

## **NASBR Symposium**

Title: Liberating, linking, and leveraging bat data for research and conservation

The goal of this symposium is to bring together parallel efforts to liberate and link bat taxonomic and ecological data from literature, museum collections, and other sources to make it accessible, standardized, scalable, and searchable by stakeholders ranging from researchers to policymakers. We are living in the Era of Big Data, where an increasing amount of data is being made digitally available each year, with many databases about bats being published for wider use. However, available data are still distributed across multiple publications and platforms, with no clear and objective guidelines for collecting or compiling data across bat species and families. Much written knowledge also remains undigitized as ‘dark data’. This situation makes it challenging for stakeholders and managers to find and use data, as well as for researchers to build upon what is already known. Identifying gaps in our understanding of bat biology, ecology, and conservation is needed to move forward efficiently. This symposium will bring together a series of speakers who are engaged in different initiatives to liberate, link, and leverage bat data for the scientific community. Speakers will range from those assembling new data sets to those using big datasets to inform conservation action.

### **Key Links in the Global Bat Data Chain**

Nancy B. Simmons<sup>1</sup>, Cullen Geiselman<sup>2</sup>, Aja C. Sherman<sup>2</sup>, DeeAnn M. Reeder<sup>3</sup>, Connor J. Burgin<sup>4</sup>, and Nathan S. Upham<sup>5</sup>

*1 Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA, 2 Bat Eco-Interactions Project, Houston, USA, 3 Department of Biology, Bucknell University, Lewisburg, USA, 4 Department of Biology, University of New Mexico, Albuquerque, USA, 5 School of Life Sciences, Arizona State University, Tempe, USA*

Effective scientific research and conservation action require diverse data collected from numerous sources. Scientists routinely work with combinations of data including information mined from old and new published literature, online repositories, museum specimens, and unpublished data inherited from past researchers as well as new data collected in the field, lab, or museum. Research is increasingly collaborative and usually requires integrating or transferring data between various databases. The links between different data sources and databases are thus of key importance. Whether working with checklist data, occurrence data, sampling metadata, or other types of information, it is essential that uniform standard identifiers be used so that different databases can be linked within a common framework. What are the key links in the global bat data chain? We would argue that the key identifiers are those that indicate “what” (species, individuals or groups), “where” (e.g., GPS coordinates), “when” (e.g., date, time), and “by whom” (e.g., collector, observer, sampler). These all seem like straightforward pieces of information, but different researchers often report these in different ways (e.g., using degrees-minutes-seconds versus decimal degrees for latitude and longitude; day-month-year versus month-day-year versus year-month-day for dates; and so on). Minimally we suggest that all researchers consider adoption of Darwin Core standards for recording their data so that it can more easily be shared and combined with other data sets. We argue that it is time to re-think how data are collected, reported, and linked at all steps in the research process.

### **Challenges and Perspectives in Collating Bat Trait Data Worldwide**

Hernani F. M. Oliveira<sup>1</sup>, Nancy B. Simmons<sup>2</sup>, Cecilia Montauban<sup>3</sup>, and Ariadna E. Morales<sup>4</sup>

*1 Department of Zoology, Universidade Federal do Paraná, Curitiba, BRA; 2 Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA; 3 Department of Life Sciences, Imperial College London, London, UK; 4 Centre for Translational Biodiversity Genomics, Senckenberg Research Institute & Goethe-University, Faculty of Biosciences, Frankfurt, DEU*

Bats are among the most taxonomically, ecologically, and functionally diverse mammals, exhibiting a widespread distribution with high local species richness in tropical regions. Their complex evolutionary history and unique adaptations have resulted in diverse forms and behaviors across species, complicating the standardized measurement and collation of trait data over space and time. Morphological traits have traditionally been the primary means of species characterization and understanding of ecological and evolutionary processes. However, to fully comprehend the functional roles of bats within ecosystems and investigate potential effects of global change, it is essential to integrate a broader range of functional traits. This requires standardized assessments and integration with updated taxonomy and correct distribution ranges, to ensure an accurate database of functional trait data for all bat species. The challenge lies in reconciling past, present, and future measurement data in a holistic framework, which demands specialized expertise. This presentation addresses strategies to overcome these challenges by identifying knowledge gaps across geographies and taxonomic scales, and proposing standardized measurement methods. By synthesizing existing data and exploring innovative approaches, we aim to enhance our understanding of bat diversity and their ecological functions, facilitating future research in bat biology and conservation.

