BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e139413
doi: 10.3897/biss.8.139413

OPEN ACCESS

Conference Abstract

Taxonomic Data Objects for Communicating the Meaning of Species Names

Nathan S. Upham[‡], Jorrit Poelen[§,|]

‡ Arizona State University, Tempe, United States of America
§ UC Santa Barbara, Santa Barbara, CA, United States of America
| Ronin Institute, Montclair, NJ, United States of America

Corresponding author: Nathan S. Upham (nathan.upham@asu.edu)

Received: 16 Oct 2024 | Published: 16 Oct 2024

Citation: Upham NS, Poelen J (2024) Taxonomic Data Objects for Communicating the Meaning of Species Names. Biodiversity Information Science and Standards 8: e139413. https://doi.org/10.3897/biss.8.139413

Abstract

A central problem in biodiversity data science remains the inability to precisely aggregate observations that originate from the same species. Current approaches still rely heavily on the '*Genus species*' pair of Linnaean taxonomy to label and group data. However, such binomial species names do not alone contain enough information to distinguish variation in a name's conceptual usage (i.e., meaning) across time, place, or investigator. This "species name-to-meaning" problem is well known, but no robust and scalable solutions yet exist (Berendsohn 1995Bisby 2000, Sutherland et al. 2000, Page 2007, Franz and Sterner 2018, Upham et al. 2021, Sterner et al. 2023, Sterner et al. 2020, Beach et al. 1993). The problem is widespread especially for well-studied groups like mammals, for which 45% more species are now recognized than 30 years ago, including many species splits and rearrangements (Burgin et al. 2018, Mammal Diversity Database 2024). To address this problem, we here propose to digitally package and sign the relevant taxonomic data associated with a given species name usage to form a 'Taxonomic Data Object' (TDO). These TDOs are conceptually similar to '*secundus*' references (i.e., species name *sec.* author, Berendsohn 1995) but differ in being machine readable and thus *operational* from the perspective of high-throughput data science. TDOs aim to move from analog secundus references, which require a human to track meaning by reading the cited article, to a digital and machine-actionable taxonomic reference. Versioned TDOs that are digitally signed using hash algorithms (e.g., md5 or sha256, see Elliott et al. 2023) will allow for precisely communicating, with verified data

provenance from sender to receiver, certain aspects of what a species name means according to a given authority at a given time. Example TDO contents include DarwinCore terms (e.g., scientificName, namePublishedIn, nameAccordingTo, nomenclaturalStatus (Darwin Core Maintenance Group 2021)) as well as meaning-rich digital assets like geographic range maps (e.g., GeoJSON format), exemplar DNA sequences (e.g., FASTA format or National Center for Biotechnology Information (NCBI) accession number), holotype specimen information (e.g., catalog number, type locality coordinates), and taxonomic treatment texts (including material citations, e.g., as digitized and extracted by Plazi TreatmentBank, Agosti and Egloff 2009; see Fig. 1). We discuss pilot examples comparing global bat species according to the v1.2 vs. v1.11 taxonomies of the Mammal Diversity Database (Sep 2020 vs. Apr 2023), showing how TDOs enable the reliable tracking of taxonomic meaning for the 63 bat species affected by splits during this period. We posit that widespread use of TDOs will enable the traceable exchange of taxonomic information across a variety of existing platforms, providing a path for accurate species-level data aggregation at global scales, at least for well-studied taxa.

