Biodiversity Data Journal 12: e133775
doi: 10.3897/BDJ.12.e133775

Research Article

Retrieving biodiversity data from multiple sources: making secondary data standardised 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[‡], Letícia Gomes da Silva[‡], Raphael Matias[‡,|], Rafael Leandro de Assis[‡], Rafael Fraga[‡], Sara Miranda Almeida[‡], Vanessa Guimarães Lopes[‡], Verônica Oliveira[‡], Rafaela Missagia[‡], Eduardo Costa Carvalho[‡], Nikolas Jorge Carneiro[‡], Ronnie Alves[‡], Pedro Souza-Filho[‡,¶], Guilherme Oliveira[‡], Margarida Miranda[‡], Valéria da Cunha Tavares[‡,#,¤]

‡ Vale Institute of Technology, Belém, Brazil
§ Universidade Estadual do Maranhão, São Luís, Brazil
| Federal University of Jataí, Jataí, Brazil
¶ Instituto de Geociências, Universidade Federal do Pará, Pará, Brazil
# Museu Paraense Emílio Goeldi, MPEG, Pós-graduação em Biodiversidade e Evolução, Belém, Brazil
¤ Pós-Graduação em Zoologia & Laboratório de Mamíferos, Departamento de Sistemática e Ecologia, Universidade Federal da Paraíba, João Pessoa, Brazil

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

Academic editor: Paulo Borges

Received: 02 Aug 2024 | Accepted: 04 Sep 2024 | Published: 20 Sep 2024

Citation: Marques N, Soares CDdeM, Casali DdeM, Guimarães E, Fava F, Abreu JMdaS, Moras L, Silva LGda, Matias R, Assis RLde, Fraga R, Almeida S, Lopes V, Oliveira V, Missagia R, Carvalho E, Carneiro N, Alves R, Souza-Filho P, Oliveira G, Miranda M, Tavares VdaC (2024) Retrieving biodiversity data from multiple sources: making secondary data standardised and accessible. Biodiversity Data Journal 12: e133775. https://doi.org/10.3897/BDJ.12.e133775

Abstract

Biodiversity data, particularly species occurrence and abundance, are indispensable for testing empirical hypothesis in natural sciences. However, datasets built for research programmes 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, standardise, filter and share biodiversity data from multiple sources have

become increasingly necessary. In this study, we propose a framework for refining and conforming secondary biodiversity data to FAIR standards to make them available for use such as macroecological modelling and other studies. We relied on a Darwin Core base model to standardise 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 standardised 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 optimising editorial processes for academic publications accelerating peer review, increasing the visibility of scientific papers and improving citation rates (Costello et al. 2013, Piwowar and Vision 2013). Efforts to build and maintain repositories of FAIR data principles have been undertaken by biodiversity data collectors and curators (Hackett et al. 2019), who strive to organise, standardise and share data from a variety of primary (e.g. fieldwork) and secondary (e.g. literature) sources, for example, global initiatives, such as the "Global Biodiversity Information Facility - GBIF" (https://www.gbif.org/), which provide access to comprehensive biodiversity datasets and facilitate their reuse. This is particularly important in the modern era of big data science, which challenges our ability to organise, filter and analyse large and complex datasets (Cao 2016).

The availability of biodiversity data is influenced by several factors, including geographic region, scientific interest and resource availability (e.g. financial and infrastructure constraints), which can affect the quality and type of data produced (Amano et al. 2016). In addition, biodiversity data are not always publicly available, requiring users to conduct extensive searches in scientific publications, technical reports or by directly contacting the researcher who collected the data (Costello et al. 2013). Therefore, conducting systematic literature reviews is often necessary to compile comprehensive databases focusing on biodiversity research programmes. This method is recommended as it provides a rigorous way to search for relevant literature, allowing for peer replication and ensuring data validity and reliability (Xiao and Watson 2019). However, the retrieved data may be in inconsistent and sometimes confusing formats, requiring additional effort from users, including finding separate metadata files, reorganising and renaming fields,

