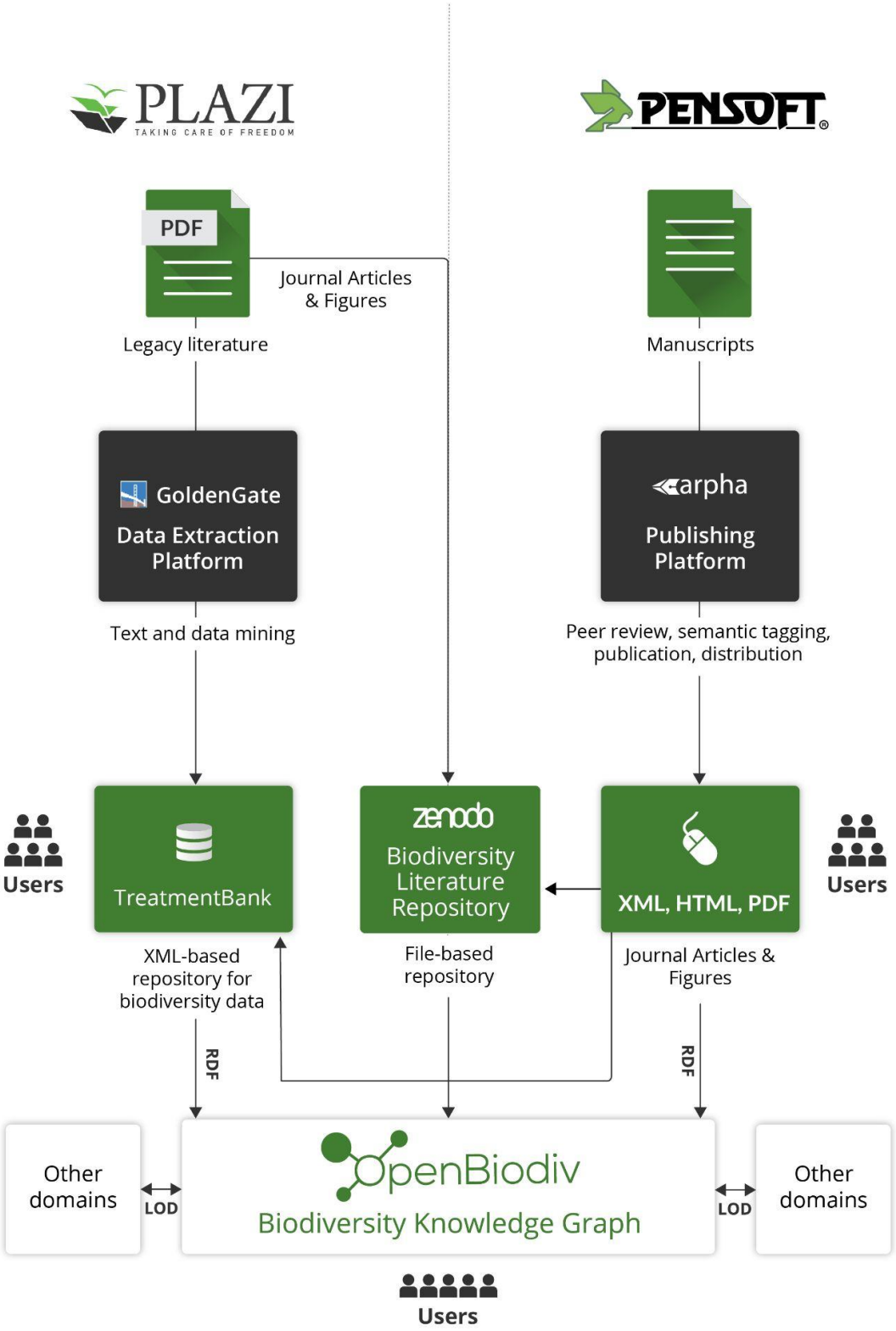

Figure 2: *The main data sources for OpenBiodiv: Pensoft's full-text article XMLs, and Plazi's taxon treatment XMLs (after Penev et al. 2019).*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

For the user feedback, OpenBiodiv has a dedicated email at datascience@pensoft.net in the Contacts page where users can report possible errors in the integrated data. In the future we will explore adding a more direct feedback tool (for example https://chat.openai.com or similar) .

What knowledge can be obtained from OpenBiodiv?

OpenBiodiv offers a broad biodiversity-related querying system to answer open-ended queries based on the data. OpenBiodiv can be used to obtain new knowledge about taxa, scientific articles and their subsections, the examined materials and their metadata, localities, sequences and a lot more. OpenBiodiv can discover hidden links within biodiversity data and can guide research into how data are used in scholarly articles.

The system is able to return information with a relevant visual representation about any one or a combination of its major data classes within a certain scope and semantic context

Data classes are:

- Taxon name (Taxon Name Usage, TNU)
- Taxon treatment
- Specimen
- Sequence
- Person (author)
- Collection/Institution

Examples of data properties are:

- Location
- Date (of publication, sample collection, etc.)
- Geo-coordinates
- Habitat

Article metadata and sections are:

- Title
- Authors
- Abstract
- Keywords
- Bibliographic metadata (DOI, publication date, journal name, article number, pages)
- Introduction
- Material and methods
- Data resources
- Results
- Taxon treatments
    - Nomenclature
    - Material citations (specimen records)

    - Type locality
    - Description
    - Diagnosis
    - Taxonomy
    - Etymology
    - Distribution
    - Molecular data
    - Ecology and biology
    - Conservation
    - Uses
    - Identification keys
- Discussion
- Conclusions
- General (or Undefined) sections
- Figures (including figure legends)
- Tables
- Appendices
- Supplementary files
- Reference lists
- Acknowledgements
- Usage rights
- Funding information
- Author contributions
- Notes

Semantic classes are article sections grouped by topic:

- Taxonomy & Nomenclature
- Diagnoses
- Identification keys
- Conservation
- Biology & Ecology
- Distribution
- Uses (e.g. ethnobotanical information)

Using OpenBiodiv one can answer complex questions like these (see Sample SPARQL queries on the website for more detail):

- Which articles contain treatments which describe specimens in forest or wood habitats?
- Specimens from which taxa have likely burned in the fire in the Museu Nacional de Rio de Janeiro in 2018, according to data from OpenBiodiv?
- Which are the most cited resources and which are the journal articles that cite them?
- Which sequence identifiers have been used in taxonomic literature to describe taxa?
- What are the storing institutions of collected holotypes from order Diptera?

