Semantic Search in Legacy Biodiversity Literature: Integrating data from different data infrastructures

Adrian Pachzelt‡, Gerwin Kasperek‡, Andy Lücking§, Giuseppe Abrami§, Christine Driller|

‡ University Library Johann Christian Senckenberg, Goethe University Frankfurt, Frankfurt am Main, Germany
§ Text Technology Lab, Goethe University Frankfurt, Frankfurt am Main, Germany
| Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany

Corresponding author: Christine Driller (christine.driller@senckenberg.de)

Abstract

Nowadays, obtaining information by entering queries into a web search engine is routine behaviour. With its search portal, the Specialised Information Service Biodiversity Research (BIOfid) adapts the exploration of legacy biodiversity literature and data extraction to current standards (Driller et al. 2020). In this presentation, we introduce the BIOfid search portal and its functionalities in a *How-To* short guide. To this end, we adapted a knowledge graph representation of our thematic focus of Central European, primarily German language, biodiversity literature of the 19th and 20th centuries. Now, users can search our text-mined corpus containing to date more than 8.700 full-text articles from 68 journals, and particularly focussing on birds, lepidopterans and vascular plants. The texts are automatically preprocessed by the Natural Language Processing provider *TextImager* (Hemati et al. 2016) and will be linked to various databases such as Wikidata, Wikipedia, the Global Biodiversity Information Facility (GBIF), Encyclopedia of Life (EoL), Geonames, the Integrated Authority File (GND) and WordNet. For data retrieval, users can filter search results and download the article metadata as well as text annotations and database links in JavaScript Object Notation (JSON) format. For example, literature that mentions taxa from certain decades or co-occurrences of species can be searched. Our search engine recognises scientific and vernacular taxon names based on the GBIF Backbone Taxonomy and offers search suggestions to support the user. The semantic network of the BIOfid search portal is also enriched with data from the EoL trait bank, so that trait data can be included in the search queries.

Thus, scientists can enhance their own data sets with the search results and feed them into the relevant biodiversity data repositories to sustainably expand the corresponding knowledge graphs with reliable data. Since BIOfid applies standard ontology terms, all data mobilized from literature can be combined with data on natural history collection objects or data from current research projects in order to generate more comprehensive knowledge. Furthermore, taxonomy, ecology and trait ontologies that have been built or extended

within this project will be made available through appropriate platforms such as The Open Biological and Biomedical Ontology (OBO) Foundry and the Terminology Service of The German Federation for Biological Data (GFBio).

Keywords

search engine, ontologies, text mining, biodiversity data

Presenting author

Adrian Pachzelt

Presented at

TDWG 2021

Grant title

DFG grants SCHN 1016/46-2, ME 2746/5-2, MO 412/54-2

References

- Driller C, Koch M, Abrami G, Hemati W, Lücking A, Pachzelt A, Kasperek G (2020) Fast and Easy Access to Central European Biodiversity Data with BIOfid. Biodiversity Information Science and Standards, 4: e59157 https://doi.org/10.3897/biss.4.59157
- Hemati W, Uslu T, Mehler A (2016) TextImager: a Distributed UIMA-based System for NLP. Proceedings of the COLING 2016 System Demonstrations