Avenues into Integration: Communicating taxonomic intelligence from sender to recipient

Beckett W. Sterner[‡], Nathan Upham[‡], Atriya Sen[‡], Nico M Franz[‡]

‡ Arizona State University, Tempe, United States of America

Abstract

"What is crucial for your ability to communicate with me… pivots on the recipient's capacity to interpret—to make good inferential sense of the meanings that the declarer is able to send" (Rescher 2000, p148).

Conventional approaches to reconciling taxonomic information in biodiversity databases have been based on string matching for unique taxonomic name combinations (Kindt 2020, Norman et al. 2020). However, in their original context, these names pertain to specific usages or taxonomic concepts, which can subsequently vary for the same name as applied by different authors. Name-based synonym matching is a helpful first step (Guala 2016, Correia et al. 2018), but may still leave considerable ambiguity regarding proper usage (Fig. 1). Therefore, developing "taxonomic intelligence" is the bioinformatic challenge to adequately represent, and subsequently propagate, this complex name/usage interaction across trusted biodiversity data networks. How do we ensure that senders and recipients of biodiversity data not only can share messages but do so with "good inferential sense" of their respective meanings?
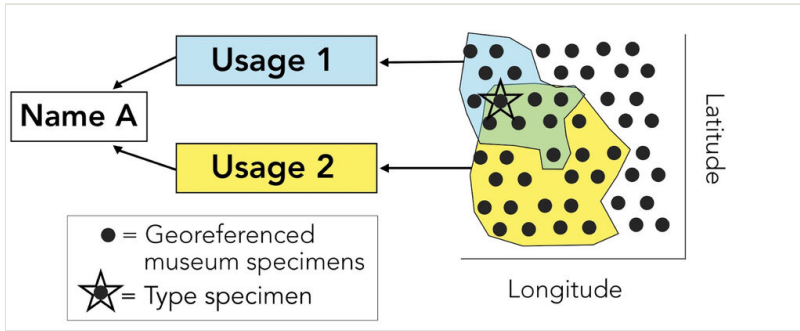
Figure 1.

The problem of taxonomic name/usage (TNU) ambiguity in biodiversity data. Two alternative usages ("1" and "2") of a species name ("A") are shown in their geospatial context as circumscribed by a set of georeferenced museum voucher specimens. Those usages were published as taxonomic opinions by given authors, both circumscribing name A for a species-level entity in different ways. Both usages share the same type specimen and locality (shown by a star) and thus the same name A, authority, and year. Name-based string matching is insufficient for parsing this type of TNU change unambiguously. Taxonomically intelligent methods, yet to be developed in a scalable fashion, are instead required.

Key obstacles have involved dealing with the complexity of taxonomic name/usage modifications through time, both in terms of accounting for and digitally representing the long histories of taxonomic change in most lineages. An important critique of proposals to use name-to-usage relationships for data aggregation has been the difficulty of scaling them up to reach comprehensive coverage, in contrast to name-based global taxonomic hierarchies (Bisby 2011). The Linnaean system of nomenclature has some unfortunate design limitations in this regard, in that taxonomic names are not unique identifiers, their meanings may change over time, and the names as a string of characters do not encode their proper usage, i.e., the name "Genus species" does not specify a source defining how to use the name correctly (Remsen 2016, Sterner and Franz 2017). In practice, many people provide taxonomic names in their datasets or publications but not a source specifying a usage. The information needed to map the relationships between names and usages in taxonomic monographs or revisions is typically not presented it in a machine-readable format.

New approaches are making progress on these obstacles. Theoretical advances in the representation of taxonomic intelligence have made it increasingly possible to implement efficient querying and reasoning methods on name-usage relationships (Chen et al. 2014, Chawuthai et al. 2016, Franz et al. 2015). Perhaps most importantly, growing efforts to produce name-usage mappings on a medium scale by data providers and taxonomic authorities suggest an all-or-nothing approach is not required. Multiple high-profile biodiversity databases have implemented internal tools for explicitly tracking conflicting or dynamic taxonomic classifications, including eBird using concept relationships from AviBase (Lepage et al. 2014); NatureServe in its Biotics database; iNaturalist using its taxon framework (Loarie 2020); and the UNITE database for fungi (Nilsson et al. 2019).