integrating aggregated data and discarding poorly documented or questionable data (Parr et al. 2012, Hunt et al. 2015, Culley 2017). The fundamental goal of the structure and standardisation of biodiversity data is to enable them to be understood and used by anyone and to be continuously updated and integrated with other datasets (Borregaard and Hart 2016). There are several guidelines and standards for biodiversity data, including the Darwin Core (DwC; Wieczorek et al. (2012); Darwin Core Maintenance Group (2014) and the DMPTool (https://dmptool.org/).

Finally, data must be shared and archived to ensure findability, accessibility and reusability. Data sharing can occur in a variety of ways, ranging from private sharing on 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. Sharing biodiversity data is essential for ecological research, conservation and management, education and policy decision-making (Ganzevoort et al. 2017). In addition, ecologists often use shared data for comparative studies, syntheses (e.g. meta-analyses), model parameterisation and reproducibility testing (Michener 2015). Data archiving is a critical final step that allows data to be reused for further analyses and syntheses to address new questions. Whitlock (2011) outlines optimal procedures for archiving ecological and evolutionary data, including selecting the appropriate repository and ensuring the accuracy of data and metadata. Good archiving practices facilitate long-term preservation of data, making them accessible for future research and applications.

Research programmes from megadiverse regions, such as the Neotropics, face difficulties in retrieving, organising and providing quality data due to their intrinsic complexity, which includes unrecognised species and unresolved taxa complexes and taxonomy. To address these challenges and to improve the reusability of secondary biodiversity data, our study had three main objectives:

1. to create a scheme or "pipeline" to improve the usability of secondary data by locating the data, performing quality control, standardising the data and archiving and sharing it;
2. to test our pipeline through a case study, demonstrating the step-by-step management of secondary data according to the FAIR principles; and
3. to evaluate if and how our approach can improve our understanding of the dynamics of regional biodiversity distribution and conservation, promote new scientific studies and knowledge and enhance our ability to generate hypotheses.

Retrieving biodiversity data is not an easy task, as the use of systematic literature searches alone does not guarantee the quality of the data. Additionally, commonly used pipelines for retrieving, standardising and making secondary data available typically overlook grey literature, despite its potential for biodiversity studies. The proposed pipeline is a combination of data retrieval and data management tools that are typically used separately, such as systematic review and the Darwin Core Standard. Following this Biodiversity Data Retrieval Pipeline ensures that secondary data are cleaned, normalised, shared and archived according to the FAIR Data Principles. Additionally, we discuss the challenges of efficient data retrieval, the potential reuse of secondary data in

future studies and its limitations. Finally, we predict that initiatives to collect biodiversity data and make it available for reuse can improve knowledge and advance conservation efforts to protect the species, communities and ecosystems of these regions.

Material and methods

The Biodiversity Data Retrieval Pipeline was built following four stages (Fig. 1): (1) Search; (2) Validate; (3) Standardise; and (4) Share and archive. All the steps are described below.

