DOI: https://doi.org/10.3897/arphapreprints.e125974

Retrieving biodiversity data from multiple sources: making secondary data standardized and accessible

Nubia Marques, Carla Soares, Daniel Casali, Erick Guimarães, Fernanda Fava, João Abreu, Ligiane Moras, Leticia Gomes, Raphael Matias, Rafael Assis, Rafael Fraga, Sara Almeida, Vanessa Lopes, Rafaela Missagia, Eduardo Carvalho, Nikolas Carneiro, Ronnie Alves, Pedro Souza, Guilherme Oliveira, Valeria Tavares

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 23/04/2024*. DOI: https://doi.org/10.3897/arphapreprints.e125974

Retrieving biodiversity data from multiple sources: making secondary data standardized and accessible

Nubia Marques[‡], Carla Danielle de Melo Soares[‡], Daniel de Melo Casali[‡], Erick Cristofore Guimarães[‡], Fernanda Guimarães Fava[‡], João Marcelo da Silva Abreu[‡], Ligiane Martins Moras[‡], Leticia Gomes[‡], Raphael Matias[‡], Rafael Leandro de Assis[‡], Rafael Fraga[‡], Sara Miranda Almeida[‡], Vanessa Guimarães Lopes[‡], Rafaela Missagia[‡], Eduardo Costa Carvalho[‡], Nikolas Jorge Carneiro[‡], Ronnie Alves[‡], Pedro Martins Souza[‡], Guilherme Oliveira[§], Valeria Da Cunha Tavares[‡]

‡ Vale Institute of Technology, Belém, Brazil
§ Instituto Tecnológico Vale, Vale, Brazil

Corresponding author: Nubia Marques (nubia.marques@pq.itv.org)

Abstract

Biodiversity data, particularly species occurrence and abundance, are indispensable for testing empirical hypothesis in natural sciences. However, datasets built for research programs do not often meet FAIR (findable, accessible, interoperable, and reusable) principles, which raises questions about data quality, accuracy, and availability. The 21st century has markedly been a new era for data science and analytics, and every effort to aggregate, standardize, filter, and share biodiversity data from multiple sources have become increasingly necessary. In this study, we propose a framework for refining and conform secondary biodiversity data to FAIR standards to make them available for valuable use such as macroecological modeling and other studies. We relied on a Darwin Core base model to standardize and further facilitate the curation and validation of data related including the occurrence and abundance of multiple taxa of a region that encompasses estuarine ecosystems in an ecotonal area bordering the easternmost Amazonia. We further discuss the significance of feeding standardized public data repositories to advance scientific progress and highlight their role in contributing to the biodiversity management and conservation.

Keywords

Darwin Core standard, FAIR data, Golfão Maranhense, secondary data

Introduction

High-quality, openly available biodiversity datasets (e.g. species occurrence, abundance, traits) are indispensable for the monitoring of species and ecosystems, and to improve the development of conservation and management policies (Wetzel et al. 2015, Wetzel et al. 2018). Biodiversity data under FAIR (findable, accessible, interoperable, and reusable) principles also help optimizing editorial processes for academic publications accelerating peer review, increase the visibility of scientific papers, and improve citation rates (Costello et al. 2013, Piwowar and Vision 2013). Efforts towards the building and maintenance of repositories of FAIR data principles have been endeavored by collectors and curators of biodiversity data (Hackett et al. 2019)striving to organize, standardize and share data from a diverse array of primary (i.e. fieldwork) and secondary sources (e.g., museums, literature). This is particularly important for the contemporary big data science age that challenges our ability to organize, filter and analyze large and complex datasets (Cao 2016).

Locating data is the first step in making it readily accessible, reusable, and available for other studies. Researchers can use or reuse data that they do not create themselves, which are called secondary data (Allen 2017). Biodiversity data is not always publicly available in open access repositories and may be found in scientific publications, technical reports, or obtained under request by directly contacting the researcher who collected it (Costello et al. 2013). Therefore, it is often necessary to conduct systematic literature reviews to gather comprehensive databases focusing on biodiversity research programs. This approach is recommended as a rigorous way to search for relevant literature, allowing for peer replication and ensuring data validity and reliability (Xiao and Watson 2019). After locating the data, researchers need to retrieve and be able to understand the metadata associated and, if necessary, standardize them by reorganizing, renaming, and optimizing entry fields.

Standardization refers to a set of procedures and norms detailing how something must be accomplished (Berg 1997). The fundamental goal of structuring and standardizing biodiversity data is to make them understandable and usable by any researcher in the world to facilitate their continuous updating and their integration with other datasets, and sharing (Borregaard and Hart 2016). Sharing biodiversity data is essential for ecological studies, nature conservation and management, education, and for decision-making policies (Ganzevoort et al. 2017). Also, ecologists often use shared data for comparative studies, synthesis (e.g., meta-analysis), parameterization of models, and to test for reproducibility (Michener 2015). Data sharing can be done in many ways, ranging from private sharing upon request to depositing data on a public platform. Often, authors make their data available as supplementary material in scientific publications and post datasets on public websites. Data archiving is a late, indispensable step, enabling data reuse for further reanalyzes and syntheses to address new questions. Whitlock (2011) outlines optimal procedures for archiving ecological and evolutionary data, including selecting the appropriate and ensuring data and working towards metadata precision.

