Mapping the Void: An In-depth Analysis of the VertNet Database to Unearth Global Gaps in Avian Specimen and Tissue Collections

Vinay K. L.

Introduction

Museums and natural history collections are important and invaluable sources of Earth history's biodiversity information. Often spanning centuries, museum collections represent a treasure trove of biological specimens and data. The specimens in natural history collections serve as the basis for numerous basic science and research fields. Natural history collections also contain deep, taxon-specific information derived from global populations of plants, animals, fungi, and microorganisms. (Hope, Sandercock, and Malaney 2018; Johnson et al. 2011; Card et al. 2021) By evaluating how species evolve, museum specimens offer crucial baseline data for studies of conservation and emerging diseases (such as Hantavirus and West Nile Virus). The importance of museum collections lies in their ability to serve as both a historical record of life on our planet and a contemporary resource for scientific inquiry.

What initially began as cabinets filled with mounted specimens has transformed into vast, digitized repositories. With the advent of modern databases and digitization efforts, these collections are now more accessible and interconnected than ever. VertNet is one such effort to combine museum collection records from over 250 natural history collections. (Constable et al. 2010) VertNet, a comprehensive repository of biodiversity data, serves as an invaluable resource for assessing the state of biodiversity worldwide. Exploiting VertNet's expansive dataset facilitates a profound comprehension of the immense diversity encompassed within avian taxa and provides the tools to identify significant gaps in our intellectual understanding. By meticulously analyzing this wealth of information, we can discern regions characterized by a need for avian tissue collections. These taxonomic groups remain underrepresented in scientific investigations and temporal gaps. These conclusions play a crucial part in determining how future specimen collection initiatives will proceed. With this background, we set out to understand a) Where are the significant data gaps regarding specimen collection and tissue samples? 2) Which are the over and under-represented groups/families of birds in the museum collection.

Methods

The compilation of museum collection data was executed through two distinct methodologies. Firstly, data acquisition transpired through direct downloads from the VertNet website and leveraged the rvertnet package,(Chamberlain 2021) explicitly employing the 'bigsearch' option. Post retrieval, a meticulous sequence of data refinement procedures was implemented. The initial step involved merging
two datasets, retaining only singular, distinct records. Notably, the VertNet dataset incorporated a spectrum of entries, including fossil specimens, eggs, and nest records, which were systematically removed from the dataset. Scrutiny of the unique country records exposed over 1100 entries, underscoring multiple nomenclatures for identical countries. Consequently, a consolidation process is employed, integrating countries and appending iso3c country codes by utilizing the countrycode package.(Arel-Bundock, Enevoldsen, and Yetman 2018) Subsequently, collection years were binned into 5-year intervals.

Furthermore, a comprehensive data cleansing regimen was implemented to eliminate any instances of missing genus and species names. Subsequently, orders and families underwent systematic refinement to standardize records in accordance with the IOC taxonomy.(n.d.) Upon completion of requisite columnar cleansing, the refined dataset was exported in CSV format, exclusively retaining selected columns for subsequent analytical procedures.

The refined dataset later served as the foundational substrate for visualizing specimen and tissue collections spatially and temporally. The rnaturalearth package (Massicotte and South 2023) facilitated the importation of world map data, enabling the visualization of bird specimens and tissue collections across various collection years. Following the visualization process, a meticulous analysis ensued, wherein standardized residuals were calculated to find the significantly underrepresented and/or overrepresented groups of birds among order and family of birds. This methodological cascade underscores the rigor and precision applied

to both the preprocessing and analytical phases of the investigation, enhancing the scholarly integrity and relevance of the ensuing findings.

Results

The merging of two datasets culminated in a dataset comprising 1.508 million raw records of museum collections. Following meticulous data cleansing procedures, the resultant dataset retained approximately 1.18 million records, with approximately 138,000 featuring tissue records. As illustrated in Figure 1, museum collections commencing from 1690 and extending to 2015, the collection data exhibits a notable acceleration from 1870 onwards, reaching its peak between 1928 and 1932 with an approximate tally of 23,000 specimens globally.