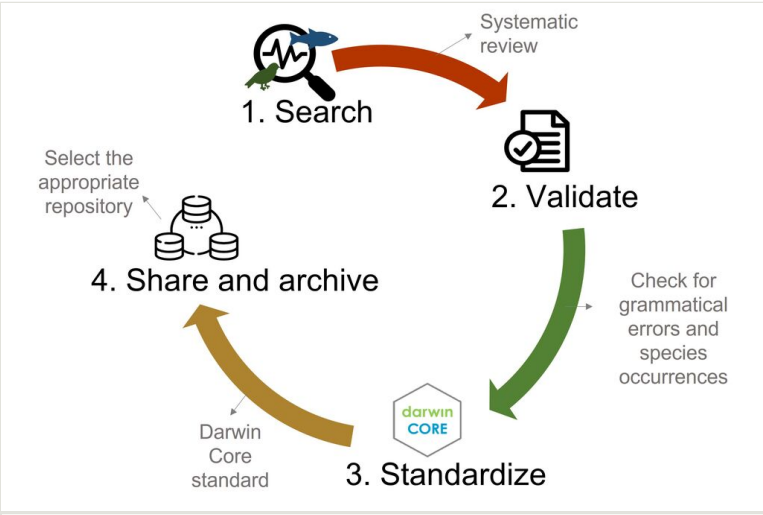

Figure 1. doi

Step-by-step guide for the proposed Biodiversity Data Retrieval Pipeline to retrieve secondary biodiversity data from various sources (e.g. scientific articles, technical reports, theses, dissertations, databases) according to the FAIR principles (findable, accessible, interoperable and reusable).

To test the pipeline, we selected the *Golfão Maranhense*, a region located in the extreme north of the Amazon (Brazil), due to its richness, ecological diversity and importance as an ecotonal mosaic between the Amazon Forest and the dry ecosystems of eastern South America and because of the scarcity of knowledge about the biodiversity in the region. Although reports on biodiversity in the region exist, they are presented in heterogeneous forms, including scientific articles and non-peer-reviewed technical reports, making it difficult to understand the true distribution of biodiversity richness in the region.

Study area

The study was conducted in the *Golfão Maranhense* (Maranhão State, Brazil) including 13 municipalities in the surroundings (Fig. 2). 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 north-western Maranhão and 2,177 km in north-eastern 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 characterised by semi-diurnal 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.

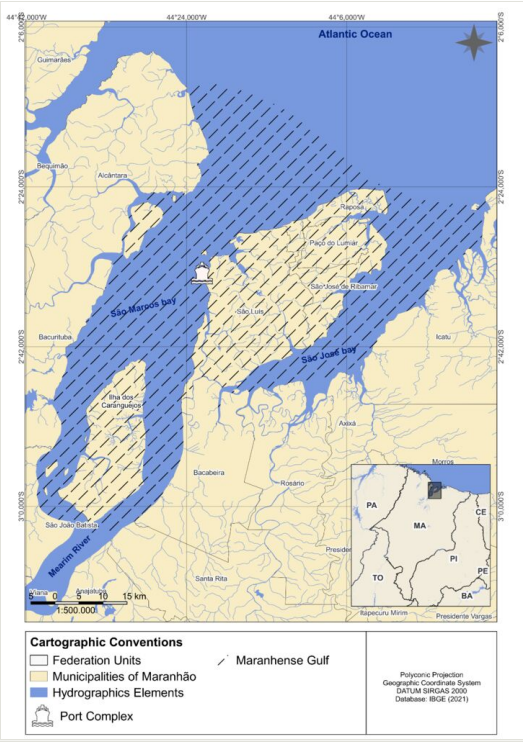

Figure 2. doi

Map showing the *Golfão Maranhense* region, an estuary in the eastern Amazon (Brazil). This is where the secondary biodiversity data was retrieved.

Step 1: Search- systematic review

The first step in retrieving secondary data is to find the data. To do this, a systematic review of the literature is recommended. We conducted a systematic review performing searches in the platforms Science Direct and Google Scholar and public data repositories such as GBIF, VertNet, Wikiaves and SpeciesLink. During the search on the

platforms, we included all the works found, such as scientific articles, books, theses and dissertations. Our search across platforms covered both published (i.e. papers and books) and unpublished literature (i.e. theses, dissertations and environmental consultancy reports focused on licensing). The searches were carried out over two months (June and July 2021) in Brazil. To ensure transparency, completeness and consistency, we followed the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)" guidelines. The PRISMA framework, with its checklist and flow diagram, facilitates reader comprehension and allows for the assessment of 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, lizards and snakes (Testudines, Squamata), birds (Aves), amphibians (Amphibia), plants (Magnoliophyta), fishes, phytoplankton and benthos.
2. Keywords - keywords were defined considering each biotic group (Suppl. material 1). The keywords must include the name of the biotic group (e.g. Amphibians) and the study area (e.g. Maranhão) and may vary according to the needs of each biotic group.
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 76 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. The species' conservation status was sourced from the International Union for Conservation of Nature (IUCN) and the Brazilian Ministry of the Environment (MMA) and applies at the species level rather than the individual level. All variables are described in Suppl. material 2.

