Addressing Global Biodiversity Challenges: Ensuring Long-Term Sustainability of Morphological Data Collection and Reuse through MorphoBank

Brooke L. Long-Fox[‡], Ana Andruchow-Colombo[§], Shreya Jariwala[|], Maureen A. O'Leary[¶], Tanya Z. Berardini[‡]

‡ Phoenix Bioinformatics, Newark, United States of America
§ University of Kansas, Lawrence, United States of America
| University of California, Berkeley, Berkeley, United States of America
¶ Stony Brook University, Stony Brook, United States of America

Corresponding author: Brooke L. Long-Fox (blongfox@morphobank.org)

Abstract

Phenotypic, especially morphological, data are highly useful in systematics, taxonomy, and phylogenetics. Despite the increased use of genetic information, phenotypic data are necessary when researching the fossil record and remain useful for living taxa by providing independent evidence for testing molecular clades. MorphoBank is a FAIR (Findable, Accessible, Interoperable, and Reusable) database providing open biodiversity data in the form of morphological characters (O'Leary and Kaufman 2011, O'Leary and Kaufman 2012), a similar concept to GenBank for open access sequence data. MorphoBank enables scientists to share morphological character data associated with their peer-reviewed publications in the form of phylogenetic matrices as TNT or NEXUS files.

MorphoBank hosts 1,738 publicly accessible projects with 173,559 images and 1,138 matrices as of July 2024. These data can be downloaded by the public, researchers, and students in the scientific community, where the data can be used for educational purposes or reused in additional phylogenetic analyses. MorphoBank encourages scientists to add content in numerous ways throughout the research process, including while actively working on a morphological matrix or in conjunction with a paper to be published that has a morphological matrix. For example, P773 (http://dx.doi.org/10.7934/P773) represents collaborative research that contains a matrix with 4,541 characters and over 12,000 annotated images. Researchers looking to replicate or utilize the data from this study, a task that would normally be extremely time and labor intensive, are able to quickly and easily download and work with the data in their own analyses.

MorphoBank has a team of part-time curators and interns who also add content post-publication. Between 2018 and 2023, MorphoBank staff accounted for 25% of project creation and 41% of project publication. The MorphoBank community members created more projects but published fewer of them in the same time frame. The MorphoBank curation team strives to add the matrices to make the data FAIR. A majority of the data are associated with publications in journals that require a subscription; MorphoBank makes the matrix data available with its complete metadata without a financial access barrier. Data standards for morphological character matrices include scored taxa, full taxonomic names, and complete character names with character state descriptions. Since NEXUS files have varying standardization and syntax (Maddison et al. 1997, Vos et al. 2012), importing a matrix can lead to data errors, which MorphoBank does not accept due to its mission to comply with the FAIR standards. Hence, users often add incomplete data as file attachments. To help ensure full data is uploaded, MorphoBank has partnered with journals to ensure instructions to authors or emails to authors of accepted manuscripts make clear the need to upload data matrices to MorphoBank.

MorphoBank has been cited over 1,500 times, with increasing citations each year (Fig. 1). We examined the use and impact of MorphoBank data on systematic and phylogenetic research and found that most data are used in phylogenetic analyses, describing new species, and examining diversification of taxonomic groups which spans a wide-range organisms from vertebrates such as dinosaurs, reptiles, and mammals (including studies of human evolution) to plants, invertebrates, and micro-organisms.

As part of its outreach work under NSF Grant EAR-2148768, MorphoBank has developed and implemented an internship program for undergraduate biology students focused on training in phylogenetic data, curation, research writing, and conference presenting. Part of this intership program involves utilizing Artificial Intelligence (AI) to increase efficiency by automating the process of extraction of character name and state data from published articles and integrating them into NEXUS files.

