Bidirectional Linking: Benefits, challenges, pitfalls, and solutions

Guido Sautter‡, Donat Agosti‡

‡ Plazi, Bern, Switzerland

Corresponding author: Guido Sautter (gsautter@gmail.com)

Abstract

Taxonomy, and biodiversity science in general, mainly revolve around four types of entities, which are available digitally in ever increasing numbers from different services: (1) Physical specimens (kept in museums and other collections around the world) and observations are available digitally via the Global Biodiversity Information Facility (GBIF). (2) DNA sequences (often derived from preserved specimens) are available from the European Nucleotide Archive (ENA) and National Center for Biotechnology Information (NCBI), having accession numbers as their primary means of citation. (3) Taxa, identified by taxon names, are increasingly registered to nomenclatural reference databases (ZooBank, International Plant Names Index (IPNI)) and aggregated in the Catalogue of Life (CoL). (4) Taxonomic treatments combine the former three; they define taxa, express scientific opinions about existing taxa, based upon specimens as well as DNA sequences derived from themand coin respective names; they are available from TreatmentBank (as well as Zenodo/Biodiversity Literature Repository (BLR) and Swiss Institute of Bioinformatics Literature Services (SIBiLS), and GBIF).

Traditionally, treatments cite specimens, taxa, and other treatments in mainly human-centric ways, describing where to find the cited object, but they are not immediately actionable in a digital sense. Specimen citations use institution and collection codes and catalog numbers (often combined with geographical and environmental data). Taxon names are a type of self-citing entities, especially when given in combination with their (bibliographic) authorship, as they represent a historical approach to human-readable taxon identifiers. Citations of treatments are very similar to those of taxon names, adding (bibliographic) information of subsequent name usages as needed. Accession numbers for DNA sequences are the closest to modern digital identifiers. However, none of these means of citation, as usually found in literature, are readily machine actionable, which makes them hard to process at scale and analyze programmatically. Identifiers coined by the various data providers, in combination with APIs to resolve them, alleviate this problem and enable computational navigation of such links. However, this alone only defers the problem, as actionable identifiers (e.g., HTTP URIs) at some point still need to be inferred from the information given in the traditional means of citation where the latter occur in data.

Recent projects, like BiCIKL, aim to add machine navigable links to the various entities (or respective data records) at scale, in pursuit of (ideally) fully intermeshed records, connecting (1) treatments to subject taxon names and concepts, cited specimens and DNA sequences, as well as cited treatments (with explicit nomenclatorial implications, e.g., taxon name synonymies or rebuttals thereof), (2) (digital) specimens to assigned taxon names, citing treatments, and any derived DNA sequences, (3) DNA sequences to source specimens (or their digital counterparts), where applicable, assigned taxon names, and citing treatments, and (4) taxon names to defining and synonymizing treatments, associated (digital) specimens, and any derived DNA sequences. This removes possible issues with transitive dependencies in a sequence of links, as an intermediate point of failure; all major data providers have been doing this to various degrees for some time, which provides a great starting point, but several challenges and pitfalls remain: For valid technical reasons, the systems of the individual data providers are (and need to be) self-contained, which comes at the cost of a certain amount of duplication (e.g., GBIF and ENA/NCBI backbone taxonomies). This is unproblematic per se, but slows down update proliferation and can incur some discrepancies. Further, traditional human-readable identifiers can be somewhat ambiguous: (1) some institution and collection codes are not unique, or authors use them in non-standard ways (some codes in the Global Registry of Scientific Collections (GrSciColl) point to half a dozen different institutions, for instance); (2) certain catalog numbers of museum specimens are also valid (resolvable) accession numbers, with actual semantics only emerging from context; (3) absence of the latter renders the semantics of data presented in tables especially hard to infer; (4) none of the providers has complete data coverage, so linking is not even technically possible in all cases at any given point, and some links can only be added over time, as coverage and thus overlap between data increases (newly published names cannot possibly be in CoL when the defining treatment gets digitized, for instance); (5) occasional full re-computation or re-processing is impractical and wasteful at best.

In this presentation, we discuss various ways of overcoming the outlined challenges and avoiding the described pitfalls, and also make related suggestions for APIs to better support respective mechanisms.

Keywords

taxonomic names, occurrences, DNA sequence, materials citations, treatments

Presenting author

Guido Sautter

Presented at

TDWG 2023

Funding program

The TreatmentBank infrastructure is supported by the Horizon Europe funded project Biodiversity Community Integrated Knowledge Library (BiCIKL), which receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492), the Arcadia Fund and the Swissuniversities funded eBioDiv project.

Conflicts of interest

The authors have declared that no competing interests exist.