Step 2: Validate- data quality and control

To ensure that the data have the lowest possible error rate, they need to go through a validation process. This process reduces the chances that the final data will contain grammatical errors, which can make it difficult to understand and species that have been incorrectly identified.

We conducted a manual validation process for the *Golfão Maranhanse* data that we optimised in two steps:

1. Identifying and fixing errors - We conducted a thorough examination of the data to identify and correct any errors or inaccuracies.

These errors included:

(a) Scientific names (e.g. “*Boana ranipcs*” [wrong] vs. “*Boana raniceps*” [correct]);

(b) Geographic coordinates (e.g. “44,321605 [typo without negative sign]” vs. “-44,321605 [correct]”);

(c) Date: we standardised the sampling dates to "Start Month" "Start Year" "End Month" and "End Year" (the original data column, verbatim date, contains the day of sampling, if available).

Additionally, our data cleaning process involved removing duplicates and standardising entries to ensure consistency. By meticulously correcting these issues, we ensured the data's integrity and reliability, making it suitable for further analysis and interpretation.

2. Checking the records - Species occurrence data are susceptible to misidentification and taxonomic inconsistencies, making this a challenging and dynamic task. To ensure the reliability and validity of the species occurrence data, we reviewed the relevant literature for known geographic species distributions and compared them with the collected points. Our team of taxa specialists meticulously checked each entry for inconsistencies and up-to-date taxonomy, according to the most recent accepted taxonomy of each group. Any mismatches between known and collected geographic distributions served as a first alert, indicating the need for further investigation. Additionally, we reviewed the literature for changes in synonymy and updated the occurrence records accordingly.

Step 3: Standardise

To standardise data from the Golfão Maranhense region, we used the Darwin Core standard (DwC) (Wieczorek et al. 2012). In addition, we manually added columns for data that are not covered by the DwC (e.g. conservation status). 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). Additionally, 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).

Step 4: Share and archive

The last step is to choose the right repository to store the data. For species occurrence data, GBIF (GBIF 2024) may be the best option, as it is a specific repository for

biodiversity data that guarantees data quality and open access. Other repositories that are popular and should be considered:

- Integrated Digitized Biocollections (iDigBio 2024);

- Dryad (2024);

- Zenodo (2024);

- Open Science Framework (OSF 2024);

Statistical analysis

To test whether the number of occurrences depended on the number of taxa in each group, a simple linear regression was performed using R software.

Results

Step 1: Search- Systematic review

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. 3). In addition, we included species occurrence records from four public repositories (GBIF, VertNet, Wikiaves, SpeciesLink). Considering only published 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).

Step 2: Validate- data quality and control

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; three other taxonomic level), amphibians (Amphibia, 55 species; nine to the genus level), reptiles (two Crocodylia; 86 Squamata; 11 Testudines); mammals (Class Mammalia; 101 species; 21 to the genus level), fish (268 species, 74 other taxonomic levels), phytoplankton (370 species; 105 other taxonomic levels), benthos (188 species; 204 other taxonomic levels) and plants (1,624 species; 292 other taxonomic levels) (Suppl. material 4). Most of the taxa were identified to species (81%) and genus (14%) level (Fig. 4). Benthos accounted for the highest number of occurrence events, with 12,510 records and reptiles had the lowest number of occurrence events recorded (570).

Data were carefully analysed 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.

