Enabling Published Taxonomic Data to be used to Address the Biodiversity Crisis: Biodiversity Literature Repository and TreatmentBank

Donat Agosti[‡], Patrick Ruch[§], Jose Benito Gonzalez Lopez[|], Lyubomir Penev[¶,#]

‡ Plazi, Bern, Switzerland
§ HES-SO, Carouge, Switzerland
| CERN, Geneva, Switzerland
¶ Pensoft Publishers, Sofia, Bulgaria
# Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences, Sofia, Bulgaria

Corresponding author: Donat Agosti (agosti@plazi.org)

Abstract

To understand the loss of species, a benchmark is needed, e.g. the status of biodiversity in 1992 when the Convention on Biological Diversity recognized biodiversity crisis to compare to its status in the successive year. Though we are far from knowning how many species there are on planet Earth, we keep track of their descriptions and number through the information kept in our libraries. Each species discovered is represented therein by at least one taxononic treatment. The library includes an estimated 500 million pages and is updated daily with an estimated 17–18,000 new species annually and over 100,000 treatments augmenting the knowledge of existing species.

In reality, we do not know how many species exist. We know that the catalogue of life is incomplete and basic knowlege of known species is often lacking and not even updated regularly. On the other hand, the scientific standard to cite previous works and facts is at the same time an ideal prerequisit to build a knowledge graph in the era of digital knowledge.

TreatmentBank (TB) and the Biodiversity Literature Repository (BLR), two European Research Infrastructures, provide a service to convert unstructured biodiversity data from scholarly publications into semantically enhanced, digital accessible knowledge (Fawcett et al. 2022) making it findable, accessible, interoperable, reusable (FAIR, Wilkinson et al. 2016) and providing long term access in collaboration with Zenodo. Data sources are either legacy publications or active journals using an XML workflow such as Zookeys or the Biodiversity Data Journal, published by the pioneer of taxonomically enhanced Taxpub XML publishing, Pensoft. As part of the Horizon 2020-funded project Biodiversity Community Integrated Knowledge Library (BiCIKL), bidirectional linking of taxonomic names with the Catalogue of Life, DNA-sequences with the International Nucleotide Sequence Database Collaboration (INSDC) and material citations with their respective

natural history collection or their depositions in the Global Biodiversity Information Facility (GBIF) is targeted. As input to allow queries including text and data mining for traits and biotic interactions, the treatments are uploaded to the Swiss Institute of Bioinformatics (SIB) Library Service (SIBiLS) using TaxPub and schema based on Journal Article Tag Suite (JATS) widely used in life science publishing and PubMed Central, allowing the use of the advanced text and data mining tools in the life science community. Another research infrastructure that ensures an RDF representation and conversion into Linked Open Data (LOD) is OpenBiodiv allowing users to explore DAK as part of growing linked open biodiverstiy and related data. All the data fit for GBIF use is reused by GBIF representing close to 60% of all published data sets, which makes TreatmentBank and BLR, for example, a sole provider of data for approximately 90,000 species not recorded in GBIF by other publishers.

The long-term strategy is to build up the momentum to be an infrastructure relevant for global biodiversity conservation and research. Long-term support will be achieved by negotiations with respective agencies and publishers, and through building up a global user community using the biodiversity digital acccessible knowledge.

Keywords

digital objects, open access, text and data mining, policies, strategies, funding, sustainability

Presenting author

Donat Agosti

Presented at

TDWG 2022

Acknowledgements

This work has been supported by the BiCIKL project receiving funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Funding program

Currently TreatmentBank and Biodiversity Literature Repository have been funded as projects by EU research awards, philanthropic foundations such as the Arcadia

Fund, voluntary work, and private commitments. BLR has long term support from Zenodo at CERN.

References

- Fawcett S, Agosti D, Cole S, Wright D (2022) Digital accessible knowledge: Mobilizing legacy data and the future of taxonomic publishing. Bulletin of the Society of Systematic Biologists 1 (1): 1-1. https://doi.org/10.18061/bssb.v1i1.8296
- Wilkinson MD, Kruuk LE, Lanfear R, Binning SA (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (160018). https://doi.org/10.1038/sdata.2016.18