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

- Which treatments describe materials stored in the Natural History Museum, London? Which taxa are described?

OpenBiodiv enables the flow of the data between international repositories for biodiversity data to Pensoft's journals, and then extracts knowledge from Pensoft's journals and Plazi's treatments and stores it in a quad store (Figure 2).

In the following section we describe in more detail the individual components of OpenBiodiv that have developed in the course of the BiCIKL project.

1.2. Backend

Data structure and resources

Prefixes

Prefixes of RDF resources in OpenBiodiv are listed in the following table:

Prefix	Expansion
c4o:	<http://purl.org/spar/c4o/>
datacite:	<http://purl.org/spar/datacite>
dc:	<http://purl.org/dc/elements/1.1/>
dcterms:	<http://purl.org/dc/terms/>
deo:	<http://purl.org/spar/deo/>
doco:	<http://purl.org/spar/doco/>
dwc:	<http://rs.tdwg.org/dwc/terms/>
dwciri:	<http://rs.tdwg.org/dwc/iri/>
fabio:	<http://purl.org/spar/fabio/>
foaf:	<http://xmlns.com/foaf/0.1/>
frbr:	<http://purl.org/spar/frbr/>
nomen:	<http://www.semanticweb.org/dmitriev/ontologies/2013/8/untitled-ontology-6#>
openbiodiv:	<https://openbiodiv.net/>
pkm:	<http://proton.semanticweb.org/protonkm#>

po:	<http://www.essepuntato.it/2008/12/pattern#>
prism:	<http://prismstandard.org/namespaces/basic/2.0>
skos:	<http://www.w3.org/2004/02/skos/core#>
sro:	<http://salt.semanticauthoring.org/ontologies/sro#>

Usage of labels

The objects in OBKMS all have unique identifiers. In addition to those identifiers, the objects have labels that are there primarily for human consumption. Labels can be things like the DOI (in the case of an article), the Latin name of a taxon (in the case of scientific names). This label is encoded with the property rdfs:label.

Article

Every article is represented in RDF using the FaBiO ontology as fabio:JournalArticle (Figure 4).

Example instantiation of an article

```
openbiodiv:e1757e0f-8fd7-543b-a915-17ea2457c102 rdf:type fabio:JournalArticle;
    prism:doi "10.3897/BDJ.10.e69685";
    prism:publicationDate "2022-06-20"^^xsd:date;
    dc:title "A new giant keelback slug of the genus Limax from the Balkans, described by citizen scientists";
    dc:publisher "Pensoft Publishers";
    datacite:hasIdentifier <https://zenodo.org/record/6681252>;
    datacite:hasIdentifier <   https://zoobank.org/f638d49d-7182-49cc-8406-2f5ee3556e38>;
    po:contains <https://openbiodiv.net/f2336f99-5594-589f-b818-be0f84de63cf>.
<https://openbiodiv.net/f2336f99-5594-589f-b818-be0f84de63cf> rdf:type openbiodiv:Treatment.
```

Key here is that the article is linked to different sub-article level elements such as treatments (e.g. https://openbiodiv.net/f2336f99-5594-589f-b818-be0f84de63cf) via po:contains.

Treatment

In OBKMS, we consider Treatment to be a rhetorical element of a taxonomic publication akin to Introduction, Methods, etc. Thus, we derive the RDF type Treatment from <http://purl.org/spar/deo/DiscourseElement> (Figure 4).

Class definition

```
@prefix openbiodiv: <https://openbiodiv.net/> .
@prefix deo: <http://purl.org/spar/deo/> .

openbiodiv:Treatment a owl:Class ;
   rdfs:label "Taxonomic Treatment"@en   rdfs:comment
"A rhetorical element of a taxonomic publication, where taxon circumscription takes place."@en
   rdfs:subClassOf deo:DiscourseElement .
```

Ontologies

OpenBiodiv introduced OpenBiodiv-O, the ontology that serves as the basis of the OpenBiodiv Knowledge Management System. The ontology is available on the OpenBiodiv website at: https://openbiodiv.net/ontologies and is described in detail by Senderov et al. (2018) with some later additions by Dimitrova et al. (2019). OpenBiodiv-O was designed to fill the gaps between ontologies for biodiversity resources, such as DarwinCore-based ontologies, and semantic publishing ontologies, such as the SPAR Ontologies.

OpenBiodiv-O introduces classes, properties, and axioms in the domains of scholarly biodiversity publishing and biological taxonomy and aligns them with several important domain ontologies (FaBiO, DoCO, DwC, Darwin-SW, NOMEN) (Figure 4). By doing so, it bridges the ontological gap across scholarly biodiversity publishing and biological taxonomy and allows for the creation of a Linked Open Dataset (LOD) of biodiversity information (a biodiversity knowledge graph) and enables the creation of the OpenBiodiv Biodiversity Knowledge Management System (OBKMS).

Data collection, conversion and indexing

OpenBiodiv is a complex ecosystem of tools and services for RDF conversion of XML narratives of biodiversity articles and taxonomic treatments into Linked Open Data (LOD). To ensure efficiency, quality control and fast tracking of all stages of the entire process of XML submission, extraction, conversion to RDF and indexing of the content, the entire workflow has been re-built. The R scripts which were used before have been replaced by Python scripts operated using an Apache Kafka platform and a Redis memory cache database (Figure 5). Processing of XMLs follows the publish-subscribe model of event driven architecture to improve performance.

As a source format OpenBiodiv works with XML files that follow the TaxPub schema (Figure 4). These files can be submitted via an API endpoint as separate tasks for processing. The tasks can be grouped in different jobs. The status of each task can be tracked. The API documentation is available at https://api.pensoft.net. Currently the two main sources are 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.

For each task, the XML is chunked into parts based on the article structure and each is sent to different Python scripts that are run as parallel jobs, processing different parts of one article. Once all scripts are finished, the RDF serialization of an article is packaged together and then it is imported into the GraphDB repository. The hierarchical structure of an article is stored in Redis, which allows assembly of the data from each parallel process into a single RDF serialization. After the process is complete, the Redis cache for an article is deleted.