The availability of biodiversity data is influenced by a range of factors, including geographical region, scientific interests, and resource availability, including financial and infrastructure constraints, which can impact the quality and type of data produced (Amano et al. 2016). Global initiatives such as the "Global Biodiversity Information Facility-GBIF" provide access to comprehensive biodiversity datasets and facilitates their reuse. However, research programs from megadiverse regions such as the Neotropics face difficulties for retrieving, organizing, and providing quality data, because these are intrinsically complex including, as an example, unrecognized species and unresolved taxa complexes and taxonomy. One strategy to increase the usability and availability of secondary data, in this case, involves efforts towards the educated compilation of dispersed data from various sources, including scientific articles, and grey literature, such as technical reports, theses, and dissertations.

Focusing on increasing the applicability of biodiversity secondary data, we had three main goals in this study: (1) to build a scheme, a "pipeline" to enhance the usability of secondary data, namely locating the data, conducting quality control, standardizing the data, archiving and sharing it; (2) to test, through a study case our pipeline demonstrating a step by step management of secondary data following the FAIR principles; and (3) to evaluate if and how our approach can enhance our understanding of the dynamics of distribution and maintenance of the regional biodiversity, promote new scientific studies, insights, and above all improve our abilities to build scientific sound hypotheses.

We herein selected the Golfão Maranhense, an area located in the extreme north of the Amazon (Brazil) as our study model due to the combination of its richness, its ecological diversity and importance as an ecotonal mosaic between the Amazon Forest and dry ecosystems to the Eastern South America and, at the same time, to the scarceness of knowledge of the biology of this area. Although there are reports on the biodiversity of Golfão Maranhense they are presented in heterogeneous forms that include scientific articles and non-peer reviewed technical reports, making it challenging to understand the actual distribution of the biodiversity richness in the region. We ultimately predict that initiatives to collect and provide biodiversity data for reuse, as in the case of Golfão Maranhense can enhance knowledge and promote conservation efforts to safeguard these region's species, communities, and ecosystems.

Material and methods

The study was conducted in the Golfão Maranhense (Maranhão State, Brazil) including 13 municipalities in the surroundings (Fig. 1). The Golfão Maranhense is a vast estuarine complex located in eastern Amazonia (Brazil) and is formed by the São Marcos and São José bays separated by the island of São Luís. This region is an area of high ecological relevance known as the "Macromaré" Mangrove Coast of Amazonia where lies the largest continuous mangrove system in the world, with about 5,414 km of mangroves in northwestern Maranhão and 2,177 km in northeastern Pará (Souza Filho 2005). The climate is tropical humid, with well-defined dry (July to December) and rainy (January to June) seasons, and average temperatures around 26 ◦C. The area is characterized by semidiurnal macrotidal with average variations of 4 m and maximum of 7 m, with maximum tidal currents exceeding 4 m/s (Rebelo-Mochel 1997). São Marcos and São José Bays are port areas that hold significant importance for maritime activities, trade, and transportation within the region.

Finding Biodiversity Data - data repositories and systematic review

We conducted a systematic review to retrieve the biodiversity data of Golfão Maranhense performing searches in the platforms Science Direct and Google Scholar, public data repositories (GBIF, VertNet, Wikiaves, SpeciesLink) and technical reports built by consultants working in the region for environmental impact licensing, fauna and flora monitoring and other programs. To ensure transparency, completeness, and consistency in reporting systematic reviews and meta-analyses we used the “Preferred Reporting Items for Systematic Reviews and Meta-Analyses- PRISMA” guidelines. The PRISMA pipeline consists of a checklist and a flow diagram, making it easier for readers to understand the process and assess the reliability and validity of the findings.

We followed four steps:

1. *Biotic group selection*- We have chosen eight biotic groups that represent a substantial proportion of the terrestrial and aquatic biodiversity: mammals (Mammalia), “reptiles” including turtles, birds, lizards and snakes (Testudines, Squamata, Aves), amphibians (Amphibia), plants (Magnoliophyta), fishes, phytoplankton, and benthos.
2. *Keywords*- relevant keywords were defined considering each biotic group (Suppl. material 1).
3. *Inclusion criteria*- Our inclusion criteria were twofold: (a) studies that were conducted in Golfão Maranhense and; (b) studies that included both the geographic coordinates and the finest possible taxonomic level.
4. *Data selection* - We selected 81 variables to be extracted from the selected studies. These variables were classified into three main categories: (a) General Information **-** data about the published work, such as title, year, keywords, and objectives of the study; (b) Sampling Events **-** information about when and where the sampling of target taxa occurred, such as date, sampling method, location, and geographic coordinates; (c) Occurrences **-** description of the collected individual, such as epithet, life stage, and conservation status (according to IUCN and MMA). All variables are described in Suppl. material 2.

Biodiversity data quality and control

We conducted a validation process for the Golfão Maranhanse data that we optimized in three steps:

1. *Identifying and fixing errors*- we checked the data to identify any errors or inaccuracies. Errors include typos, wrong spelling of names, missing values,

inconsistencies, outliers, and other data anomalies. We cleaned the data, by removing duplicates and addressing inconsistencies in the data entries.

2. *Checking units and variables-* Data may involve variables with different units of measurement. We checked all the geographical coordinates and standardized them to decimal degrees and standardize the sampling dates to "Start Month" "Start Year" "End Month" and "End Year".
3. *Checking the records-* Species occurrence data are susceptible to misidentification and taxonomic changes make this a challenging, dynamic task. To ensure the reliability and validity of the species occurrence data, we reviewed the relevant literature for known geographic species distributions and compare them with the collected points. We also relied in our team of specialists in the taxa groups to meticulously check each entry for inconsistencies and up to date taxonomy. Mismatches between the known and collected geographic distributions indicate caution, a first alert and further investigation on that entry. Also, we reviewed the literature for any changes in synonymy and updated the occurrence record accordingly.