Other ongoing projects incorporating taxonomic intelligence include the Flora of Alaska (Flora of Alaska 2020), the Mammal Diversity Database (Mammal Diversity Database 2020) and PollardBase for butterfly population monitoring (Campbell et al. 2020).

Keywords

taxonomic name, taxonomic concept, logic reasoning, artificial intelligence, data integration, Darwin Core, information retrieval, extended specimen

Presenting author

Beckett W. Sterner

Presented at

TDWG 2020

Funding program

National Science Foundation Science and Technology Studies Program

Grant title

Productive Ambiguity in Classification ID#1827993

Conflicts of interest

None declared.

References

- Bisby F (2011) The Catalogue of Life - building a taxonomic backbone for the world biota. TDWG 2011 Annual Conference. URL: https://mbgocs.mobot.org/index.php/tdwg/2011/paper/view/225
- Campbell D, Thessen A, Ries L (2020) A novel curation system to facilitate data integration across regional citizen science survey programs. PeerJ 8 https://doi.org/10.7717/peerj.9219
- Chawuthai R, Takeda H, Wuwongse V, Jinbo U (2016) Presenting and preserving the change in taxonomic knowledge for linked data. Semantic Web 7 (6): 589-616. https://doi.org/10.3233/sw-150192

- Chen M, Yu S, Franz N, Bowers S, Ludäasher B (2014) Euler/X: A toolkit for logic-based taxonomy integration. arXiv:1402.1992 [cs] URL: http://arxiv.org/abs/1402.1992
- Correia R, Jarić I, Jepson P, Malhado AM, Alves J, Ladle R (2018) Nomenclature instability in species culturomic assessments: Why synonyms matter. Ecological Indicators 90: 74-78. https://doi.org/10.1016/j.ecolind.2018.02.059
- Flora of Alaska (2020) About the new Flora of Alaska – Flora of Alaska. URL: https://floraofalaska.org/about/
- Franz N, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B (2015) Reasoning over taxonomic change: Exploring alignments for the *Perelleschus* use case. PLOS One 10 (2). https://doi.org/10.1371/journal.pone.0118247
- Guala G (2016) The importance of species name synonyms in literature searches. PLOS One 11 (9). https://doi.org/10.1371/journal.pone.0162648
- Kindt R (2020) WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online Taxonomic Backbone data. bioRxiv https://doi.org/10.1101/2020.02.02.930719
- Lepage D, Vaidya G, Guralnick R (2014) Avibase – a database system for managing and organizing taxonomic concepts. ZooKeys 420: 117-135. https://doi.org/10.3897/zookeys.420.7089
- Loarie S (2020) Taxon Frameworks · iNaturalist. URL: https://www.inaturalist.org/pages/taxon_frameworks
- Mammal Diversity Database (2020) About the Mammal Diversity Database. URL: https://mammaldiversity.org/
- Nilsson RH, Larsson K, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research 47 (D1). https://doi.org/10.1093/nar/gky1022
- Norman KA, Chamberlain S, Boettiger C (2020) taxadb: A high-performance local taxonomic database interface. Methods in Ecology and Evolution https://doi.org/10.1111/2041-210X.13440
- Remsen D (2016) The use and limits of scientific names in biological informatics. ZooKeys 550: 207-223. https://doi.org/10.3897/zookeys.550.9546
- Rescher N (2000) Pluralism: Against the demand for consensus. Clarendon Press, Oxford. [ISBN 978-0-19-823601-6]
- Sterner B, Franz N (2017) Taxonomy for humans or computers? Cognitive pragmatics for big data. Biological Theory 12 (2): 99-111. https://doi.org/10.1007/s13752-017-0259-5