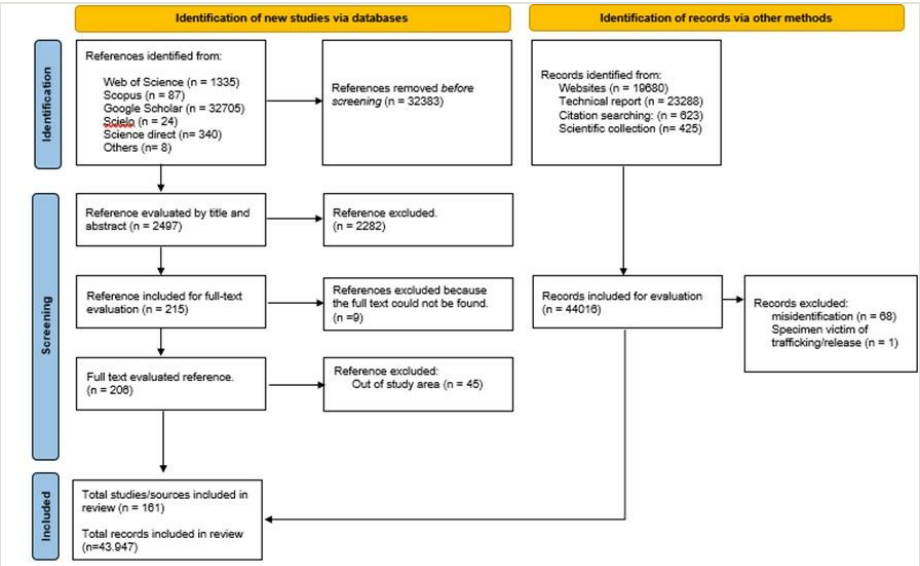

Figure 3. doi

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.

Step 3: Standardise

Of the 76 variables extracted from the studies, 59 were standardised using DwC terms and 17 were adapted due to the lack of appropriate terms for these variables within the current DwC models (Suppl. material 2).

Step 4: Share and archive

We decided not to publish the database in a specialised database such as GBIF because it contains secondary data that includes information extracted from other databases, including GBIF itself. This would result in duplication of information. We decided to store the data in the Open Science Framework (OSF). OSF is an open access repository that maintains version control, allowing them to track changes to their projects over time. The OSF also assigns Digital Object Identifiers (DOIs), making the data citable and ensuring its long-term preservation. The database is publicly available on the link (Marques 2024). In addition, the database will be accessible through an online platform developed using

PowerBI software (Microsoft corporation 2024). This platform will be developed and will be freely available, promoting the dissemination of knowledge related to the biodiversity of the Golfão Maranhense region.

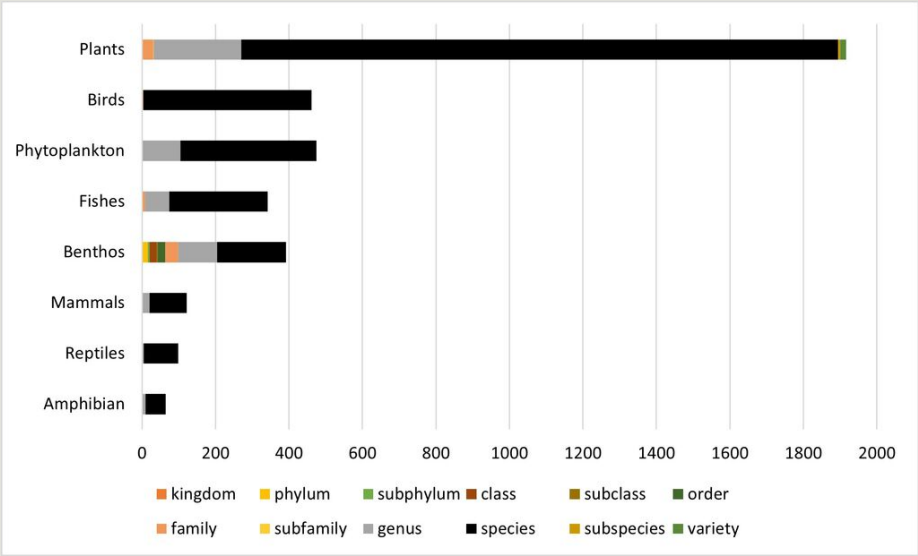

Figure 4. doi

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

Statistical analysis

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 represented 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. 5).