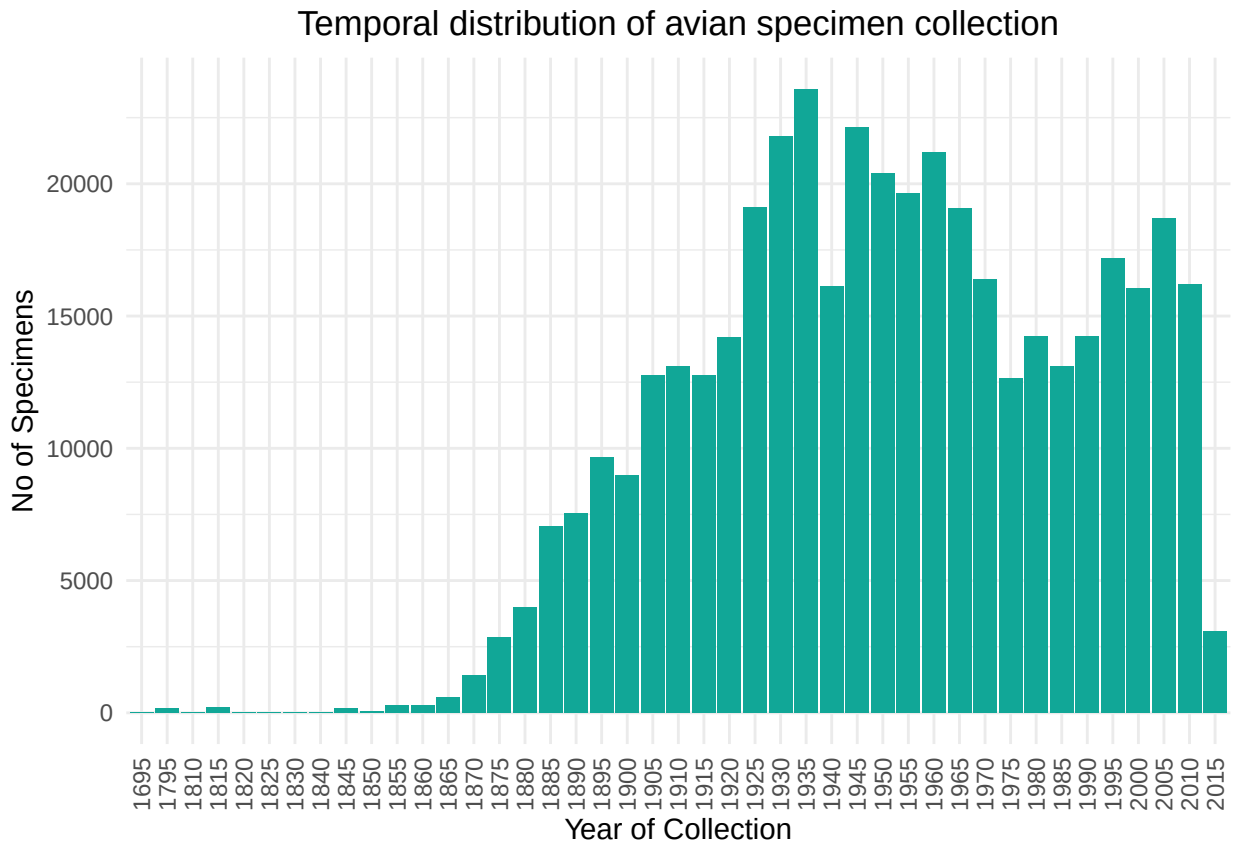

Figure 1: Avian Specimen collection in museums across years binned to 5 years winddow

Among the 237 countries catalogued, records are distributed across 190 nations, with conspicuous concentrations in the United States of America, as depicted in Figure 2. In contrast, the central region of Africa emerges as an undersampled center devoid of recorded specimens. The taxonomic exploration of class Aves, delineated by the IOC Bird taxonomy, elucidates a taxonomy comprising 41 orders and 255 families. The collection records, spanning 40 orders and 232 families underscore a comprehensive representation of avian biodiversity.

Coming to the tissue collection, a pronounced surge in "modern" tissue collection is discernible from the late 1970s, as depicted in Figure 3.

The United States of America assumes a prominent role in tissue collection, contributing slightly over 60,000 specimens. In contrast to the broader specimen collections, tissue collection encompasses only 145 countries, revealing a pronounced void in global representation. The taxonomic spectrum of tissue collection spans 37 orders and 197 families of avian taxa, accentuating the specificity and selectivity inherent in this facet of scientific exploration.