Standardizing biodiversity data

To standardize the data obtained from the Golfão Maranhense area, we used the Darwin Core standard (DwC) (Wieczorek et al. 2012) with adaptations (e.g. some column naming). DwC is one of the most widely used standards for biodiversity data used as a language for sharing biodiversity data that can be understood by human users and interpreted by computational systems. The DwC provides a straightforward, stable standard that simplifies the process of publishing biodiversity data, promoting the sharing, use, and reuse of openly accessible biodiversity data (Wieczorek et al. 2012). Also, DwC allows users to adapt terms that name the columns for various applications, including the checklists of species in an area (DoNascimiento et al. 2017).

Data sharing and archiving

The secondary data retrieved from the Golfão Maranhense area will be accessible through an online platform developed using PowerBI software. This platform is being developed and will be freely available, promoting the dissemination of knowledge related to the biodiversity of the Golfão Maranhense region.

Results

Of the 81 variables extracted from the studies, 59 were standardized using DwC terms and 13 were adapted due to the lack of appropriate terms for these variables within the current DwC models (Suppl. material 2). Considering all biotic groups, a total of 161 bibliographical references, including papers and technical reports were included in the systematic review of the literature (Fig. 2). In addition, we included species occurrence records from four public repositories (GBIF, VertNet, Wikiaves, SpeciesLink). Considering only published

Author-formatted, not peer-reviewed document posted on 23/04/2024. DOI: https://doi.org/10.3897/arphapreprints.e125974

papers, the group included in the largest number of published papers and reports was plants (N = 59), and the group with the least data sources was benthos (N = 11) (Suppl. material 3 "Preferred Reporting Items for Systematic Reviews and Meta-Analyses – PRISMA", separated by groups).

Taxa occurrence data

A total of 2,070 occurrence events were obtained from bibliographic references and 43,947 were obtained by public repositories (n = 46,017) from 3,871 taxa. These include birds (Aves, 458 species; 3 other taxonomic level), amphibians (Amphibia, 55 species; 9 to the genus level), reptiles (2 Crocodylia; 86 Squamata; 11 Testudines); mammals (Class Mammalia; 101 species; 21 to the genus level), fish (268 species, 74 other taxonomic level), phytoplankton (370 species; 105 other taxonomic level), benthos (188 species; 204 other taxonomic level), and plants (1,624 species; 292 other taxonomic level). (Suppl. material 4). Most of the taxa were identified to species (81%) and genus (14%) level (Fig. 3). Benthos accounted for the highest number of occurrence events, with 12,510 records, and non-bird reptiles had with the lowest number of occurrences events recorded (570).

Data were carefully analyzed by specialists in each group to check for inconsistencies in identification, spelling, and, as much as possible, potentiality of identification correctness (e.g. check if the geographic locations were within expected known geographic distribution for each taxon, checking vouchers when possible). A total of 93 occurrence events were deleted, including 92 from taxa that were not correctly identified (76 birds and 16 mammals) and one bird specimen that was a victim of animal trafficking.

The number of occurrences was dependent on the number of taxa in each sampled group (R^2 = 0.47, P = 0.03). While amphibians and non-bird reptiles were represented by low numbers of both taxa and occurrences, plants, birds, and phytoplankton were highly representated for both occurrences and richness. On the other hand, the group "benthos" had a high number of occurrences, and a low number of taxa (Fig. 4).

Discussion

We proposed a workflow to improve our abilities of recovering biodiversity data of better quality while using secondary data sources based on databases compiled from a megadiverse ecotonal estuarine region in the easternmost border of Amazonia, the Golfão Maranhense. We were able to extract a large amount of information about the biodiversity of the Golfão Maranhense and transform this unrelated data into organized and re-usable data. This systematic approach ensured data accuracy and reliability, facilitating the potential reuse of information in future studies. A step further that we started advancing for some groups is the systematic survey in museum collections and analyzes focusing on relevant questions we have found along the way (e.g. general patterns of occurrence of migratory birds, sampling bias and gaps for many groups and so on).

Researchers can use existing datasets, such as those obtained through our biodiversity data retrieval method, to conduct a wide range of studies for advancing scientific research. Secondary data can be used, for example, to do meta-analyses (e.g. *Biggs et al. 2020*) for comparative studies across different geographic regions and time, to support ecological modeling applied to species distributions (e.g. Fletcher Jr. et al. 2019), habitat preferences, and potential impacts of environmental changes (Bayraktarov et al. 2019) as long as it is used judiciously. However, finding high-quality secondary data can be challenging, as showed in a recent survey that most researchers reported that data discovery can be arduous (73%) or difficult (19%) (Gregory et al. 2020). Several initiatives have been made to collect, standardize, store, and make biodiversity data openly available (Costello et al. 2013). For example, international repositories, such as “GBIF” (Global Biodiversity Information Facility 2022), and “Freshwater Biodiversity Data Portal- BioFresh” (http://data.freshwaterbiodiversity.eu/) (for other repositories see Culina et al. 2018).

