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

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

[1] Pensoft Publishers, Sofia, Bulgaria [2] Bulgarian Academy of Sciences, Sofia, Bulgaria

## Background

**The OpenBiodiv project set to establish a semantic knowledge graph of biodiversity statements extracted from taxonomic articles, published by Pensoft and treatments, extracted by Plazi. Starting with a conceptualisation of the biodiversity publishing domain, several agreed vocabularies like DarwinCore and the SPAR ontologies were combined into a single ontology, OpenBiodiv-O. Semantic enhancement of articles, published in eXtensible Markup Language (XML) enabled their transformation into the machine-readable Resource Description Framework (RDF), which gave rise to the Linked Open Dataset. We also converted GBIF's backbone to RDF and mapped scientific names from taxonomic articles to it. Storing and managing linked statements in the OpenBiodiv knowledge graph allows easy traversal through the statements. This facilitates the answering of complex queries related to biodiversity and biodiversity publishing by institutions, taxonomists, curators, conservation experts and funding organisations.**

## Motivation

**Increasing accessibility of biodiversity knowledge to stimulate scientific research and conservation efforts.**

## Aim

**Developing OpenBiodiv into a knowledge graph capable of answering complex biodiversity questions.**

**Figure 1: The OpenBiodiv architecture**

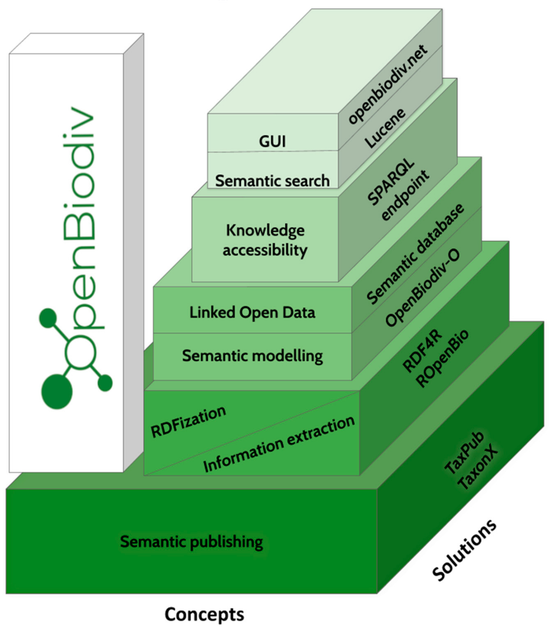
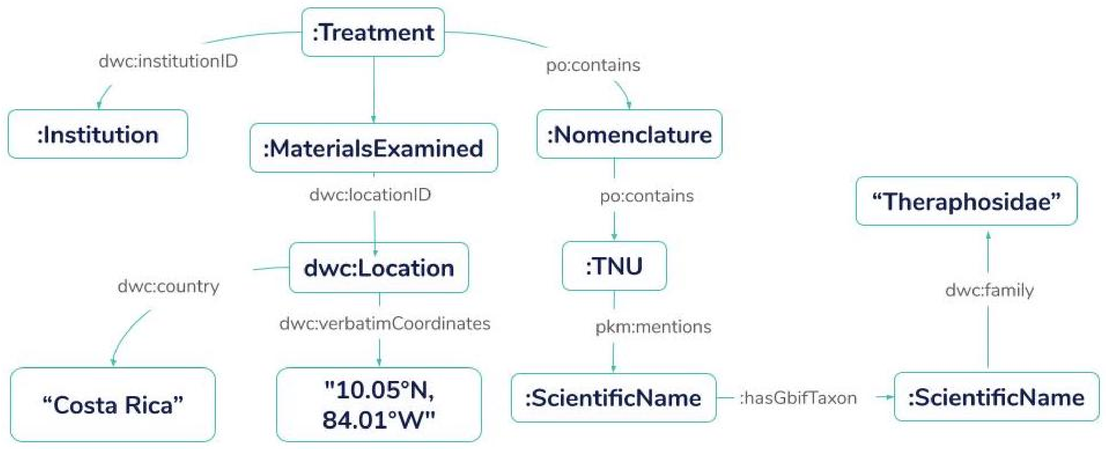



**Figure 2: Semantic relationships between resources in the graph database enable complex queries**




## Example use cases

| Use Case | User Group | SPARQL |
|---|---|---|
| Find papers which describe a taxon, the type material for which is held in the NHM | Researcher<br>Institution<br>Funding body | |
| Find GenBank accession numbers associated with taxa with type material deposited in the NHM | Researcher | |
| Find the storing institution of collected holotypes from family Theraphosidae | Researcher<br>Curator<br>Institution<br>Conservation expert<br>Tarantula owner | |
| Find institutions storing type material specimens of the genus Prosopistoma from various literature sources | Researcher<br>Curator<br>Institution | |

**Table 1: Example use cases of OpenBiodiv and the user groups which might benefit from them.** Answering these questions is done by executing the corresponding SPARQL queries at the OpenBiodiv SPARQL endpoint, available at: http://graph.openbiodiv.net/sparql. Here, you can execute the queries corresponding to the last 2 questions yourself by scanning the QR codes and specifying the username and password to be both "biodiversity_next".

## Challenges

- **Disambiguation of author names, taxon names, institution names and codes**
  - **GRSciColl institution list and Wikidata could provide data to disambiguate institutions**
  - **ORCID identification system could be used to disambiguate authors**
- **Information extraction (e.g. catalog numbers, institution abbreviations) from unstructured text**
  - **Different NLP techniques, such as gazetteer use, supervised machine learning and rule-based solutions, could be of help.**

## Acknowledgements & Funding

We are grateful to Plazi for contributing ideas and are looking forward to many more years of fruitful collaboration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840

## References

1. Senderov, V.; Penev, L. The Open Biodiversity Knowledge Management System in Scholarly Publishing. Res. Ideas Outcomes 2016, 2, e7757.
2. Senderov, V.; Simov, K.; Franz, N.; Stoev, P.; Catapano, T.; Agosti, D.; Sautter, G.; Morris, R.A.; Penev, L. OpenBiodiv-O: Ontology of the OpenBiodiv knowledge management system. J. Biomed. Semant. 2018, 9, 5.
3. TaxonX. Available online: https://sourceforge.net/projects/taxonx/
4. Penev, L.; Catapano, T.; Agosti, D.; Georgiev, T.; Sautter, G.; Stoev, P. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2012.Available online: https://www.ncbi.nlm.nih.gov/books/NBK100351/

SCAN ME