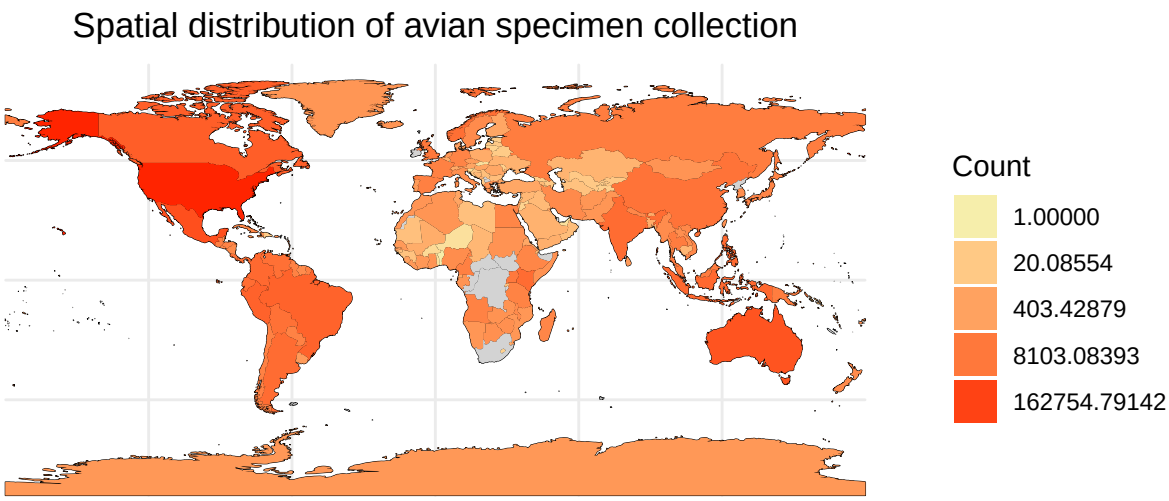

Figure 2: Spatial distribution of the avian specimen collections across globe. Red indicates high and yellow indicates low numbers whereas light grey shows the absence of specimen in logarithmic scale