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

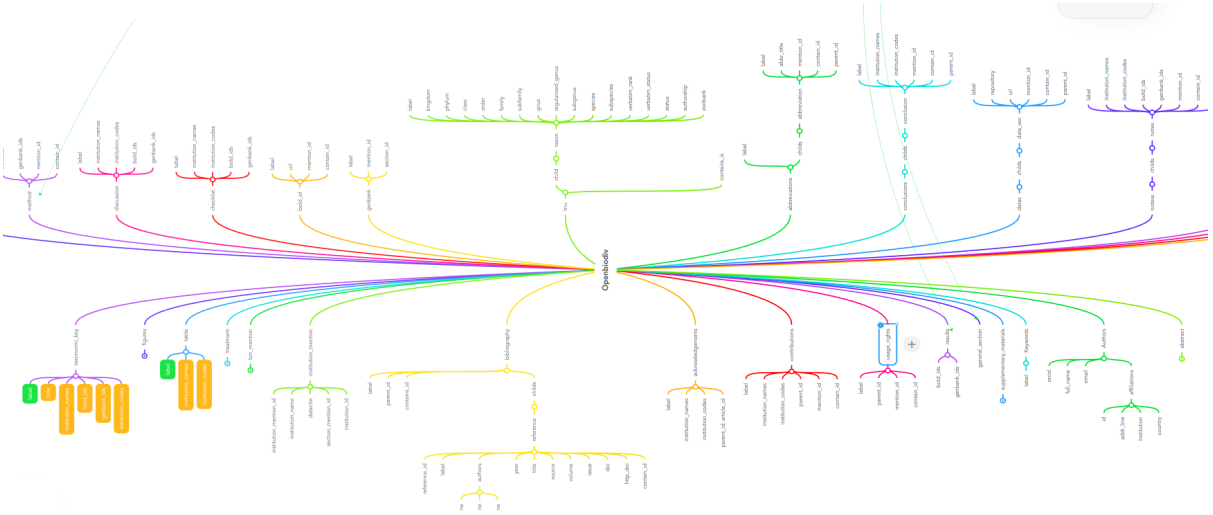

Figure 3: *Hierarchical model of the entities extracted in OpenBiodiv.*

The full text content of each entity (each section) is stored in a MongoDB collection from where it can easily be retrieved. Thus, the GraphDB repository stores predominantly the relationships between entities (e.g. sections) and the MongoDB database stores the literal contents and identifiers for entities.

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) (Figure 4).

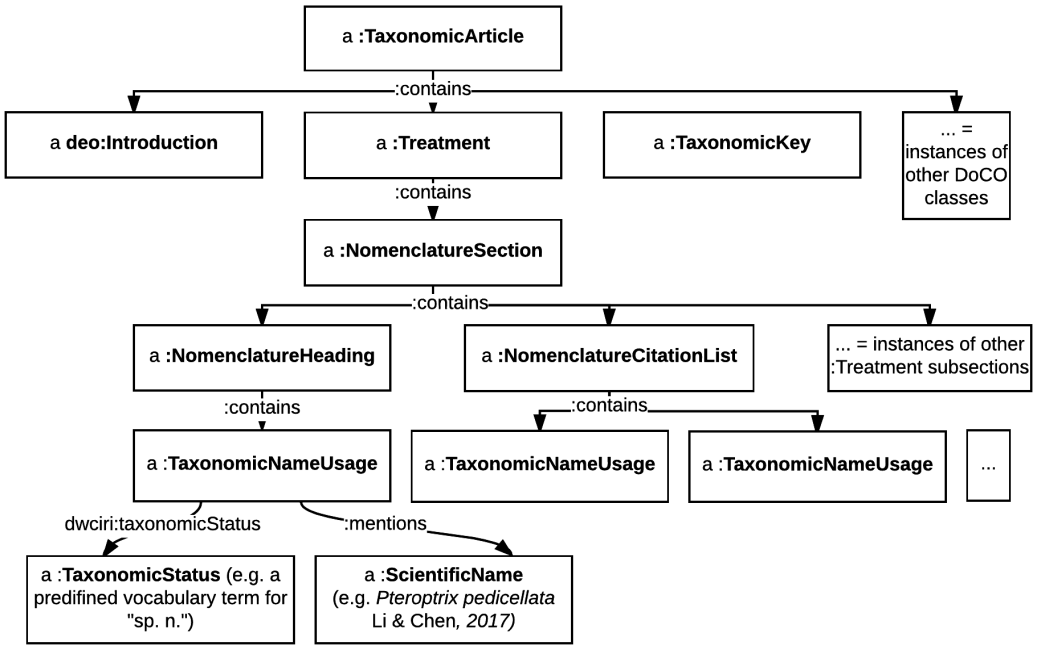

Figure 4: *Modelling of the structure of the taxonomic article for the purpose of OpenBiodiv (after Senderov et al. 2018).*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

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. The documentation of this API is available at https://api.openbiodiv.net/