Long-term monitoring datasets can help understand patterns and changes in ecological variables over time (Costa and Magnusson 2010, Magnusson et al. 2021). This can help identify ecological shifts and potential drivers of biodiversity change (e.g. Marques et al. 2022). We are currently using our database to expand our knowledge on species of mammals, birds, fishes, amphibians, non-bird reptiles, marine phytoplankton and benthos in the Golfão Maranhense. We are also exploring the patterns and determinants of floristic variation in the region and the temporal variation of migratory birds in the São Marcos Bay region. Whereas data retrieved in standardized monitoring programs such as LTER (Vanderbilt et al. 2015, Kaplan et al. 2021) can be directly linked to FAIR standardized data repositories, other secondary data that are not may be important, rescued, treated and used, given that they may be previously thoroughly revised and curated and kept with some rule of error estimates to build robust hypotheses to investigate and to understand biodiversity patterns.

While secondary data can be a valuable resource for scientific research, it is crucial to recognize and address its limitations, and ideally estimate the errors within. Common challenges include species identification accuracy, geographic coordinate precision, and data entry errors. In addition, datasets from different studies may differ in their sampling methods, data structure, and definitions of key variables, making direct comparisons difficult. Finally, some datasets may not be openly accessible, which has implications for data availability and complicates data access and sharing policies.

Other limitations are the sampling and temporal biases, which can arise when working with secondary data, making data interpretation more challenging. Sampling bias occurs when the data sampling disproportionately favors certain species or areas over others. For example, the concentration of specimen records in more easily accessible sites, such as major cities, roads, and navigable rivers (Boakes et al. 2010). Also, logistics and human interference are factors that can explain research probability (e.g. 64% of research probability in Amazon; Carvalho et al. 2023). Temporal bias, on the other hand, refers to the uneven distribution of data across time periods. Secondary data sources may include data collected over different time spans, reflecting historical variations in research focus, funding availability, or changes in data recording practices. Consequently, certain time periods may be overrepresented, while others may be sparsely covered or entirely absent. Additionally, the difficulty of conducting research in regions with limited accessibility introduces challenges that restrict the ability to gather data from remote areas. Thus, remote regions potentially hosting unique biodiversity hotspots are often underrepresented, or completely absent from the dataset.

In our study, sampling bias is evident in the São Marcos Bay area, where an industrial ship port is located. Many technical survey reports were produced in this area due to the need for port companies to carry out mandatory environmental licensing processes. These reports conducted in port area inherently prioritize certain species and ecological aspects more relevant to the licensing process, overlooking other important components of biodiversity. Within our database, it becomes apparent that some species records originate from technical reports not easily available. For example, we found 365 species and varieties of phytoplankton in technical reports, but 101 were not previously catalogued on the Brazilian Biodiversity Platform REFLORA (REFLORA 2013) for the Maranhão region. This underscores the fact that the retrieval of biodiversity data can yield enhancements in the comprehension of species composition existing within the defined geographical area.

Conclusions

The workflow that we employed has facilitated the retrieval of biodiversity data from the ecologically rich and megadiverse Golfão Maranhense region in Maranhão, Brazil. By combining a systematic review approach with standardized worksheets with a Darwin Core base, we were able to effectively search and explore a wide range of scientific articles, technical reports, and specialized public repositories. The potential use of secondary data for the advancement of scientific research is significant although it must be taken with care and analyzed with the lens of precaution observing all bias limitation and filters involved. Many technical survey reports were produced in the Golfão Maranhense to carry out mandatory environmental licensing for the port and surrounding activities. By using existing datasets, researchers can carry out a wide range of activities which include meta-analyses, comparative studies, ecological modelling, and most of all, building sound hypotheses and produce sound experiment designs to monitor diversity in a standardized based. Our study highlights the value of systematic review methods, and the need for an approach to address data limitations and biases. Likewise, this method can facilitate collaboration among researchers, enable comparative analyses across different datasets, and support evidence-based conservation strategies and policymaking.

Acknowledgements

We are grateful to Tereza Cristina Giannini, Leonardo Trevelín, and André Luis Acosta for their valuable contributions with the Darwin Core standard.