### **Bat Data for Conservation Planning: The What, How, and Why**

Winifred F. Frick<sup>1,2</sup>, Brian E. Reichert<sup>3</sup>, Cecilia Montauban<sup>4</sup>, Caitlin J. Campbell<sup>1</sup>, and M. Teague O'Mara<sup>1,5,6</sup>

*1 Bat Conservation International, Austin, USA; 2 Department of Ecology and Evolutionary Biology, University of Santa Cruz, Santa Cruz, USA; 3 USGS Fort Collins Science Center, Fort Collins, USA; 4 Department of Life Sciences, Imperial College London, London, UK; 5 Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, DEU; 6 Smithsonian Tropical Research Institute, Balboa, PAN*

Conservation planning relies on information to guide priority setting, identify species in need, evaluate efficacy of actions, and monitor long-term outcomes. Numerous databases exist that collate or curate data useful for bat research and conservation. Some of these databases focus on bats, while others have a broader biodiversity or conservation purpose. Much progress has been made in the past decades toward implementing open data standards and improving the availability of information through a suite of different applications, tools, and data inventories. However, data deficiency remains one of the vexing challenges for bat conservation globally, with significant geographical and taxonomic gaps in knowledge. We provide an overview of existing and developing databases and tools, and discuss the need for integrating bat data into platforms that feed directly into conservation decision making. We evaluate existing tools, such as Inventory Biodiversity Assessment Tool used for impact assessments, to highlight the importance of collating and sharing bat data to improve bat conservation planning and outcomes. We provide examples from the North American Bat Monitoring Program as one model for coordinated data collection and analysis to inform conservation planning, and highlight the efforts of the Global Union of Bat Diversity Networks (GBatNet) to create databases that enable research and conservation for bats. Open data sharing and integrating bat data into existing biodiversity databases and platforms can enable conservation efforts to protect threatened and vulnerable bat populations.

### **Why Taxonomic Splits Matter for Bat Biodiversity and Viral Risk Analyses**

Nathan S. Upham<sup>1</sup>, Norma Alicia Hernández Hernández<sup>1</sup>, Ángel Luis Robles Fernández<sup>1</sup>, Beckett Sterner<sup>1</sup>, Nico Franz<sup>1</sup>, Aja Sherman<sup>2</sup>, Cullen Geiselman<sup>2</sup>, Nancy B. Simmons<sup>3</sup>, Kendra Phelps<sup>4</sup>, Anna Willoughby<sup>5</sup>, Quentin Groom<sup>6</sup>, Donat Agosti<sup>7</sup>, DeeAnn M. Reeder<sup>8</sup>, Jorrit H. Poelen<sup>9</sup>, and Connor J. Burgin<sup>10</sup>

*1 School of Life Sciences, Arizona State University, Tempe, USA; 2 Bat Eco-Interaction Project, Houston, USA; 3 Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA; 4 EcoHealth Alliance, New York, USA; 5 Odum School of Ecology, University of Georgia, Athens, USA; 6 Meise Botanic Garden, Meise, BEL; 7 Plazi, Berne, SWZ; 8 Department of Biology, Bucknell University, Lewisburg, USA; 9*

*Ronin Institute for Independent Scholarship, Montclair, USA; 10 Department of Biology, University of New Mexico, Albuquerque, USA*

