BIOfid, a Platform to Enhance Accessibility of Biodiversity Data

Claus Weiland[1], Christine Driller[1], Markus Koch[1], Marco Schmidt[1], Giuseppe Abrami[2], Sajawel Ahmed[2], Alexander Mehler[2], Adrian Pachzelt[3], Gerwin Kasperek[3], Angela Hausinger[3], Thomas Hörnschemeyer[4]

[1]Senckenberg Biodiversity and Climate Research Centre, Germany; [2]Text-technology Lab, Faculty of Computer Science and Mathematics, Goethe-University Frankfurt, Germany; [3]University Library Johann Christian Senckenberg, Goethe-University Frankfurt, Germany; [4]Senckenberg Research Institute and Natural History Museum Frankfurt, Germany

Corresponding author(s) e-mail: claus.weiland@senckenberg.de; christine.driller@senckenberg.de

ABSTRACT:

With the ongoing loss of global biodiversity, long-term recordings of species distribution patterns are increasingly becoming important to investigate the causes and consequences for their change. Therefore, the digitization of scientific literature, both modern and historical, has been attracting growing attention in recent years. To meet this growing demand the *Specialised Information Service for Biodiversity Research* (BIOfid) was launched in 2017 with the aim of increasing the availability and accessibility of biodiversity information. Closely tied to the research community the interdisciplinary BIOfid team is digitizing data sources of biodiversity related research and provides a modern and professional infrastructure for hosting and sharing them. As a pilot project, German publications on the distribution and ecology of vascular plants, birds, moths and butterflies covering the past 250 years are prioritized. Large parts of the text corpus defined in accordance with the needs of the relevant German research community have already been transferred to a machine-readable format and will be publicly accessible soon. Software tools for text mining, semantic annotation and analysis with respect to the current trends in machine learning are developed to maximize bioscientific data output through user-specific queries that can be created via the BIOfid web portal (https://www.biofid.de/). To boost knowledge discovery, specific ontologies focusing on morphological traits and taxonomy are being prepared and will continuously be extended to keep up with an ever-expanding volume of literature sources. Here we present the key elements of the BIOfid pipeline with emphasis on a practical approach to develop domain-specific ontologies, to deal with the dynamic nature of taxonomies, and to promote interoperability and standardization of terms and definitions. The taxonomic ontologies are designed to fetch any taxon-related literature entry. In this respect, scientific names, synonyms, vernaculars, taxonomic ranks and other classifications peculiar for the respective organism group are compiled. The sources to be used are common open access platforms like *Catalogue of Life* and the *Global Biodiversity Information Facility* (GBIF), complemented by databases more narrowly focusing on certain taxa, such as the *International Ornithological Congress* (IOC) *World Bird List*. We further give insight into our work on the OBO Foundry *Flora Phenotype Ontology* (FLOPO) and the *Lepidoptera Anatomy Ontology* (LepAO). The latter is build on the already existing *Hymenoptera Anatomy Ontology* (HAO) and developed within the framework of an international collaboration that aims at designing a unified *Insect Anatomy Ontology*. BIOfid is co-funded by the Deutsche Forschungsgemeinschaft (DFG) under project number 326061700.

KEYWORDS: Bio-Ontologies, Text Mining, Machine Learning, Specialised Information Service, Biodiversity Knowledge Base