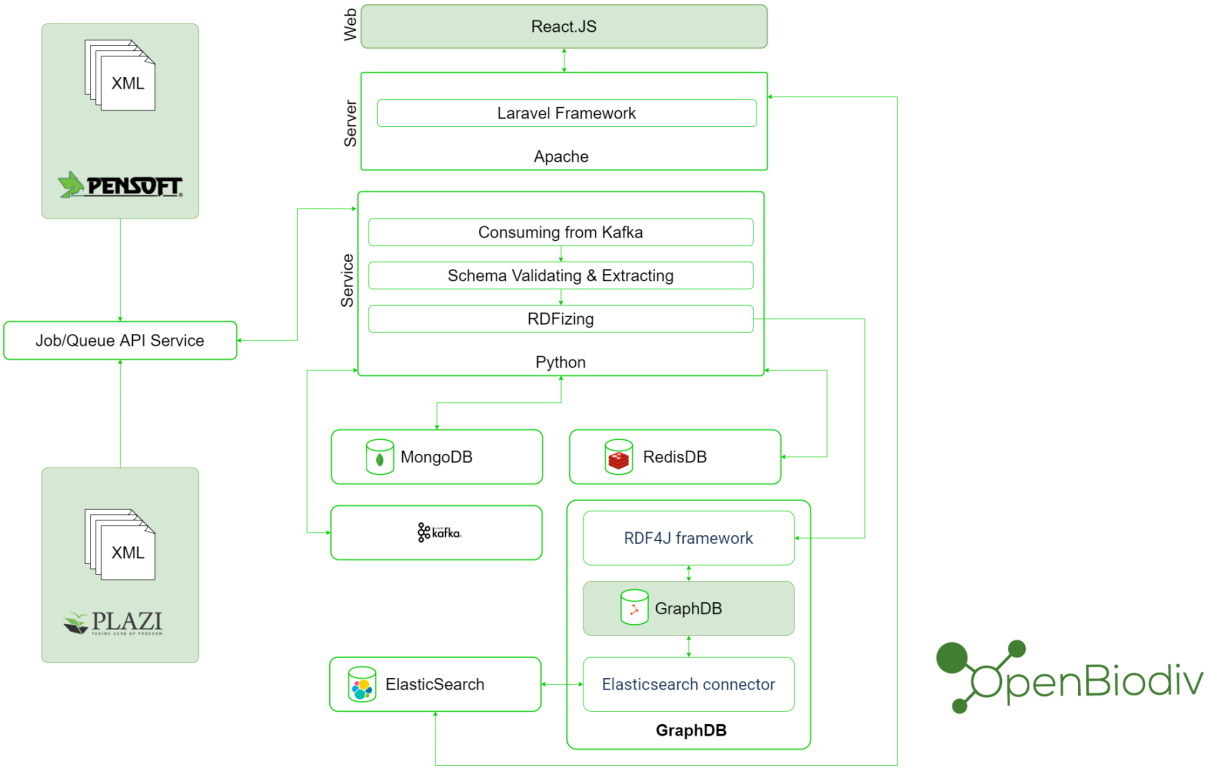

Figure 5: *Data extraction, RDF conversion and indexing workflow of OpenBiodiv (after Penev et al 2022).*

1.3. Frontend

Website

The website of OpenBiodiv is available at: https://openbiodiv.net. It contains an user interface, explanatory pages (About), Terms of service, API and SPARQL endpoints, General search and User applications (Figure 6).

How to find information about biodiversity in OpenBiodiv?

There are four approaches for exploration of data stored in the graph:

- General search
- API
- User applications, based on query algorithms
- SPARQL queries in a thematic context

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

General search

The general search is available on the homepage of OpenBiodiv and allows exploration of the knowledge graph based on key terms like taxonomic names, persons, articles. The user only needs to type the name of an entity of interest belonging to one of the above-mentioned types and the system finds information about it. Misspelling the name is not a problem because the Elasticsearch index supports fuzziness for maximum edit distance allowed for matching. It can also automatically determine the semantic type of the searched entity.

Application Programing Interface (API)

OpenBiodiv can be explored by an unlimited number of various SPARQL queries, however it also provides an API for programmatic access to the data. The documentation of the API is described in Swagger. The API construction and functionalities follow the recommendations elaborated by the Technical Research Infrastructures forum of the BiCIKL project.

User applications based on query algorithms

This section aims at specifying generalised types of queries that can be applied for any data class in OpenBiodiv.

The approach will be based on defining exact relationships between element type (e.g. Taxon name) and section type - the section where this element can be searched to avoid 'No results were found' output.

Figure 6: *The OpenBiodiv homepage.*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Application 1: Literature exploration

Rationale: This application is designed to answer the following general question: Find me information about an entity mentioned within a certain article section in OpenBiodiv. The results will show the number of mentions of this entity (e.g. taxonomic name) in each section of interest (e.g. Titles (X), Abstracts (Y), Treatments (Z), etc.) and aggregated by articles (Figure 7).