Hosting institution

Vale Institute of Technology

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Allen M (2017) Secondary data. SAGE Encyclopedia of Communication Research Methods1578-1579. https://doi.org/10.4135/9781483381411.n557
- Amano T, Lamming JL, Sutherland W (2016) Spatial gaps in global biodiversity information and the role of citizen science. BioScience 66 (5): 393-400. https://doi.org/10.1093/biosci/biw022
- Bayraktarov E, Ehmke G, O'Connor J, Burns E, Nguyen H, McRae L, Possingham H, Lindenmayer D (2019) Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution 6 https://doi.org/10.3389/fevo.2018.00239
- Berg M (1997) Problems and promises of the protocol. Social Science & Medicine 44 (8): 1081-1088. https://doi.org/10.1016/S0277-9536(96)00235-3
- Biggs C, Yeager L, Bolser D, Bonsell C, Dichiera A, Hou Z, Keyser S, Khursigara A, Lu K, Muth A, Negrete Jr. B, Erisman B (2020) Does functional redundancy affect ecological stability and resilience? A review and meta-analysis. Ecosphere 11 (7). https://doi.org/10.1002/ecs2.3184
- Boakes E, McGowan PK, Fuller R, Chang-qing D, Clark N, O'Connor K, Mace G (2010) Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLOS Biology 8 (6). https://doi.org/10.1371/journal.pbio.1000385
- Borregaard MK, Hart E (2016) Towards a more reproducible ecology. Ecography 39 (4): 349-353. https://doi.org/10.1111/ecog.02493
- Cao L (2016) Data science and analytics: a new era. International Journal of Data Science and Analytics 1 (1): 1-2. https://doi.org/10.1007/s41060-016-0006-1
- Carvalho R, Resende A, Barlow J, França F, Moura M, Maciel R, Alves-Martins F, Shutt J, Nunes C, Elias F, Silveira J, Stegmann L, Baccaro F, Juen L, Schietti J, Aragão L, Berenguer E, Castello L, Costa FC, Guedes M, Leal C, Lees A, Isaac V, Nascimento R, Phillips O, Schmidt FA, Ter Steege H, Vaz-de-Mello F, Venticinque E, Vieira ICG, Zuanon J, Ferreira J, Carvalho R, Resende A, Barlow J, França F, Moura M, Maciel R, Alves-Martins F, Shutt J, Nunes C, Elias F, Silveira J, Stegmann L, Baccaro F, Juen L, Schietti J, Aragão L, Berenguer E, Castello L, Costa FC, Guedes M, Leal C, Lees A, Isaac V, Nascimento R, Phillips O, Schmidt FA, Ter Steege H, Vaz-de-Mello F, Venticinque E, Vieira ICG, Zuanon J, Ferreira J, Geber Filho ANDS, Ruschel A, Calor AR, De Lima Alves A, Muelbert AE, Quaresma A, Vicentini A, Piedade ARD, Oliveira AAD, Aleixo A, Casadei-Ferreira A, Gontijo A, Hercos A, Andriolo A, Lopes A, Pontes-Lopes A, Santos APMD, Oliveira ABDSD, Mortati AF, Salcedo AKM, Albernaz AL, Fares AL, Andrade AL, Oliveira Pes AM, Faria APJ, Batista APB, Puker A, Bueno A, Junqueira

AB, Holanda De Andrade ALR, Ghidini AR, Galuch A, Menezes ASOD, Manzatto AG, Correa AS, Queiroz AM, Zanzini ACDS, Olivo Neto AM, Melo AWFD, Guimaraes AF, Castro AB, Borges A, Ferreira AB, Marimon B, Marimon-Junior BH, Flores B, De Resende BO, Albuquerque BW, Villa B, Davis B, Nelson B, Williamson B, Melo BSBD, Cintra BL, Santos BB, Prudente BDS, Luize BG, Godoy BS, Rutt C, Duarte Ritter C, Silva CJ, Ribas CR, Peres C, Azevêdo CASD, Freitas C, Cordeiro CL, Brocardo CR, Castilho C, Levis C, Doria CRDC, Arantes C, Santos CAD, Jakovac C, Silva CA, Benetti CJ, Lasmar C, Marsh C, Andretti CB, Oliveira CPD, Cornelius C, Alves Da Rosa C, Baider C, Gualberto C, Deus CPD, Monteiro Jr. CDS, Santos Neto CRD, Lobato CMC, Santos CRMD, Penagos CCM, Costa DDS, Vieira DLM, Aguiar DPPD, Veras DS, Pauletto D, Braga DDL, Storck-Tonon D, Almeida DDF, Douglas D, Amaral DDD, Gris D, Luther D, Edwards D, Guimarães DP, Santos DCD, Campana DRDS, Nogueira DS, Silva DRD, Dutra DBDS, Rosa DCP, Silva DASD, Pedroza D, Anjos D, Melo Lima DV, Silvério D, Rodrigues DDJ, Bastos D, Daly D, Barbosa EM, Arenas ERC, Oliveira EAD, Santos EAD, Santana ECCD, Guilherme E, Vidal E, Campos-Filho EM, Van Den Berg E, Morato EF, Da Silva E, Marques E, Pringle E, Nichols E, Andresen E, Farias EDS, Siqueira ELSD, De Albuquerque EZ, Görgens EB, Cunha EJRD, Householder E, Novo EMMDL, Oliveira FFD, Roque FDO, Coletti F, Reis F, Moreira FF, Todeschini F, Carvalho FA, Coelho De Souza F, Silva FAB, Carvalho FG, Cabeceira FG, d'Horta FM, Mendonça F, Florêncio FP, Carvalho FRD, Arruda FVD, Nonato FADS, Santana FD, Durgante F, Souza FKSD, Obermuller FA, Castro FSD, Wittmann F, Sales FMDS, Neto FV, Salles FF, Borba GC, Damasco G, Barros GG, Brejão GL, Jardim GA, Prance G, Lima GR, Desidério GR, Melo GDCD, Carmo GHPD, Cabral GS, Rousseau GX, Da Silva GC, Schwartz G, Griffiths H, Queiroz HLD, Espírito-Santo HV, Cabette HSR, Nascimento HEM, Vasconcelos H, Medeiros H, Aguiar HJACD, Leão H, Wilker I, Gonçalves IC, De Sousa Gorayeb I, Miranda IPDA, Brown IF, Santos ICS, Fernandes IO, Fernandes I, Delabie JHC, De Abreu JC, Gama Neto JDL, Costa JBP, Noronha JC, De Brito JG, Wolfe J, Santos JC, Ferreira-Ferreira J, E Gomes JO, Lasky J, De Faria Falcão JC, Costa JG, Cravo JS, Guerrero JEB, Muñoz Gutiérrez JA, Carreiras J, Lanna J, Silva Brito J, Schöngart J, Mendes Aguiar JJ, Lima J, Barroso J, Noriega JA, Pereira JLDS, Nessimian JL, Souza JLPD, De Toledo JJ, Magalhães JLL, Camargo JL, Oliveira J, Ribeiro JMF, Silva JODA, Da Silva Guimarães JR, Hawes J, Andrade-Silva J, Revilla JDC, Da Silva JS, Da Silva Menger J, Rechetelo J, Stropp J, Barbosa JF, Do Vale JD, Louzada J, Cerqueira Silva JC, Da Silva KD, Melgaço K, Carvalho KS, Yamamoto KC, Mendes KR, Vulinec K, Maia LF, Cavalheiro L, Vedovato LB, Demarchi LO, Giacomin L, Dumas LL, Maracahipes L, Brasil LS, Ferreira LV, Calvão LB, Maracahipes-Santos L, Reis LP, Da Silva LF, De Oliveira Melo L, Carvalho LCDS, Casatti L, Amado LL, De Matos LS, Vieira L, Prado LPD, Alencar L, Fontenele L, Mazzei L, Navarro Paolucci L, Zanzini LP, Carvalho LN, Crema LC, Brulinger LFB, Montag LFDA, Naka LN, Azara L, Silveira LF, Nunes LGDO, Rosalino LMDC, Mestre LM, Bonates LCDM, Coelho LDS, Borges LHM, Lourenço LDS, Freitas MAB, Brito MTDS, Pombo MM, Da Rocha M, Cardoso MR, Guedes MC, Raseira MB, Medeiros MBD, Carim MDJV, Simon MF, Pansonato MP, Dos Anjos MR, Nascimento MT, Souza MRD, Monteiro MGT, Da Silva MJ, Uehara-Prado M, Oliveira MAD, Callisto M, Vital MJS, O Santos MPD, Silveira M, Oliveira M, Pérez-Mayorga MA, Carniello MA, Lopes MA, Silveira MAPDA, Esposito MC, Maldaner ME, Passos M, Anacléto MJP, Costa MKS, Martins MP, Piedade MTF, Irume MV, Costa MMSD, Maximiano MFDA, Freitas MG, Cochrane M, Gastauer M,