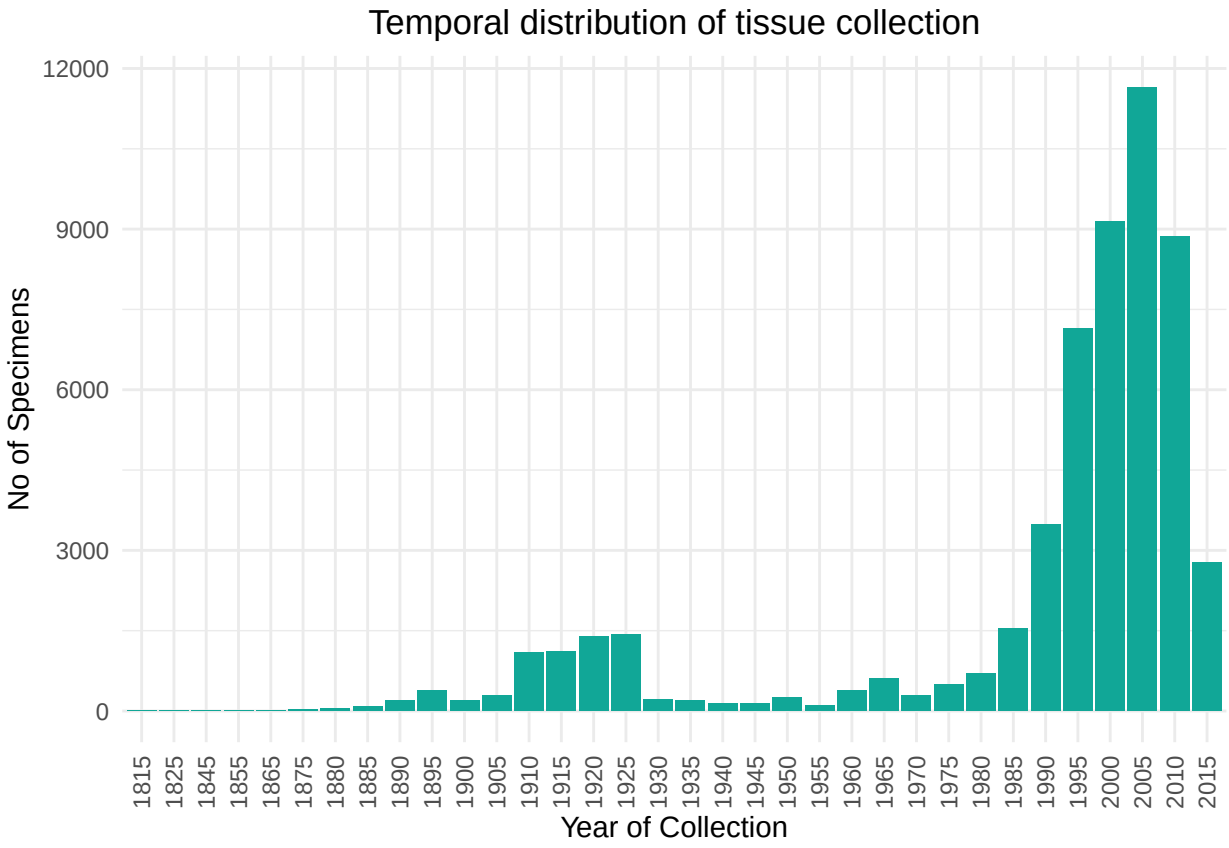

Figure 3: Avian tissue only collection in museums across years binned to 5 years winddow

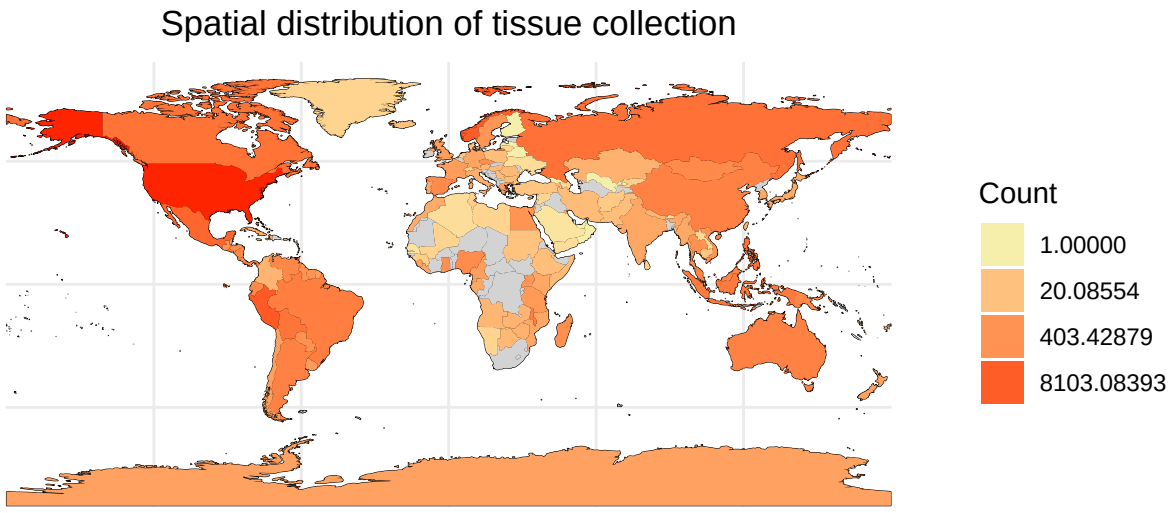

Figure 4: Spatial distribution of the avian tissue collections across globe. Red indicates high and yellow indicates low numbers whereas light grey shows the absence of specimen in logarithmic scale

Figure 5: Standardized Residuals - Orders. Chi-sqaure test p-value $< 2.2e\text{-}16$. Positive value indicates the overrepresentation and vice versa