Taxonomy is generally viewed as external to ecology, a fixed framework rather than a variable — the taxonomic variable — that needs tuning to understand its impact on inferences. However, since taxonomy is a human-imposed perspective rather than something innate, biodiversity scientists have the responsibility to quantify its impact on derived knowledge. Here we study how the 35% increase in recognized bat species (Chiroptera) since 1993 has impacted our knowledge of bat-virus interactions and, in turn, inferences of viral spillover risk to humans. We focus on change in the geographic concepts of bat species globally due to taxonomic splits between two periods: (i) 2008-2020 using IUCN range maps (based on Mammal Species of the World, 3rd edition) relative to Mammal Diversity Database (MDD) v1.2 range maps; and (ii) 2020-2023 comparing MDD v1.2 to newly produced v1.11 maps. We then intersect these conceptual changes with known bat-virus interactions from databases to assess their impact upon risks of cross-species viral transmission. We find that taxonomic splits affected 248 bat species across both periods (185 and 63 species, respectively), which has impacted 16.9% of currently recognized bat species globally. Those taxonomic splits have rendered ambiguous 3,249 bat-virus interactions — 21.9% of all digitally known observations — since these data are indexed by host species name rather than observation location, which leads to ambiguity when species are split. We discuss high-throughput solutions for accurately translating the taxonomy of these data, and the impact of inaction upon estimates of viral spillover risk.

### **Bat Taxonomic Alignment Tool: Valid Names for Conservation Assessments**

Aja C. Sherman<sup>1</sup>, Cullen Geiselman<sup>1</sup>, Nancy B. Simmons<sup>2</sup>, Jorrit H. Poelen<sup>3,4</sup>, Donat Agosti<sup>5</sup>, DeeAnn M. Reeder<sup>6</sup>, Nathan S. Upham<sup>7</sup>, and Kendra Phelps<sup>8</sup>

*1 Bat Eco-Interaction Project, Houston, USA; 2 Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York; 3 Ronin Institute for Independent Scholarship, Montclair, USA; 4 Center for Biodiversity and Ecological Restoration, UC Santa Barbara, Santa Barbara, USA; 5 Plazi, Berne, SWZ; 6 Department of Biology, Bucknell University, Lewisburg, USA; 7 School of Life Sciences, Arizona State University, Tempe, USA; 8 EcoHealth Alliance, New York, USA*

We created a dynamic and versioned alignment tool, the Bat Taxonomic Alignment (BTA), to address the extensive modifications in taxonomic names of 1,450+ bat species. The BTA integrates data from three of the most utilized treatments, The Handbook of the Mammals of the World Volume 9 Bats (HMW), Bat Species of the World: A Taxonomic and Geographic Database (Batnames, Version 2022B), and Mammal Diversity Database (MDD, Version 1.9.1). This compilation contains 1,532 taxonomic treatments and reveals the three authorities differed in a third of their taxonomic assignments in 2022. The BTA allows users to accurately integrate and compare taxonomic treatments while maintaining data provenance. Additionally, it includes a comprehensive list of alternative names and synonyms. The BTA also integrates the IUCN Red List (version 2022) which lacked 181 bat species currently accepted by Mammal Diversity Database and Batnames, as well as 124 superseded epithets. This tool was used by the GBatNet Working Group to easily identify missing species and propose updates to the Red List to facilitate more accurate species status assessments, especially as species complexes are split resulting in significant range changes.

### **Leveraging Bat Genomic Data: Exceptional Adaptations, Biodiversity, and Human Health**

Ariadna E. Morales<sup>1</sup>, Hernani F. M. Oliveira<sup>2</sup>, Cecilia Montauban<sup>3</sup>, Nancy B. Simmons<sup>4</sup>, Cullen Geiselman<sup>5</sup>, Aja C. Sherman<sup>5</sup>, Nathan S. Upham<sup>6</sup>, Michael Hiller<sup>1</sup>, and Bat1K Consortium<sup>7</sup>