Almeida MRN, Souza MFD, Catarino M, Costa Batista M, Massam M, Martins MFDO, Holmgren M, Almeida M, Dias M, Espírito Santo NB, Benone NL, Ivanauskas NM, Medeiros N, Targhetta N, Félix NS, Ferreira N, Hamada N, Campos N, Giehl NFDS, Metcalf OC, Silva OGMD, Cerqueira PV, Moser P, Miranda PN, Peruquetti PSF, Alverga PPDP, Prist P, Souto P, Brando P, Pompeu PDS, Barni PE, Graça PMDA, Morandi P, Cruz PV, Da Silva PG, Bispo P, Camargo PBD, Sarmento PDM, Souza P, Andrade RBD, Braga RB, Boldrini R, Bastos RC, Assis RLD, Salomão R, Leitão RP, Mendes RG, Carpanedo RDS, Melinski RD, Ligeiro R, E Pérez REP, Barbosa RI, Cajaiba RL, Silvano RAM, Salomão RP, Hilário RR, Martins RT, Perdiz RDO, Vicente RE, Silva RJD, Koroiva R, Solar R, Silva RDC, S De Lima RB, Silva RDSAD, Mariano R, Ribeiro RAB, Fadini RF, Oliveira RLCD, Feitosa RM, Matavelli R, Mormul RP, Da Silva RR, Zanetti R, Barthem R, Almeida RPS, Ribeiro SC, R Costa Neto SVD, Nienow S, Oliveira SAVD, Borges SH, Milheiras S, Ribeiro SP, Couceiro SRM, Sousa SAD, Rodrigues SB, Dutra SL, Mahood S, Vieira SA, Arrolho S, Silva SSD, Triana SP, Laurance S, Kunz SH, Alvarado S, Rodrigues THA, Santos TFD, Machado TLDS, Feldpausch T, Sousa T, Michelan TS, Emilio T, Brito TDF, André T, Barbosa TAP, Miguel TB, Izzo TJ, Laranjeiras TO, Mendes TP, Silva TSF, Krolow TK, Begot TO, Baker T, Domingues T, Giarrizzo T, Bentos TV, Haugaasen T, Peixoto U, Pozzobom UM, Korasaki V, Ribeiro VS, Scudeller VV, Oliveira VHF, Landeiro VL, Santos Ferreira VR, Silva VDNG, Gomes VHF, Oliveira VCD, Firmino V, Santiago WTV, Beiroz W, Almeida WRD, Oliveira WLD, Silva WCD, Castro W, Dáttilo W, Cruz WJAD, Silva WFMD, Magnusson W, Laurance W, Milliken W, Paula WSD, Malhi Y, Shimabukuro YE, Lima YGD, Shimano Y, Feitosa Y (2023) Pervasive gaps in Amazonian ecological research. Current Biology 33 (16). https://doi.org/10.1016/j.cub.2023.06.077