Discussion

We proposed a workflow to improve our ability to recover higher quality biodiversity data using secondary data sources. We were able to extract a large amount of information about the biodiversity of the *Golfão Maranhense* and transform this unrelated data into organised and re-usable data. This systematic approach ensured data accuracy and reliability, facilitating the potential reuse of information in future studies. A further step that we have begun to take for some groups is the systematic survey of museum collections and analyses, focusing on relevant questions that we have identified along the way (e.g. general patterns of occurrence of migratory birds, sampling biases and gaps for many groups etc.).

Researchers can use existing datasets, such as those obtained through our biodiversity data retrieval method, to conduct a wide range of studies to advance scientific research (Pernat et al. 2024). For example, secondary data can be used to conduct meta-analyses (e.g. Biggs et al. (2020)) for comparative studies across different geographic regions and time, to support ecological modelling of species distributions (e.g. Fletcher et al. (2019)), habitat preferences and potential impacts of environmental change (Bayraktarov et al. 2019) as long as they are used judiciously. However, finding high-quality secondary data can be challenging, as evidenced by a recent survey in which most researchers reported that data finding can be arduous (73%) or difficult (19%) (Gregory et al. 2020). Several initiatives have been launched to collect, standardise, store and make biodiversity data openly available (Costello et al. 2013). For example, international repositories, such as "GBIF" (GBIF 2024) and "Freshwater Biodiversity Data Portal- BioFresh" (http://data.freshwaterbiodiversity.eu/) (for other repositories, see Culina et al. (2018)).

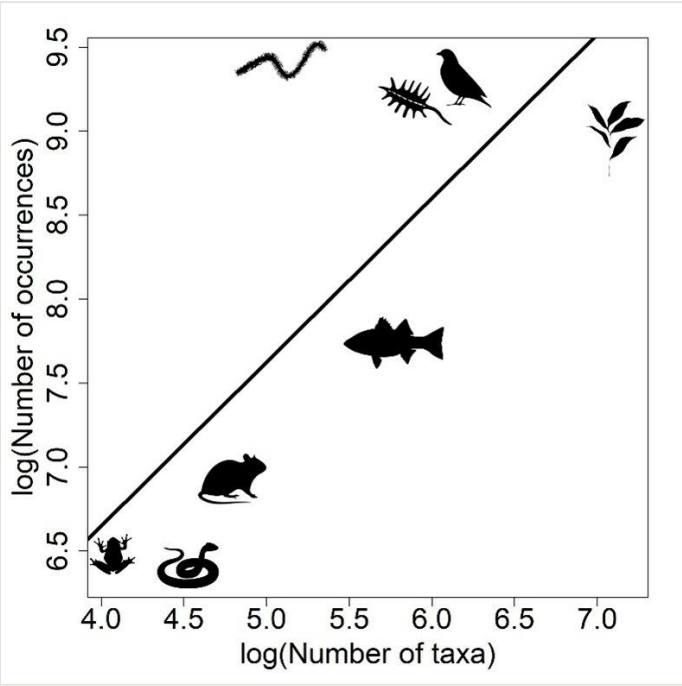

Figure 5. doi

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.

Long-term monitoring datasets can help to 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 increase our knowledge of mammal, bird, fish, amphibian, reptile, marine phytoplankton and benthic species in the

Golfão Maranhense. We are also studying the patterns and determinants of floristic variation in the region and the temporal variation of migratory birds in the *São Marcos* Bay region. While data collected in standardised monitoring programmes, such as LTER (Vanderbilt et al. 2015), can be directly linked to FAIR's standardised data repositories, other secondary data that are not standardised may be important to rescue, treat and use, as they can be thoroughly revised and curated beforehand and stored with some rule of error estimation to build robust hypotheses to investigate and understand biodiversity patterns.