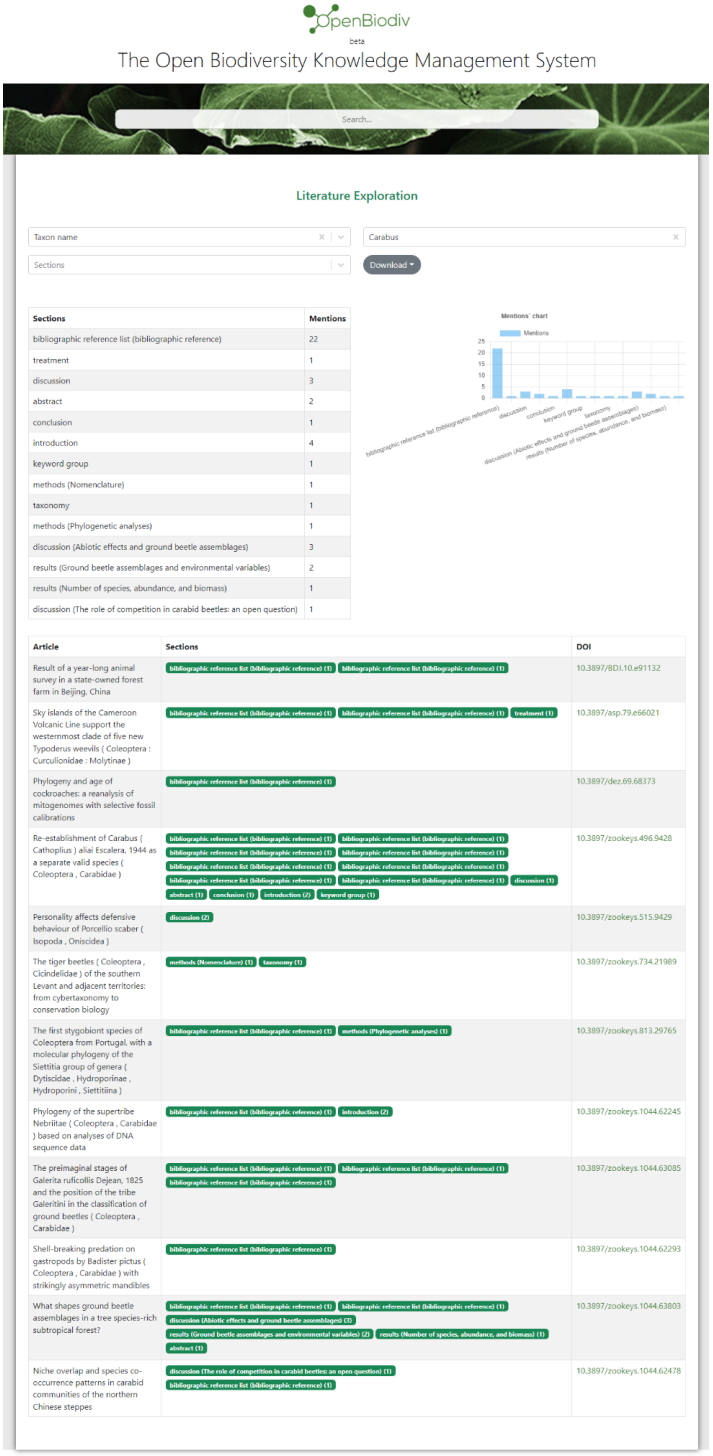

Figure 7: *Application 1: Literature exploration.*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

By clicking on the hyperlinked section label, the user is redirected to the article section where that entity is mentioned (Figure 8).

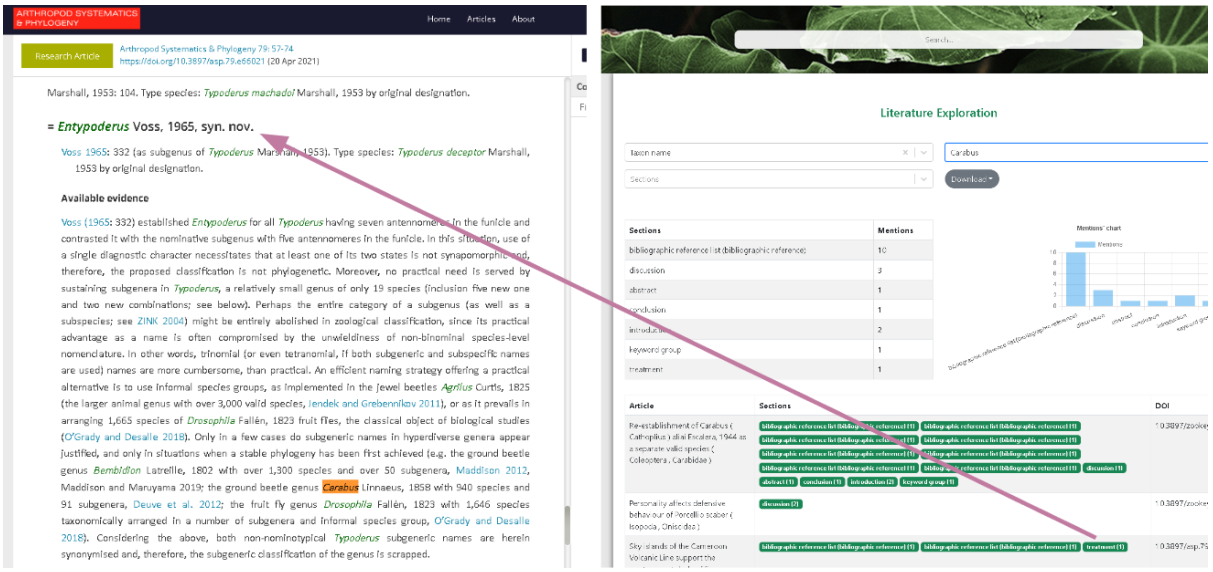

Figure 8: *Back linking from the OpenBiodiv Literature exploration result page to the respective data element in the original article, provided through the persistent identifiers (PIDs) in the article full-text XML (after Penev et al, 2022, see also Agosti et al. 2022).*

Output: A simple graphic representation of the information, for example, about Element X mentioned in Y titles and Z abstracts (plot comparison) illustrates the distributions of the element in the searched sections.

In addition to being visualised in the web page, the results can be exported to a CSV file for further use.

Use: The user will be able to identify articles and their sections where a certain element of interest is mentioned. The data gathered are visualised on screen or can be exported to a CSV file for further use.

Query: Give me all papers where a certain element [Taxon name, Taxon Concept, Specimen, Collection/Institution Code, Sequence, Person) is mentioned in the:

- Title
- Abstract
- Introduction
- Material and methods
- Results
- Discussion
- Conclusions
- *…. please see the full list above in the section "Article metadata and sections"*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Planned extended query: The same query as above but within a certain context: Give me the papers or sections where a certain element [Taxon name, Person, Sequence, Collection/Institution Code, etc.] is mentioned in the literature:

- For a Taxon Name including all child taxa (e.g. all names that belong to a higher taxon, for example all taxa within family Carabidae
- Within a certain time frame (publication date)
- Within a region
    - Response: Titles (X mentions), Abstracts (Y mentions). etc.

  Note: This extension will be developed, once the basic functionality gathers sufficient testing and feedback from the users, so that contexts can be added as options to the above basic query.

Application 2: Co-occurrences

Rationale: This application extends the functionality of the Literature exploration app by adding two or more data elements (named entities), e.g. taxon names, sequences, specimens, specific terms, etc. to be searched together within a given context (Figure 9). For example, some possible questions are:

Query: Simple co-occurrence of two or more terms in a given context

- Give me article sections [Treatment sections, Paragraphs, where Data element 1 and Data element 2 (Taxon name X Taxon name; Taxon name x Sequence, etc.) are mentioned together.
- In the interface you can add as many entities you need (Figure 9), e.g.:
    - Taxon name A
      AND
    - Taxon name B
      AND
    - Sequence C
      AND
    - etc.

The data gathered are visualised on screen or can be exported to a CSV file for further use. By clicking on the hyperlinked section label, the user is redirected to the article section where these entities co-occurred.