*1 Centre for Translational Biodiversity Genomics, Senckenberg Research Institute & Goethe-University, Faculty of Biosciences, Frankfurt, DEU; 2 Department of Zoology, Universidade Federal do Paraná, Curitiba, BRA; 3 Department of Life Sciences, Imperial College London, London, UK; 4 Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA 10024; 5 Bat Eco-Interactions Project,*

Houston, USA; **6** School of Life Sciences, Arizona State University, Tempe, Arizona, USA; **7** BatIK Consortium, Frankfurt, DEU

Genomic and transcriptomic data represent crucial resources for decoding genotype information into phenotypes. They not only aid in identifying, classifying, and understanding biodiversity but also offer significant insights that can inform our understanding of evolutionary processes and even human health. Developments in sequencing technologies have accelerated data generation, but systematic data quality assessments (e.g., completeness and reusability) have been left behind. We present an overview of numerous bat genomes that have been assembled and annotated under the BatIK sequencing consortium. We discuss the scope of questions that can be addressed by different types of data (e.g. short- vs. long-read genomes, or annotations based on orthology projections-only vs. transcriptome-based), including potential discoveries linked to skin coloration, vocal learning, wing development, dietary adaptations, and tolerance of viruses that can be deadly for other mammals. It becomes evident that unified data gathering is not just a suggestion but a necessity. The establishment of unified public repositories, housing not only samples, tissues, and specimens but also the associated data and comprehensive guidelines for replicating analyses, is crucial. This approach will not only foster collaboration but also ensure the availability of comprehensive data and current discoveries to leverage future research and conservation efforts. Morpho-genomic integration and coordinated collaborations with current efforts by BatIK and other networks and working groups united by the GBatNet are strongly needed to promote a collaborative community and maximize data usability.

### **Building Datasets for Machine Learning: Lessons from the Bat Co-Roosting Project**

Cullen Geiselman<sup>1</sup>, Aja C. Sherman<sup>1</sup>, Nancy B. Simmons<sup>2</sup>, Nathan S. Upham<sup>3</sup>, Jorrit H. Poelen<sup>4,5</sup>, DeeAnn M. Reeder<sup>6</sup>, Quentin Groom<sup>7</sup>, Kendra Phelps<sup>8</sup>, Donat Agosti<sup>9</sup>, and Anna R. Willoughby<sup>10,11</sup>

*1* Bat Eco-Interactions Project, Houston, USA; *2* Department of Mammalogy, American Museum of Natural History, New York, USA; *3* School of Life Sciences, Arizona State University, Tempe, USA; *4* Ronin Institute for Independent Scholarship, Montclair, USA; *5* Cheadle Center for Biodiversity and Ecological Restoration, UC Santa Barbara, Santa Barbara, USA; *6* Department of Biology, Bucknell University, Lewisburg, USA; *7* Meise Botanic Garden, Meise, BEL; *8* EcoHealth Alliance, New York, USA; *9* Plazi, Berne, SWZ; *10* Odum School of Ecology, University of Georgia, Athens, USA; *11* Center for the Ecology of Infectious Diseases, University of Georgia, Athens, USA

Bat roosts could be critical locations for infectious disease spillover events, but they have been widely overlooked in part due to the disaggregated nature of roosting reports in the literature. Most roost studies have narrow research scopes focusing on roost type, species presences, and population sizes, while a significant amount of roosting information is buried in the methods sections of studies unrelated to roosting behavior. We tested the feasibility of aggregating roosting information from a curated list of 333 publications spanning 1877-2021 and built a training dataset of 17,000+ ecological interactions (including co-roosting, trophic, anthropogenic, and parasitic interactions) involving 630 bat species in 118 countries. We followed open-access and FAIR (Findable, Accessible, Interoperable, and Reusable) data principles for manually mining data, which took one person approximately 575 hours to collect. This open-access dataset, available at the Coronavirus-Host community at Zenodo, contains roosting events that are aligned with multiple ontologies (interaction terms, taxonomies, administrative regions) and are linked with their original metadata, including the verbatim text describing the interaction. The scientific names of subject and target taxa were later verified/corrected using the Bat Taxonomic Alignment tool and Global Biodiversity Information Facility (GBIF), which added another 75 hours of manual work. Though the current dataset reveals geographic and taxonomic patterns in roosting behavior, its main purpose is as a gold-standard dataset for training Large Language Models (LLMs) to automatically extract interaction data from literature for rapid recovery of bat ecological knowledge.