- Costa FRC, Magnusson WE (2010) The need for large-scale, integrated studies of biodiversity - the experience of the program for biodiversity research in Brazilian Amazonia. Natureza & Conservação 08 (01): 3-12. https://doi.org/10.4322/natcon.00801001
- Costello M, Michener W, Gahegan M, Zhang Z, Bourne P (2013) Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution 28 https://doi.org/10.1016/j.tree.2013.05.002
- Culina A, Baglioni M, Crowther T, Visser M, Woutersen-Windhouwer S, Manghi P (2018) Navigating the unfolding open data landscape in ecology and evolution. Nature Ecology & Evolution 2 (3): 420-426. https://doi.org/10.1038/s41559-017-0458-2
- DoNascimiento C, Herrera-Collazos EE, Herrera-R. G, Ortega-Lara A, Villa-Navarro F, Oviedo JSU, Maldonado-Ocampo J (2017) Checklist of the freshwater fishes of Colombia: a Darwin Core alternative to the updating problem. ZooKeys25-138. https://doi.org/10.3897/zookeys.708.13897
- Fletcher Jr. R, Hefley T, Robertson E, Zuckerberg B, McCleery R, Dorazio R (2019) A practical guide for combining data to model species distributions. Ecology 100 (6). https://doi.org/10.1002/ecy.2710
- Ganzevoort W, van den Born RG, Halffman W, Turnhout S (2017) Sharing biodiversity data: citizen scientists’ concerns and motivations. Biodiversity and Conservation 26 (12): 2821-2837. https://doi.org/10.1007/s10531-017-1391-z
- Global Biodiversity Information Facility (2022) What is GBIF? URL: https://www.gbif.org/what-is-gbif

- Gregory K, Groth P, Scharnhorst A, Wyatt S (2020) Lost or found? discovering data needed for research: Supplementary materials. Harvard Data Science Review https://doi.org/10.1162/99608f92.e38165eb
- Hackett R, Belitz M, Gilbert E, Monfils A (2019) A data management workflow of biodiversity data from the field to data users. Applications in Plant Sciences 7 (12). https://doi.org/10.1002/aps3.11310
- Kaplan N, Baker K, Karasti H (2021) Long live the data! Embedded data management at a long-term ecological research site. Ecosphere 12 (5). https://doi.org/10.1002/ecs2.3493
- Magnusson WE, Lima AP, Aragón S, Rosa CAd, Brocardo CR, Fadini R (2021) Long-term standardized ecological research in an amazonian savanna: A laboratory under threat. Volume 93, Número e20210879. URL: https://repositorio.inpa.gov.br/handle/1/38327
- Marques NC, Machado R, Aguiar LM, Mendonca-Galvão L, Tidon R, Vieira E, Marini-Filho O, Bustamante M (2022) Drivers of change in tropical protected areas: Long-term monitoring of a Brazilian biodiversity hotspot. Perspectives in Ecology and Conservation 20 (2): 69-78. https://doi.org/10.1016/j.pecon.2022.02.001
- Michener W (2015) Ecological data sharing. Ecological informatics 29: 33-44. https://doi.org/10.1016/j.ecoinf.2015.06.010
- Piwowar H, Vision T (2013) Data reuse and the open data citation advantage. PeerJ 1 URL: https://peerj.com/articles/175/?report=reader
- Rebelo-Mochel F (1997) Mangroves on São Luís Island, Maranhão, Brazil. In: Kjerve B, Lacerda LD, Diop ES (Eds) Mangrove ecosystem studies in Latin America and Africa. UNESCO, Paris, 145-154 pp.
- REFLORA (2013) Flora e Funga do Brasil. https://floradobrasil.jbrj.gov.br/reflora/PrincipalUC/PrincipalUC.do. Accessed on: 2024-4-19.
- Souza Filho PWM (2005) Costa de manguezais de macromaré da Amazônia: cenários morfológicos, mapeamento e quantificação de áreas usando dados de sensores remotos. Revista Brasileira de Geofísica 23: 427-435. https://doi.org/10.1590/S0102-261X2005000400006
- Vanderbilt K, Lin C, Lu S, Kassim AR, He H, Guo X, Gil IS, Blankman D, Porter J (2015) Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network. Ecosphere 6 (10): 1-18. https://doi.org/10.1890/ES14-00281.1
- Wetzel F, Saarenmaa H, Regan E, Martin C, Mergen P, Smirnova L, Tuama ÉÓ, García Camacho F, Hoffmann A, Vohland K, Häuser C (2015) The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case study. Biodiversity 16 (2-3): 137-149. https://doi.org/10.1080/14888386.2015.1075902
- Wetzel F, Bingham H, Groom Q, Haase P, Kõljalg U, Kuhlmann M, Martin C, Penev L, Robertson T, Saarenmaa H (2018) Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. Biological Conservation 221: 78-85. https://doi.org/10.1016/j.biocon.2017.12.024
- Whitlock M (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution 26 (2): 61-65. https://doi.org/10.1016/j.tree.2010.11.006
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data

standard. PLOS One 7 (1). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029715

- Xiao Y, Watson M (2019) Guidance on conducting a systematic literature review. Journal of Planning Education and Research 39 (1): 93-112. https://doi.org/10.1177/0739456X17723971

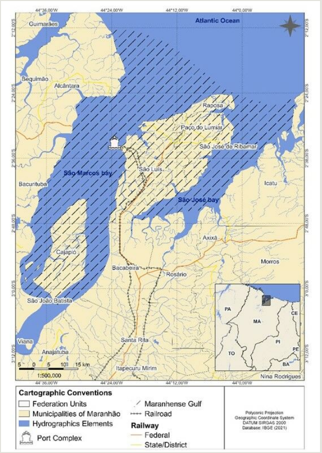

Figure 1.

Map showing the Golfão Maranhense area, estuarine region of eastern Amazon, Brazil.