Use cases:

- Co-occurrences related to a particular term, e.g. Taxon name
    - Give me all article sections where a Sequence X is mentioned together with Taxon Name Y
    - Give me all treatments of Taxon X where Sequence (Y) are mentioned within the treatment text

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

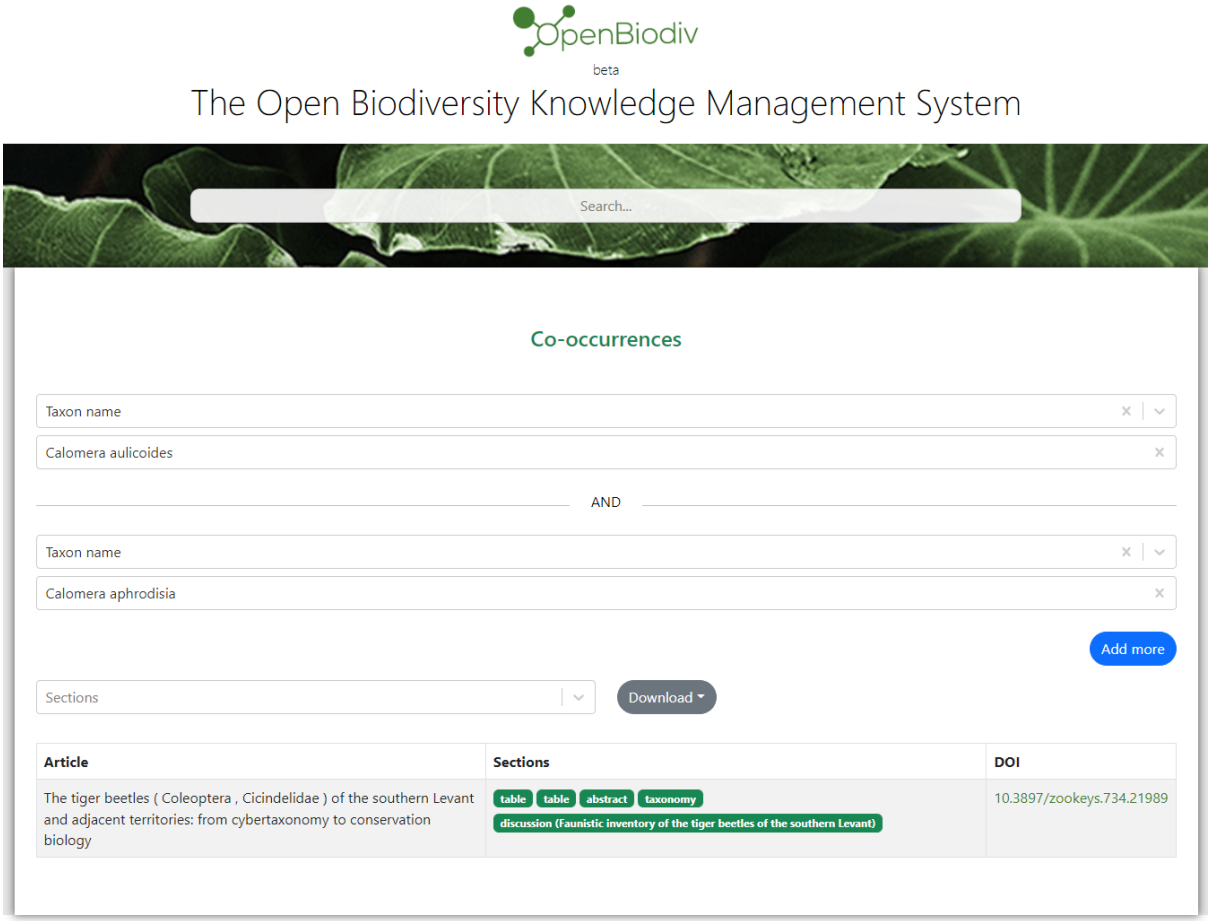

Figure 9: *Application 2: Co-occurrences.*

Application 3: External links

The basic aim of this data discovery application is to search, discover and display data available from trusted external resources, for example specimens, collections, sequences, taxon names, literature, persons and others. The element of interest may be present also in OpenBiodiv.

This service is planned as an additional step to other apps. For example, when one is making a bibliographic exploration about a certain named entity, it could have the option to ask for additional information about that entity available from external resources.

The data records and their identifiers obtained as a result of the search across various resources will be stored as CSV file or RDF using the SKOS ontology.

This application will be developed at a later stage as it requires testing and feedback from the biodiversity community.

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

Application 4: Alerts

OpenBiodiv performs a number of queries at regular intervals to generate reports and send these to the users subscribed to the RSS & E-mail Alert service. The queries can deliver for example:

- All mentions of specimens from a collection or institution based on either citations of a particular collection/Institution code or use of specimen identifiers in the examined materials (material citations).
- All taxon treatments (new taxa, re-descriptions, nomenclatural changes and others) published within a particular taxon.
- All newly published literature that mentions a certain taxon or other named entity of interest (e.g. sequence).

This application will be developed at a later stage as it requires testing and feedback from the biodiversity community.

SPARQL queries in a thematic context

OpenBiodiv provides a SPARQL endpoint through the Ontotext GraphDB enterprise solution at: http://graph.openbiodiv.net/ (Figure 10). The GraphDB software platform has been selected among others (e.g. NEO4J, Virtuoso etc.) because of high performance rates, technical support available “next door” (GraphDB owner Ontotext is based in Sofia, Bulgaria) and for the affordable pricing.

Several Sample SPARQL queries are also available on the website (Figure 11).

