(Re)Discovering Known Biodiversity: Introduction

Donat Agosti ‡

‡ Plazi, Bern, Switzerland

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

Abstract

Biodiversity sciences, including taxonomy, are empirical sciences where all results are published in scholarly publications as part of the research life cycle. This creates a corpus of an estimated 500 million printed pages (Kalfatovic 2010) including billions of facts such as traits, biotic interactions, observations characterizing all the estimated 1.9 million known species (Costello et al. 2013). This library is continually reused, cited and extended, for example with more than an estimated 15,000–20,000 new species annually (Polaszek 2005). All of these figures are estimates because we neither know how many species have been discovered, nor how many are being discovered every day, let alone what we know about them.

Following standard scientific practice, previous publications, specimens, gene sequences, or taxonomic treatments (Catapano 2019) are cited more or less explicitly. In the pre-digital age, these links were meant for the human reader to be understood. For example, "L. 1758" is an established reference and links to both, Carolus Linnaeus and Linnaeus 1758, understandable at least by an expert human, and in the digital age, provides access to the respective digital representation. These data within the hundreds of millions of printed and now increasingly digitally published pages form a seamless, albeit implicit knowledge graph. Unfortunately, most of these publications are in print—the Biodiversity Heritage Library digitized about 50 million pages (Kalfatovic 2010)—or in many cases, closed access publications, and thus this knowledge is not readily accessible in the digital age.

However, in today's digital age, each of these kinds of implicit links is an expensive stumbling block to access and reuse of the referenced data, its parent publications and the cited referenced data therein. Inadequate formats, language and access to taxonomic information were already recognized in 1992 at the Rio Summit (Taxonomic Impediment). The consequences of these impediments are only now obvious with the realization of the daunting amount of human resources needed to digitally catalogue and index this unknown (not discoverable and inaccessible) known knowledge, let alone making the data itself findable, accessible, interoperable and reusable (FAIR). This is a formidable and complex scientific challenge.

Plazi is taking on this challenge. Its vision is to promote and enable the discovery and liberation of data to transform the unknown known data into digitally accessible knowledge, i.e., to build a digital knowledge base aimed at discovering all the species (and other taxa) we know, and what we know about them. Taxonomic publications with their highly standardized taxonomic names, taxonomic treatments, treatment citations, material citations and illustrations are well suited to machine extraction. Together they include the entire catalogue of life with all the discovered species and their synonyms, often tens to hundreds of treatments, and figures that depict the myriad forms that comprise the world's biodiversity. Once these data are FAIR, it allows bidirectional linking, for example of taxonomic names to the referenced taxonomic treatment, other digital resources such as gene sequences or digital specimens. At the same time, each datum is an entry point to the wealth of information that can be followed by the human user by clicking the links, but more importantly, analysed by machines. Here, digitally accessible knowledge will be defined in the context of discovering known biodiversity, including strategies of how to approach the challenge, which then will be detailed in subsequent talks in this symposium.

This symposium is based on Plazi's ongoing data liberation and discovery supported by the European Union (e.g. Biodiversity Community Integrated Knowledge Library BiCIKL), United States (e.g. NIH) and Swiss research funding (e.g. e-BioDiv and the Arcadia Fund), collaboration with publishers (e.g. Pensoft, Muséum national d'Histoire naturelle, Consortium of European Taxonomic Facilities Publications, the Zenodo repository, Biodiversity Heritage Library), and data reusers like the Global Biodiversity Information Facility, Ocellus, Synospecies and openBiodiv. Currently, over 500,000 taxonomic treatments and 300,000 illustrations have been liberated and are accessible through TreatmentBank and the Biodiversity Literature Repository.

Keywords

data liberation, digitally accessible knowledge, FAIR

Presenting author

Donat Agosti

Presented at

TDWG 2021

Funding program

Swiss universities; Arcadia Fund; The BiCIKL project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492

Conflicts of interest

References

- Catapano T (2019) TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Zenodo https://doi.org/10.5281/zenodo.3484285
- Costello MJ, May RM, Stork NE (2013) Can We Name Earth's Species Before They Go Extinct? Science 339 (6118): 413-416. https://doi.org/10.1126/science.1230318
- Kalfatovic M (2010) BHL Australia Kick Off Meeting: Melbourne Museum. 1 June 2010. Melbourne, Australia. URL: https://www.slideshare.net/Kalfatovic/3-years-on-the-biodiversity-heritage-library?qid=3a0bdbbc-8b89-4260-a69d-93b58c8c6885&v=&b=&from_search=19
- Linnaeus C (1758) Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 10. Laurentii Salvii https://doi.org/10.5962/bhl.title.542
- Polaszek A, et al. (2005) A universal register for animal names. Nature 437 (7058): 477-477. https://doi.org/10.1038/437477a