Towards "Biodiversity PMC"

Emilie Pasche[‡], Donat Agosti[§], Lyubomir Penev[|,¶], Quentin Groom[#], Alexandre Flament[‡], Julien Gobeill[‡], Patrick Ruch[¤,«]

‡ SIB & HES-SO, Geneva, Switzerland
§ Plazi, Bern, Switzerland
| Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
¶ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria
# Meise Botanic Garden, Meise, Belgium
¤ SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
« HES-SO, HEG Geneva, Geneva, Switzerland

Corresponding author: Patrick Ruch (patrick.ruch@sib.swiss)

Abstract

The Swiss Institute of Bioinformatics (SIB) Literature services (SIBiLS, Gobeill et al. 2020) provides powerful search capabilities to explore the life and health sciences literature by mirroring the United States National Institute of Health's National Library of Medicine (NIH/NLM) (MEDLINE) and National Center for Biotechnology Information (NCBI) PubMed Central® contents.

In the course of the BiCIKL project, SIBiLS started indexing a larger set of biodiversity-related contents in the broad sense including environmental sciences and ecology, to build a new literature database called "Biodiversity PMC". In addition to MEDLINE and PubMed Central, SIBiLS is now providing a unique entry point to half a million taxonomic treatments extracted by Plazi, as well as to a growing set of full-text article XMLs from Pensoft, which were not included into the original PubMed Central. The services can be accessed via a new Graphic User Inteface and an OpenAPI. In addition to usual search operators (using the Apache Lucene syntax), the contents are normalized using a large collection of life sciences terminologies and ontologies. Each instance of a term (or its synonym) is normalized with a unique accession number to support a semantically richer search experience. Of particular interest for the biodiversity communities, SIBiLS contents are normalized using ENVO (Environmental Ontology). Further, taxonomic names are normalized using both the NCBI Taxonomy and the Open Tree of Life, which include names from the Catalogue of Life. The resulting data graph contains 12 billion normalized descriptors and supports access via keyword search, as well as via an original question answering interface, which can help provide new perspectives when navigating the life and health sciences. The data (Journal Publishing Tag Set, JATS, and BioC) are fully available under CC-BY 4.0 licences.

©
.

Keywords

literature services, information retrieval, named entity recognition, question answering interface

Presenting author

Patrick Ruch

Presented at

TDWG 2023

Acknowledgements

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

Hosting institution

SIB Swiss Institute of Bioinformatics & HES-SO

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Gobeill J, Caucheteur D, Michel P, Mottin L, Pasche E, Ruch P (2020) SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. Nucleic Acids Research 48 https://doi.org/10.1093/nar/gkaa328