Three additional outreach activities help raise awareness and increase community contributions to MorphoBank. (1) A partnership with the American Museum of Natural History (AMNH) was established in Summer 2024 to train volunteer curators. (2) MorphoBank workshops have been developed for in-person (i.e. 12th North American Paleontological Convention in Ann Arbor, Michigan) and virtual (i.e. 3rd Joint Congress on Evolutionary Biology through sponsorships from the Society of Systematic Biologists) conferences. (3) Virtual workshops will be offered quarterly to educate the scientific community on ways to add their own phylogenetic data to MorphoBank.

The long-term sustainability of MorphoBank depends on success in three areas.

1. Financial sustainability: MorphoBank is currently supported by membership fees from academic institutions and museums, institutional support from Phoenix Bioinformatics, the non-profit running the resource, and US NSF grants. Its future depends on continued and growth in membership.

2. Technical sustainability: The over 20 year old MorphoBank codebase is being completely overhauled to provide better performance, add longer term software stability, and enable easier addition of new features.

3. Scientific sustainability: The outreach efforts to increase community awareness and contributions aim to ensure the continued relevance and utility of the resource. Growth in data depth and breadth feeds into making MorphoBank indispensable for research in this scientific domain.

Keywords

morphological matrix, phylogenetics, database, repository, FAIR data

Presenting author

Brooke L. Long-Fox

Presented at

SPNHC-TDWG 2024

Acknowledgements

We acknowledge the use of Google Gemini Large Language Model for the integration of matrix and character information. We would also like to thank the Phoenix Bioinformatics Tech Team members Xingguo Chen, Swapnil Sawant, and Kartik Khosa as well as Kenzley Alphonse for their technical development and support of the web resource.

Funding program

MorphoBank is currently supported by the non-profit Phoenix Bioinformatics, NSF-DBI-2049965, NSF-EAR-2148768, and membership fees from academic institutions and museums around the world.

Hosting institution

Phoenix Bioinformatics

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Maddison D, Swofford D, Maddison W (1997) NEXUS: An Extensible File Format for Systematic Information. Systematic Biology 46 (4). https://doi.org/10.2307/2413497
- O'Leary M, Kaufman S (2011) MorphoBank: phylophenomics in the "cloud". Cladistics 27 (5): 529-537. https://doi.org/10.1111/j.1096-0031.2011.00355.x
- O'Leary MA, Kaufman SG (2012) MorphoBank 3.0: Web application for morphological phylogenetics and taxonomy. URL: http://www.morphobank.org
- Vos R, Balhoff J, Caravas J, Holder M, Lapp H, Maddison W, Midford P, Priyam A, Sukumaran J, Xia X, Stoltzfus A (2012) NeXML: Rich, Extensible, and Verifiable Representation of Comparative Data and Metadata. Systematic Biology 61 (4): 675-689. https://doi.org/10.1093/sysbio/sys025

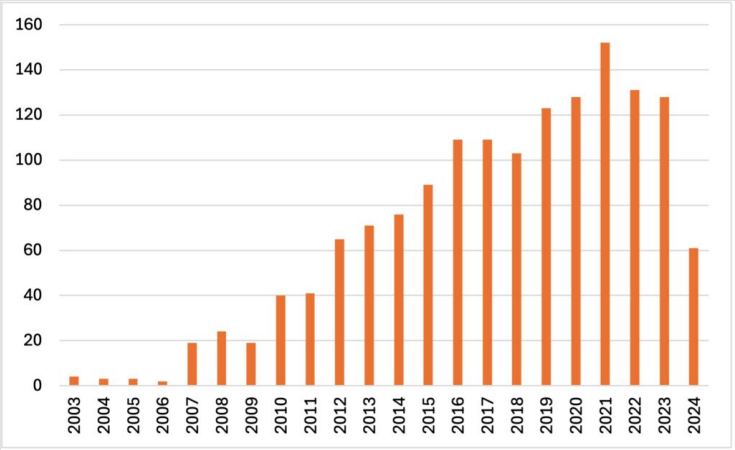

Figure 1.

MorphoBank citations over time (2003 - present) as found in Google Scholar.