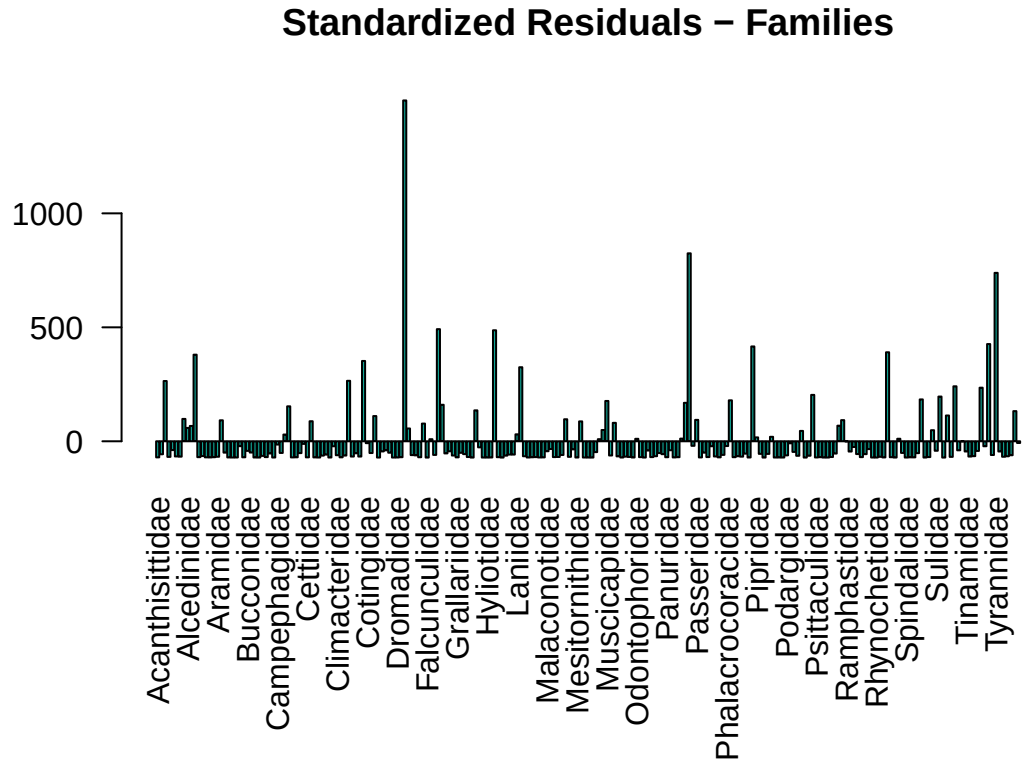

Figure 6: Standardized Residuals - Families. Chi-sqaure test p-value $<$ 2.2e-16. Positive value indicates the overrepresentation and vice versa

Figure 5 and Figure 6 show the Standardized residuals of the Chi-square test for order and families of class aves, respectively.

Discussion

The temporal analysis of the dataset revealed intriguing patterns in collection dynamics over the last three centuries, with a significant acceleration post-1870s. This period corresponds to heightened global exploration and scientific expeditions, emphasizing the historical context that shaped museum collections.(Bakker et al. 2020) The distribution of records across 190 nations underscores the global reach of these collections, with concentrations in the United States and notable undersampling in central Africa. These patterns prompt consideration of historical, geopolitical, and ecological factors influencing collection efforts, offering a nuanced perspective on the evolution of biodiversity documentation.(Bakker et al. 2020)

Zooming into the taxonomic exploration of class Aves, the study reveals a comprehensive representation of avian biodiversity across 41 orders and 255 families. This rich taxonomic spectrum underscores the importance of museum collections in providing a detailed understanding of the diversity within avian taxa. However, the analysis also highlights gaps and biases in the representation of certain orders and families, raising questions about the factors influencing collection priorities. A few potential reasons for the underrepresentation of certain taxa could be habitat inaccessibility, research focus, conservation status and religious believes as explained in Bakker et al., 2020 (Bakker et al. 2020). This exploration of taxonomic nuances enhances the depth of understanding regarding the completeness and comprehensiveness of the available avian specimen collections.

Transitioning to tissue collections, the results shed light on the delayed inception in the 19th century due to challenges in reliable tissue preservation methods. The pronounced surge in “modern” tissue collection from the late 1970s, particularly in the United States, points towards technological advancements and shifting scientific paradigms. This shift’s implications impact molecular research, genetic studies, and the understanding of diseases affecting avian populations.(Bi et al. 2013; Besnard et al. 2014) The global void

in tissue representation, encompassing only 145 countries, raises questions about accessibility, infrastructure, and collaborative efforts in scientific research. Data shows a considerable amount of tissue repository pre-1950s, which is unlikely given that there were no reliable methods to preserve the tissue samples.(Seutin, White, and Boag 1991) Hence, the records could be either due to human error or toe pad clippings cut much later in the collection assigned as a tissue. Thus, this opens the scope of the global landscape of scientific exploration, emphasizing the need for inclusivity and equity in the representation of biological diversity in tissue collections.

In a broader context, this study contributes to the ongoing dialogue on the significance of museum collections as invaluable reservoirs of biodiversity information. It reinforces the dual role of collections as historical records and contemporary scientific resources, urging continued support for the digitization and accessibility initiatives that enhance the utility of museums. These findings call attention to the importance of targeted specimen and tissue collection initiatives to fill gaps in representation, fostering a more holistic understanding of avian specimen collections. Ultimately, this study could catalyze future discussions and actions to advance the integrity and relevance of natural history collections in the face of evolving scientific landscapes and global challenges.

