Wanted: Standards for FAIR taxonomic concept representations and relationships

Beckett Sterner‡, Nathan Upham‡, Prashant Gupta‡, Caleb Powell‡, Nico M Franz‡

‡ Arizona State University, Tempe, United States of America

Corresponding author: Beckett Sterner (bsterne1@asu.edu)

Abstract

Making the most of biodiversity data requires linking observations of biological species from multiple sources both efficiently and accurately (Bisby 2000, Franz et al. 2016). Aggregating occurrence records using taxonomic names and synonyms is computationally efficient but known to experience significant limitations on accuracy when the assumption of one-to-one relationships between names and biological entities breaks down (Remsen 2016, Franz and Sterner 2018). Taxonomic treatments and checklists provide authoritative information about the correct usage of names for species, including operational representations of the meanings of those names in the form of range maps, reference genetic sequences, or diagnostic traits. They increasingly provide taxonomic intelligence in the form of precise description of the semantic relationships between different published names in the literature. Making this authoritative information Findable, Accessible, Interoperable, and Reusable (FAIR; Wilkinson et al. 2016) would be a transformative advance for biodiversity data sharing and help drive adoption and novel extensions of existing standards such as the Taxonomic Concept Schema and the OpenBiodiv Ontology (Kennedy et al. 2006, Senderov et al. 2018). We call for the greater, global Biodiversity Information Standards (TDWG) and taxonomy community to commit to *extending and expanding* on how FAIR applies to biodiversity data and include practical targets and criteria for the publication and digitization of taxonomic concept representations and alignments in taxonomic treatments, checklists, and backbones.

As a motivating case, consider the abundantly sampled North American deer mouse—*Peromyscus maniculatus* (Wagner 1845)—which was recently split from one continental species into five more narrowly defined forms, so that the name *P. maniculatus* is now only applied east of the Mississippi River (Bradley et al. 2019, Greenbaum et al. 2019). That single change instantly rendered ambiguous ~7% of North American mammal records in the Global Biodiversity Information Facility (n=242,663, downloaded 2021-06-04; GBIF.org 2021) and ⅓ of all National Ecological Observatory Network (NEON) small mammal samples (n=10,256, downloaded 2021-06-27). While this type of ambiguity is common in name-based databases when species are split, the example of *P. maniculatus* is particularly striking for its impact upon biological questions ranging from hantavirus

surveillance in North America to studies of climate change impacts upon rodent life-history traits. Of special relevance to NEON sampling is recent evidence suggesting deer mice potentially transmit SARS-CoV-2 (Griffin et al. 2021).

Automating the updating of occurrence records in such cases and others will require operational representations of taxonomic concepts—e.g., range maps, reference sequences, and diagnostic traits—that are FAIR *in addition to* taxonomic concept alignment information (Franz and Peet 2009). Despite steady progress, it remains difficult to find, access, and reuse authoritative information about how to apply taxonomic names even when it is already digitized. It can also be difficult to tell without manual inspection whether similar types of concept representations derived from multiple sources, such as range maps or reference sequences selected from different research articles or checklists, are in fact interoperable for a particular application. The issue is therefore different from important ongoing efforts to digitize trait information in species circumscriptions, for example, and focuses on how already digitized knowledge can best be packaged to inform human experts and artifical intelligence applications (Sterner and Franz 2017). We therefore propose developing community guidelines and criteria for ***FAIR taxonomic concept representations*** as "semantic artefacts" of general relevance to linked open data and life sciences research (Le Franc et al. 2020).

Keywords

FAIR Principles, open data, taxonomic intelligence

Presenting author

Beckett Sterner

Conflicts of interest

References

- Bisby FA (2000) The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289 (5488): 2309-2312. https://doi.org/10.1126/science.289.5488.2309
- Bradley R, Francis J, Platt R, Soniat T, Alvarez D, Lindsey L (2019) Mitochondrial DNA sequence data indicate evidence for multiple species within Peromyscus maniculatus. Spec. Publ. Mus. Tex. Tech Univ.
- Franz N, Pier N, Reeder D, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B, et al. (2016) Two Influential Primate Classifications Logically Aligned. Systematic Biology 65 (4): 561-582. https://doi.org/10.1093/sysbio/syw023

- Franz NM, Peet RK (2009) Perspectives: Towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity 7 (1): 5-20. https://doi.org/10.1017/s147720000800282x
- Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database 2018 https://doi.org/10.1093/database/bax100
- GBIF.org O (2021) Occurrence Download. The Global Biodiversity Information Facility https://doi.org/10.15468/dl.phjg43
- Greenbaum IF, Honeycutt RL, Chirhart SE (2019) Taxonomy and phylogenetics of the *Peromyscus maniculatus* species group. Spec. Publ. Mus. Tex 17.
- Griffin B, Chan M, Tailor N, Mendoza E, Leung A, Warner B, Duggan A, Moffat E, He S, Garnett L, Tran K, Banadyga L, Albietz A, Tierney K, Audet J, Bello A, Vendramelli R, Boese A, Fernando L, Lindsay LR, Jardine C, Wood H, Poliquin G, Strong J, Drebot M, Safronetz D, Embury-Hyatt C, Kobasa D, et al. (2021) SARS-CoV-2 infection and transmission in the North American deer mouse. Nature Communications 12 (1). https://doi.org/10.1038/s41467-021-23848-9
- Kennedy J, Hyam R, Kukla R, Paterson T, et al. (2006) Standard Data Model Representation for Taxonomic Information. OMICS: A Journal of Integrative Biology 10 (2): 220-230. https://doi.org/10.1089/omi.2006.10.220
- Le Franc Y, Parland-von Essen J, Bonino L, Lehväslaiho H, Coen G, Staiger C, et al. (2020) D2.2 FAIR Semantics: First recommendations. Zenodo https://doi.org/10.5281/zenodo.3707984
- 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
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris R, Penev L, et al. (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (1). https://doi.org/10.1186/s13326-017-0174-5
- 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
- Wagner JA (1845) Arch. Naturgesch. 11(1): 148.
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18