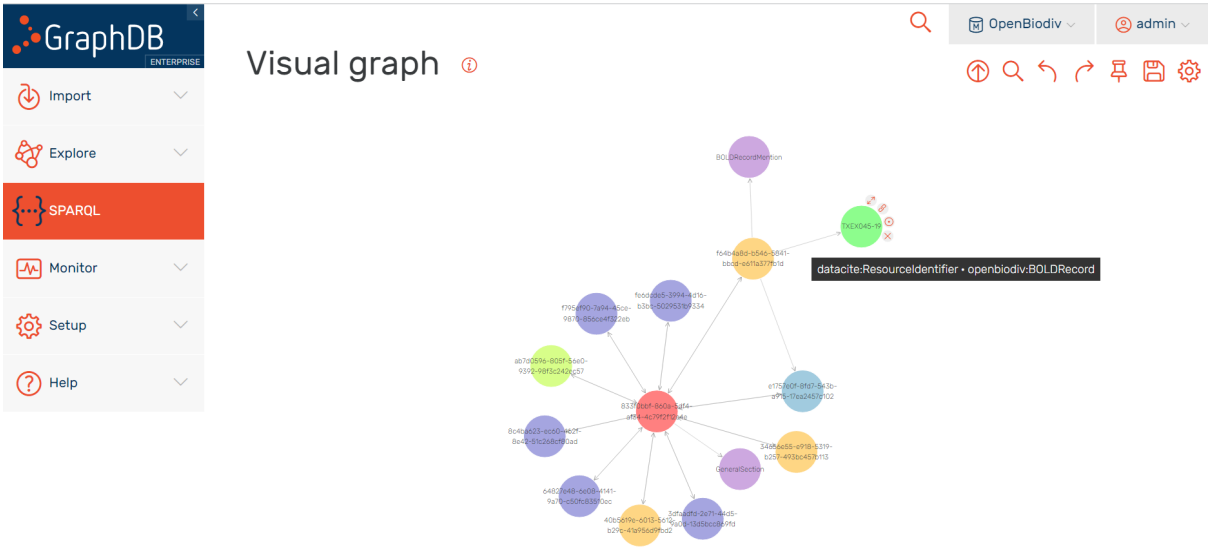

Figure 10: *The GraphDB-based SPARQL endpoint of OpenBiodiv.*

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102612

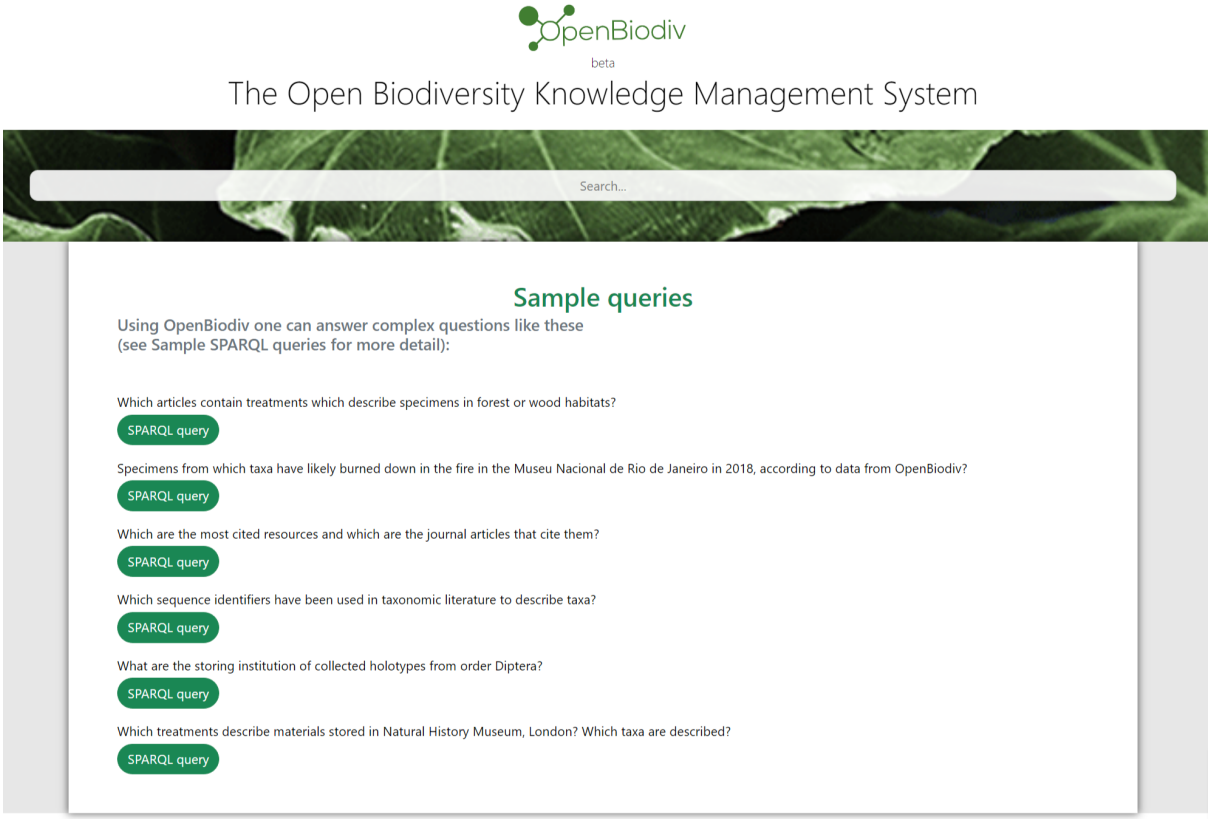

Figure 11: *Sample SPARQL queries at OpenBiodiv.*

At the time of submitting this report (29.12.2022), OpenBiodiv consisted of more than 32,396 processed articles, 47,608 taxon treatments, 1,067 institutions, 367,296 taxon names, 77,296 sequences, 188,181 bibliographic references, 268,862 author names, and 2,105,336 article sections and subsections. It also has a total of 37,555,727 statements (26,899,774 explicit and 10,655,953 inferred).

The below use cases can be searched through OpenBiodiv through predefined SPARQL queries. The list is open and far from complete, hence new queries of this kind can be added. Their purpose is to illustrate the functionalities of the graph and the opportunities it offers for higher-level exploration of the literature. The SPARQL query should also be visible (optionally) to the user for educational purposes and/or possible modification.

Input: The hyperlinked query sentence + the element field which is searched for. In this sense the queries are actually variants of APP1 and can use the same interface

Output: Same as in APP 1 or 2, depending on the case.