Caveats

As with any scientific study, this study is subject to certain constraints and may only comprehensively address some pertinent questions. The reliance on a publicly available database, limited to data until 2015, is acknowledged as a potential limitation. Subsequent to this period, concerted efforts have been made to enhance the comprehensiveness of global collection databases. A notable proportion of records originate from the pre-colonial era, introducing a potential source of imprecision in determining the precise specimen collection locations. Furthermore, the dynamic nature of the Tree of Life taxonomy is recognized, with ongoing updates and revisions. It is important to note that the lower-level taxonomy employed in this study needed to be standardized, introducing a potential source of bias in the results. Thus, the inferences drawn from this study should be exercised cautiously, considering these inherent limitations.

Acknowledgments

I thank Dr Brant Faircloth for helping me conceptualise the project and discuss the results. I would also like to thank Naman Goyal, Eamon Corbett and Oluwaseun Akinsulire for the meaningful discussion. I thank Dr Dijing Li for the opportunity to conduct this study as a part of class project.

Appendix

Other R packages used in the study are listed below tidyverse(Wickham et al. 2019), dplyr(Wickham et al. 2023), lubridate(Grolemund and Wickham 2011), stringr(Wickham 2023), readr(Wickham, Hester, and Bryan 2023), rvest(Wickham 2022), sf(Pebesma 2018), ggplot2(Wickham 2016), ggthemes(Arnold 2021), cowplot(Wilke 2020)

References

n.d. *IOC World Bird List*. https://www.worldbirdnames.org/new/.

Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman. 2018. "Countrycode: An r Package to Convert Country Names and Country Codes." *Journal of Open Source Software* 3 (28): 848. https://doi.org/10.21105/joss.00848.

Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'*. https://CRAN.R-project.org/package=ggthemes.

Bakker, Freek T, Alexandre Antonelli, Julia A Clarke, Joseph A Cook, Scott V Edwards, Per GP Ericson, Søren Faurby, et al. 2020. "The Global Museum: Natural History Collections and the Future of Evolutionary Science and Public Education." *PeerJ* 8: e8225.

Besnard, Guillaume, Pascal-Antoine Christin, Pierre-Jean G Malé, Emeline Lhuillier, Christine Lauzeral, Eric Coissac, and Maria S Vorontsova. 2014. "From Museums to Genomics: Old Herbarium Specimens Shed Light on a C3 to C4 Transition." *Journal of Experimental Botany* 65 (22): 6711–21.

Bi, Ke, Tyler Linderoth, Dan Vanderpool, Jeffrey M Good, Rasmus Nielsen, and Craig Moritz. 2013. "Unlocking the Vault: Next-Generation Museum Population Genomics." *Molecular Ecology* 22 (24): 6018–32.

Card, Daren C, Beth Shapiro, Gonzalo Giribet, Craig Moritz, and Scott V Edwards. 2021. "Museum Genomics." *Annual Review of Genetics* 55: 633–59.

Chamberlain, Scott. 2021. *Rvertnet: Search 'Vertnet', a 'Database' of Vertebrate Specimen Records*. https://CRAN.R-project.org/package=rvertnet.

Constable, Heather, Robert Guralnick, John Wieczorek, Carol Spencer, A Townsend Peterson, and VertNet Steering Committee. 2010. "VertNet: A New Model for Biodiversity Data Sharing." *PLoS Biology* 8 (2): e1000309.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hope, Andrew G, Brett K Sandercock, and Jason L Malaney. 2018. "Collection of Scientific Specimens: Benefits for Biodiversity Sciences and Limited Impacts on Communities of Small Mammals." *BioScience* 68 (1): 35–42.

Johnson, Kenneth G, Stephen J Brooks, Phillip B Fenberg, Adrian G Glover, Karen E James, Adrian M Lister, Ellinor Michel, et al. 2011. "Climate Change and Biosphere Response: Unlocking the Collections Vault." *BioScience* 61 (2): 147–53.

Massicotte, Philippe, and Andy South. 2023. *Rnaturalearth: World Map Data from Natural Earth*. https://CRAN.R-project.org/package=rnaturalearth.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Seutin, Gilles, Bradley N White, and Peter T Boag. 1991. "Preservation of Avian Blood and Tissue Samples for DNA Analyses." *Canadian Journal of Zoology* 69 (1): 82–90.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.