While secondary data can be a valuable resource for scientific research, it is crucial to recognise 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 favours 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). In addition, 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 over-represented, 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 under-represented 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. Thus, most of the data were obtained from environmental monitoring reports in the region linked to the environmental licensing process. These reports conducted in a port area inherently prioritise 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 that are 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 (Jardim Botânico do Rio de Janeiro 2024) 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 standardised worksheets with a Darwin Core base, we were able to effectively search and explore a wide range of scientific articles, technical reports and specialised public repositories. The potential use of secondary data for the advancement of scientific research is significant although it must be taken with care and analysed with precautions observing all bias limitation and filters involved. Many technical survey reports were produced in the *Golfão Maranhense* linked to environmental licensing process 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 hypotheses and producing experiment designs to monitor diversity in a standardised base. 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 amongst researchers, enable comparative analyses across different datasets and support evidence-based conservation strategies and policy-making.

Acknowledgements

We would like to thank the Environmental Management of the Ponta da Madeira Maritime Terminal for their support in developing the project. We are grateful to the reviewer Pedro Cardoso for his suggestions for improving the manuscript and to the collectors of the data we retrieved from the literature. Daniel M. Casali is currently being funded by the grant #2022/00044-7, São Paulo Research Foundation (FAPESP)

Hosting institution

Vale Institute of Technology

Conflicts of interest

The authors have declared that no competing interests exist.

References

- 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
- 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
- Culley T (2017) The frontier of data discoverability: Why we need to share our data. Applications in Plant Sciences 5 (10). https://doi.org/10.3732/apps.1700111
- 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
- Dryad (2024) Data Dryad. https://datadryad.org/
- Fletcher 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
- GBIF (2024) 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
- Hunt V, Jacobi S, Knutson M, Lonsdorf E, Papon S, Zorn J (2015) A data management system for long-term natural resource monitoring and management projects with multiple cooperators. Wildlife Society Bulletin 39 (3): 464-471. https://doi.org/10.1002/wsb.547
- iDigBio (2024) Integrated Digitized Biocollections. https://www.idigbio.org/
- Jardim Botânico do Rio de Janeiro (2024) Flora e Funga do Brasil. http://floradobrasil.jbrj.gov.br/. Accessed on: 2024-7-31.
- 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, et al. (2024) https://osf.io/gh9ym/
- 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
- Microsoft corporation (2024) Power BI. Release date: 2024-6-01. URL: https://powerbi.microsoft.com
- OSF (2024) Open Science Framewor. https://osf.io/
- Parr C, Guralnick R, Cellinese N, Page RM (2012) Evolutionary informatics: unifying knowledge about the diversity of life. Trends in Ecology & Evolution 27 (2): 94-103. https://doi.org/10.1016/j.tree.2011.11.001
- Pernat N, Canavan S, Golivets M, Hillaert J, Itescu Y, Jarić I, Mann HR, Pipek P, Preda C, Richardson D, Teixeira H, Vaz AS, Groom Q (2024) Overcoming biodiversity blindness: Secondary data in primary citizen science observations. Ecological Solutions and Evidence 5 (1). https://doi.org/10.1002/2688-8319.12295
- 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 Sao Luís Island, Maranhão. In: Kjerfve B, Lacerda L, Diop E (Eds) Mangrove ecosystem studies in Latin America and Africa. UNESCO, Paris, 145-154 pp.
- 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
- Zenodo (2024) https://zenodo.org/

Supplementary materials

Suppl. material 1: Keywords doi

Authors: Nubia Marques
Data type: table
Brief description: Keywords used in the systematic review of each biotic group.
Download file (16.13 kb)

Suppl. material 2: Table Darwin Core (DwC) doi

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) doi

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 (395.77 kb)

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

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)