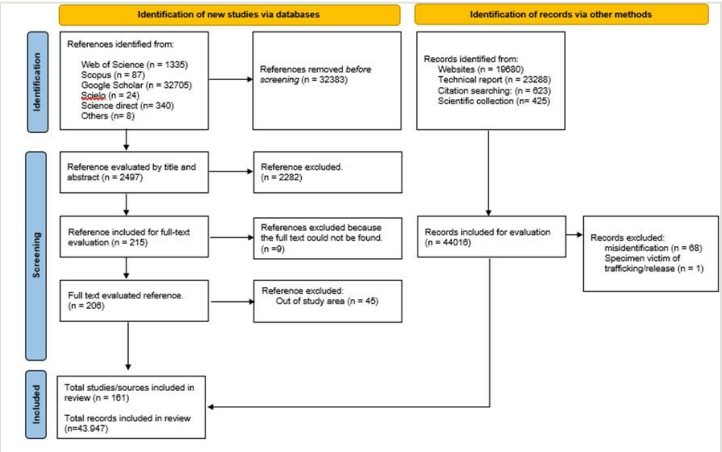
Identification of new studies via databases
Identification of records via other methods
Identification
Screening
Included
References identified from:
Web of Science (n = 1335)
Scopus (n = 87)
Google Scholar (n = 32705)
Scielo (n = 24)
Science direct (n= 340)
Others (n= 8)
References removed *before screening* (n = 32383)
Records identified from:
Websites (n = 19680)
Technical report (n = 23288)
Citation searching: (n = 623)
Scientific collection (n= 425)
Reference evaluated by title and abstract (n = 2497)
Reference excluded. (n = 2282)
Reference included for full-text evaluation (n = 215)
References excluded because the full text could not be found. (n =9)
Records included for evaluation (n = 44016)
Records excluded:
misidentification (n = 68)
Specimen victim of trafficking/release (n = 1)
Full text evaluated reference. (n = 206)
Reference excluded:
Out of study area (n = 45)
Total studies/sources included in review (n = 161)
Total records included in review (n=43.947)

Figure 2.

Flowchart of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for all groups showing the process of selecting studies throughout systematic review**.** The selection process includes three stages: (1) identifying the database and choosing the papers; (2) scanning the references and selecting the papers to be included; (3) including the selected papers.

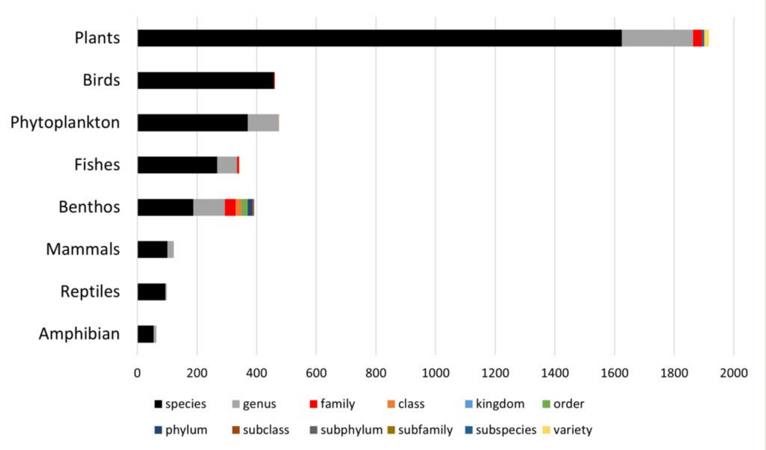

Figure 3.

Proportion of each taxonomic level identified for each biological group in the secondary data recovered from the Golfão Maranhense area.

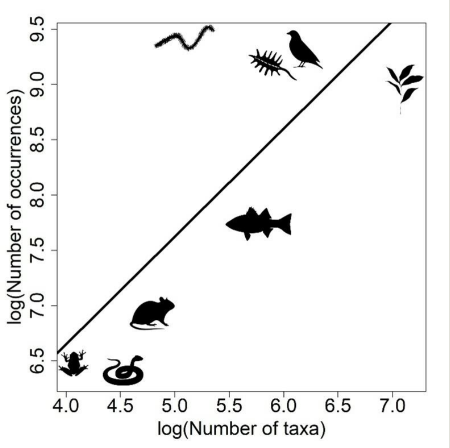

Figure 4.

Relationships between numbers of taxa and occurrences gathered through an extensive review of secondary biodiversity data from the Golfão Maranhense area, in the estuarine regions of eastern Amazonia.

Supplementary materials

Suppl. material 1: Keywords

Authors: Nubia Marques
Data type: Table
Brief description: Keywords used in the systematic review of each biotic group
Download file (16.09 kb)

Suppl. material 2: Table Darwin Core (DwC)

Authors: Nubia Marques
Data type: Table
Brief description: Table containing the Darwin Core (DwC) standard terms that were used to make the table and extract the information from the bibliographic references previously selected in the systematic review. Label= name of the column in the DwC standard; Definition= Brief definition of what each column means.
Download file (28.87 kb)

Suppl. material 3: Flowchart of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)

Authors: Nubia Marques
Data type: Images
Brief description: Flowchart of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) separated by groups showing the process of selecting studies throughout the systematic review. The selection process includes three stages: (1) identifying the database and choosing the papers; (2) scanning the references and selecting the papers to be included; (3) including the selected papers.
Download file (1.87 MB)

Suppl. material 4: List of species from the Golfão Maranhense (Maranhão State, Brazil)

Authors: Nubia Marques
Data type: Table
Brief description: List of species from the Golfão Maranhense (Maranhão State, Brazil) that were retrieved through the systematic literature review.
Download file (395.77 kb)