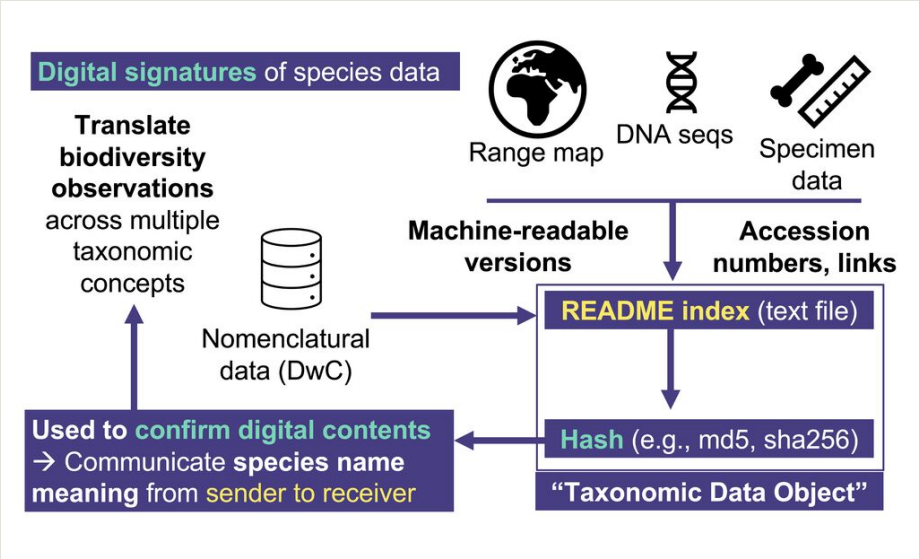

Figure 1.

Construction of Taxonomic Data Objects (TDOs) using algorithmic hashes for precise verification and comparison of TDO contents when versions are sent between different computers.

Keywords

taxonomy, semantic knowledge, databases, reproducibility, hashes

Presenting author

Nathan S. Upham

Grant title

Intelligently predicting viral spillover risks from bats and other wild mammals (1R21AI164268-01/02)

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2 (1). https://doi.org/10.1186/1756-0500-2-53
- Beach J, Pramanik S, Beaman J (1993) Hierarchic taxonomic databases. Ch. 15 (pp. 241-256) in: Fortuner, R.(ed.): Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision. John Hopkings University Press, Baltimore.
- Berendsohn WG (1995) The concept of “potential taxa” in databases. Taxon 44 (2): 207-212. https://doi.org/10.2307/1222443
- 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
- Burgin CJ, Colella JP, Kahn PL, Upham NS (2018) How many species of mammals are there? Journal of Mammalogy 99 (1): 1-14. https://doi.org/10.1093/jmammal/gyx147
- Darwin Core Maintenance Group (2021) Darwin Core Quick Reference Guide. Biodiversity Information Standards (TDWG). URL: https://dwc.tdwg.org/terms/
- Elliott MJ, Poelen JH, Fortes JA (2023) Signing data citations enables data verification and citation persistence. Scientific Data 10 (1): 1. https://doi.org/10.1038/s41597-023-02230-y
- Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database https://doi.org/10.1093/database/bax100
- Mammal Diversity Database (2024) v1.13. Zenodo. https://doi.org/10.5281/zenodo.12738010
- Page R (2007) Towards a Taxonomically Intelligent Phylogenetic Database. Nature Precedings 1-1 https://doi.org/10.1038/npre.2007.1028.1
- Sterner B, Upham N, Sen A, Franz N (2020) Avenues into Integration: Communicating taxonomic intelligence from sender to recipient. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59006
- Sterner B, Elliott S, Gilbert EE, Franz NM (2023) Unified and pluralistic ideals for data sharing and reuse in biodiversity. Database048. https://doi.org/10.1093/database/baad048

- Sutherland I, Robinson J, Brandt SM, Jones AC, Embury SM, Gray WA, White RJ, Bisby FA (2000) Assisting the integration of taxonomic data: The LITCHI toolkit. Proceedings of 16th International Conference on Data Engineering (Cat (00CB37073))679-680. https://doi.org/10.1109/ICDE.2000.839491
- Upham NS, Poelen JH, Paul D, Groom QJ, Simmons NB, Vanhove MP, Bertolino S, Reeder DM, Bastos-Silveira C, Sen A, Sterner B, Franz NM, Guidoti M, Penev L, Agosti D (2021) Liberating host-virus knowledge from biological dark data. The Lancet Planetary Health 5 (10): 746-750. https://doi.org/10.1016/S2542-5196(21)00196-0