### **BatLit - Mobilizing Bat Literature through Existing Collaborations, Platforms, and Open Source Tools**

Jorrit H. Poelen<sup>1</sup>, Aja C. Sherman<sup>2</sup>, Anna Willoughby<sup>3</sup>, Donat Agosti<sup>4</sup>, Nancy B. Simmons<sup>5</sup>, Nathan S. Upham<sup>6</sup>, and Cullen Geiselman<sup>2</sup>

*1* Cheadle Center for Biodiversity and Ecological Restoration, UC Santa Barbara, Santa Barbara, USA; *2* Bat Eco-Interactions Project, Houston, USA; *3* Odum School of Ecology, University of Georgia, Athens, USA; *4* Plazi, Bern, SWZ; *5* Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA; *6* School of Life Sciences, Arizona State University, Tempe, USA

Bat researchers rely on access to a vast corpus of bat literature to help advance our understanding of bats and the ecosystems they live in. Many researchers build and organize their personal literature collections using mainstream digital tools like Zotero and EndNote, whereas others use homegrown digital methods or even manage their collections manually. However, all researchers routinely encounter roadblocks to literature access including paywalls and older literature resources that have not yet been digitized. To help provide access to bat research literature for all, Plazi (<https://plazi.org>) and the GBatNet Bat Eco-Interactions Working Group are compiling the Bat Literature Corpus (BatLit). BatLit is an actively managed, digital, versioned, and citable collection of bat research literature and associated metadata compiled from existing literature contributed by bat researchers. BatLit is designed to be used in manual (e.g., point-and-click) as well as automated workflows (e.g., text mining, language model training), and can be accessed in many ways, including, but not limited to, external storage media, Zenodo and GitHub. As BatLit continues to improve and grow, we aim to continue to democratize access to bat literature, accelerate research, and help reduce the barrier to knowledge for bat researchers around the world. We invite you to contribute your reference library, especially the PDFs, to BatLit and help increase information access for all.

### **Increasing Accessibility of Historical Nomenclatural Information for Bat Systematic Research**

Connor J. Burgin<sup>1</sup>, Nathan S. Upham<sup>2</sup>, and Nancy B. Simmons<sup>3</sup>

*1* Department of Biology, University of New Mexico, Albuquerque, USA; *2* School of Life Sciences, Arizona State University, Tempe, USA; *3* Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, USA

Integrative taxonomic revisions hinge on the ability of systematists to access historical taxonomic and nomenclatural information associated with the scientific names of their study organisms. This includes species-level synonyms, original citations and descriptive texts, type localities, and type specimen identities and locations, which is particularly difficult to access for names initially described in older literature or with type specimens in museums without online repositories. With the aim of making this information more accessible, we combed the primary literature and taxonomic compendia to compile nomenclatural data for every scientific name applied to a bat on the Mammal Diversity Database (MDD; v1.13) in collaboration with the Batnames and Hesperomys databases. So far, we have identified 4,030 valid and synonymous names applicable to the 1,483 currently recognized bat species since 1758, including 3,471 available names, 263 unavailable names, 257 spelling alterations, and 39 name combinations. The original name combination and authority are available for every name, with other nomenclatural data available for the majority of applicable names: verified original citations (3,533), online links to original publications (3,321), type localities (3,617), and type specimens (3,064). Integration of these data into the MDD makes this normally difficult to access information more accessible to taxonomic researchers, potentially accelerating the process by which integrative taxonomic revisions are proposed and implemented for broader applications in conservation litigation and biological research.