Topic 1: Taxonomy, nomenclature, treatments

- Give me all taxon treatments and taxonomic / nomenclatural novelties (new species, related names, replacement names) for a particular taxon including its child taxa (e.g. all taxa within the family Carabidae)
    - *only names that are treatment titles or are related names in the nomenclature section of a treatment and belong to the family Carabidae are searched for*
- For taxon X (or also including its child taxa), give me a timeline graph showing the number of publications where the taxon is mentioned for a specific time period
    - Any mention of this taxon name in their publications (any context)
    - Treatments of that taxon name in their publications and/or TNU in nomenclatural sections of treatments (taxonomic context)
    - Also mark the top 5 most cited works on the graph
- For taxon X (or including its child taxa), give me all authors who have published on it in context:
    - Any mention of this taxon name in their publications (any context)
    - Treatments of that taxon name in their publications and/or TNU in nomenclatural sections of treatments (taxonomic context)
- Search for the first description of a certain taxon name and all consequent treatments of that taxon name
- Is Taxon name X a validly published taxonomic name in a nomenclatural sense (for example, has it a designated holotype in the first description)?
    - Search for the original description if any
    - If found, display the material citations section of the treatment that bear Holotype or Paratype or Lectotype or other type status
- Which original treatments (first publication) are creating homonyms because they are referring to different type material?
- In which treatments a given taxon name has been used for equivalent taxon concepts?
- Which treatments use different names for the same taxon concepts?
- Which treatments are related in that they describe overlapping taxon concepts?
- Which treatments are linked in a nomenclatural sense (including homonyms!) to another treatment?
- Help me find diagnostic characters for a particular taxon (or including its child taxa as well)
    - Identify and display all treatments of that taxon that contain diagnostic sections or ID keys where a taxon name is mentioned
- Explain me etymology of the Taxon name X
    - Identify and display (1) all treatments of that name that contain etymology sections + all etymology sections that contain the name or similarly spelled names (fuzzy match)
- Find me images or videos for a particular taxon (or including its child taxa)
    - Find (1) treatments of a given taxon name that contain images/videos + (2) all figure/video legends that mention that name

- For a given taxon, give me a graph showing the cumulative number of (accepted) species + the number of names through the years - species rate of discovery
- Find all corresponding taxon treatments and articles which mention catalog number XYZT.1234.5678
- Which scientific names have been mentioned as synonyms within a taxon of a certain rank (for example “Porifera”)?
    - Step 1: Find all taxon treatments of all child taxa of certain taxon
    - Step 2: Find and display all names mentioned in the nomenclature sections of those treatments

Topic 2: Specimens and collections

- Give me the papers that have been using materials from a particular natural history collection
    - Step 1: Ask for a collection code
    - Step 2: Give all mentions of a collection code in the OpenBiodiv articles and treatments
- Give mentions of particular specimens from a collection -> search for a collection code within the material citations of all treatments
    - Step 1: Ask for Specimen ID (Catalog number)
    - Step 2: Find mentions of that specimen in the material citation sections of treatments
- Give me all papers that mention a specimen from GBIF (or BOLD, or iDigBio, or DiSSco) published in the OpenBiodiv literature
    - Step 1: Ask for a GBIF or BOLD or iDigBio ID of a specimen
    - Step 2: Find and display material citations or material citation sections where this ID is mentioned
- Which institutions hold the most specimens of taxon X published in the OpenBiodiv literature?
    - Step 1: Find all treatments of a taxon X
    - Step 2: Count and display the numbers of collection codes used in their material citation sections
- What specimens have been published under different names?
    - Step 1: Find all treatments which contain IDs of the same specimen
    - Step 2: Display treatments of different taxon names where the same specimen is present
- Determine the intensity of use of a collection or collections (e.g. to help policy makers make decisions and justify future funding)?
    - Step 1: Count all articles that mention a collection
    - Step 2: Sort out and display collection (codes and full names) by numbers of use
    - OR
    - Step 1: Count all specimen records (material citation sections of treatments) that mention collection codes

    - Step 2: Sort out and display collections (codes and full names) by numbers of use of their specimens
- Determine the intensity of use of a collection over time span (trend)
    - Step 1: Ask for a collection code
    - Step 2: Execute the above queries for that particular collection
    - Step 3: Present the results in a graph over time (years)
- Give me the list and intensity of use of natural history collections located in a certain region (e.g., Europe or in a country)
    - Step 1: Define the region
    - Step 2: Define the collections located in the region
    - Step 3: Find list of collections and display it with numbers of mentions
- Find me all specimens from a particular collection that have been subject of a taxonomic treatment and consequent revision
    - Step 1: Ask for a collection code and a taxon
    - Step 2: Find all treatments for that taxon
    - Step 3: Identify and display treatments that use the same specimens but are published under different taxon names
- What happens if a specimen voucher held in a collection and identified as taxon X, is a subject to taxonomic revision?
    - The taxonomy of species X is revised and it is now in a new genus.
    - The specimen is then a voucher for a different species than the original X but is that reflected in the collection metadata?
    - If this is a type specimen and the new revision in the taxonomy of species X was made according to a new examination of that specimen, then it would be easy to reflect the change
    - If the specimen ID is recorded in the taxonomic database which reflects the taxonomic revision, no taxon name change has to happen (assuming everything is linked) - ideal case scenario

Topic 3: Distribution, biological interactions, biology and ecology

- See Type of Question No 3 above: Co-occurrence of terms, taxa and other named entities; this type of queries will use Pensoft's Annotator tool and access functions to the articles full texts, available via OpenBiodiv

Topic 4: Locations and geographic scope

- What are the taxa living in a particular location?
- What taxons have been mentioned to live in a specific location and what are the articles which mention them?
- For location X, give me all the (GBIF, BOLD or of a particular collection) specimens which have been mentioned in OpenBiodiv

2. Acknowledgements

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

3. References

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: e97374. https://doi.org/10.3897/rio.8.e97374

Dimitrova M, Senderov V, Simov K, Georgiev T, Penev L (2019) OpenBiodiv-O ontology: Bridging the gap between biodiversity data and biodiversity publishing. Biodiversity Information Science and Standards 3: e35773. https://doi.org/10.3897/biss.3.35773

Dimitrova M, Senderov VE, Georgiev T, Zhelezov G, Penev L (2021) Infrastructure and Population of the OpenBiodiv Biodiversity Knowledge Graph. Biodiversity Data Journal 9: e67671. https://doi.org/10.3897/BDJ.9.e67671

Senderov V, Penev L (2016) The Open Biodiversity Knowledge Management System in scholarly publishing. Research Ideas and Outcomes 2: e7757. https://doi.org/10.3897/rio.2.e7757

Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L (2018) OpenBiodiv-O: Ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (5). https://doi.org/10.1186/s13326-017-0174-5

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: 38. https://doi.org/10.3390/publications7020038

Penev L, Dimitrova M, Zhelezov G, Georgiev T (2022) 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: e91357. https://doi.org/10.3897/biss.6.91357