Ocean and Coastal Research

Original Article
Special Article Collection
Di Domenico, M.; Andriguetto-Filho, J.M. and Lopes, R.M. (Eds)
An ocean of science, worms, and humaneness:
a tribute to Paulo Lana
http://doi.org/10.1590/2675-2824072.23102
ISSN 2675-2824

New perspectives for the Annelida collection (National Museum/UFRJ) database: using data visualization to analyze and manage biological collections

Camila Simões Martins de Aguiar Messias [1,*], Carlos Cesar de Oliveira Fonseca [2], Monique Cristina dos Santos [1,3], Asla M. Sá [2], Joana Zanol [1]

[1] Laboratório de Biodiversidade de Annelida – Departamento de Invertebrados – Museu Nacional – Universidade Federal do Rio de Janeiro (Quinta da Boa Vista, s/n – Horto Botânico – CEP 20940-040 – São Cristóvão, RJ, Brazil).
[2] Escola de Matemática Aplicada – Fundação Getúlio Vargas (Praia de Botafogo, 190 – CEP 22253-900, Botafogo, Rio de Janeiro, RJ, Brazil).
[3] Pós-graduação em Biologia Celular e Molecular – Fundação Oswaldo Cruz (Avenida Brasil, 4365 – Manguinhos, Rio de Janeiro, RJ, Brazil).

* Corresponding author: camila.messias@mn.ufrj.br

Abstract

Collection management faces many challenges in keeping stored items preserved and the information associated with them accurate and organized. It is essential for the expansion and use of this biodiversity repository that the database is unambiguous and that errors are quickly identified and corrected. This work aims to show the use of interactive visual representations (IVRs) of the collection's metadata as tools to inspect the data and help solve these challenges. To do this, we used the Annelida collection database from the National Museum (MN) of the Federal University of Rio de Janeiro (UFRJ). Interactive graphs of the metadata within this database (catalog date, taxonomic identification and determiners, sampling, depth, geographic localization, and collector data) were created with the Altair library in the Python 3 language. Data analyses using these graphs made it possible to identify anomalous patterns in the data and fill in missing records. They also provided an understanding of the spatial and bathymetric distribution of the specimens deposited over time, and the growth rate of the collection in each family, thus projecting future growth and solutions for the physical organization of vials. Graphs are an ally in the management of collections with digital entry forms and aim to facilitate the availability of metadata associated with cataloged specimens. Likewise, IVRs can even be used to give credit to the researchers involved in building biological collections. Thus, visualization tools are efficient in recognizing global patterns present in databases and solving biological collection management tasks.

Keywords: Polychaetes, Biological collection, Management, Interactive visual representations

INTRODUCTION

Different public and private institutions host biological collections with a wide range of origins that allow the study of the biodiversity on this planet throughout time. Collections in natural history museums (NHM) are repositories of biodiversity that can be consulted for the development of taxonomic, ecological, public health, anthropological, and historical studies (Krishtalka and Humphrey, 2000; Suarez and Tsutsui, 2004; Johnson et al., 2023). In addition to the specimens, large databases are created

Submitted: 29-jun.-2023
Approved: 06-Dec-2023
Associate Editor: José Milton Andriguetto Filho

with the metadata associated with each specimen, from the collectors' field notes to information on the scientific publications featuring the specimens (Krishtalka and Humphrey, 2000; Peterson et al., 2004; Page et al., 2015).

Data digitization has become a reality for many collections across the world, using technology to optimize collection management and data availability and switching, whenever possible, to born-digital files (Krishtalka and Humphrey, 2000; Beaman and Cellinese, 2012; Blagoderov et al., 2012; Page et al., 2015; Scott et al., 2019; Hedrick et al., 2020; Medeiros e Sá et al., 2022). This change in the logic of work, from physical to virtual, raises different challenges, such as the increase in errors. These may be due to the expansion of the amount of metadata about each specimen, different people working with the same database, and difficulties interpreting forms that were previously handwritten (Krishtalka and Humphrey, 2000; Graham et al., 2004; Page et al., 2015; He et al., 2021). The management of collections must include the identification and correction of errors in order to make it possible to consult specimens and associated metadata, and make high-quality data available for research and education (Cook et al., 2014; National Academies of Sciences, Engineering, and Medicine, 2020).

Biological collections are consistently expanding their holdings. Therefore, global detection of errors in databases becomes humanly impossible without the help of tools that work with the data in a global way (Peterson et al., 2004; Wang et al., 2015; Scott et al., 2019; Ribeiro et al., 2022). The incorporation of automated tools (e.g., OpenRefine, Miller and Vielfaure, 2022) that perform database cleaning into the curatorial routine is one way to improve the data (Ribeiro et al., 2022). However, identifying errors in data that lack facets or global patterns or contain outliers that are indeed rare facts and not errors, as is common in biological collections databases, cannot be done without the supervision of an expert in the history of the biological collection and the nature of the specimens (Medeiros e Sá et al., 2022). Thus, the inspection of biological collection databases benefits from other tools that are not completely automated.

Data visualizations of biological collections help to inspect the data and understand biological or historical patterns present in the data (Liu et al., 2014; Scott et al., 2019; Medeiros e Sá et al., 2022). Data visualizations transform data of different variables into a set of visual representations, summarizing information and revealing patterns. Data visualization tools can be categorized as static or interactive. In static visual representations (SVRs), a single figure encompasses all the information presented in the figure. In contrast, interactive visual representations (IVRs) allow the same figure to offer various data visualizations. IVRs can include options such as interactive filters that reduce the amount of plotted data based on the user's selections or the presentation of additional data related to the plotted point, such as the use of tooltips, thereby expanding the information available (Medeiros e Sá et al., 2022).

Data inspection by data visualization allows the recognition of unidimensional errors based on one variable or object at a time, but more importantly of errors that become clear from patterns revealed by the combination of multiple variables or objects and their consistency with the curatorial team's knowledge. Biological collections databases are usually managed by specialists who understand the nature of the data and the history of the collections, relevant knowledge that needs to be considered in data inspection and which cannot be automated as the human-in-the-loop paradigm suggests (Shnneiderman, 1996; Liu et al., 2018; Medeiros e Sá et al., 2022). Data visualizations also allow for different analyses of the metadata and the development of new hypotheses, help solve problems in the management of the collection, and offer more information to the researcher who is carrying out research with that group, adding scientific value to the collection (Keim, 2002; Shiravi et al., 2012; Wilson et al., 2021; Medeiros e Sá et al., 2022).

In this article, we show how IVRs can act as tools for inspecting large databases and quickly detecting anomalous patterns, which, combined with the knowledge of the curatorial team, enables global verification of the metadata. The database

used in this work is that of the Annelida collection of the National Museum (MN) of the Federal University of Rio de Janeiro (UFRJ). This collection is an important testimonial to the diversity of the phylum. Annelida is one of the most diverse phyla of terrestrial and marine invertebrates. They are distributed worldwide and dominate marine macrobenthic communities in terms of richness and abundance (e.g. Lana and Bernardino, 2018). Some species play a key role in communities by altering the environment, reworking sediments, or even participating in the structuring of new environments (Hutchings, 1998). The entire Annelida collection has 7488 lots, the majority of which are traditional polychaetes, but there are also sipunculans, echiurans (Thalassematidae), and oligochaetes. The data presented in the IVRs, besides allowing the identification of anomalous patterns, have opened up new perspectives for understanding the diversity of specimens deposited in the collection on geographic and temporal scales, the importance of investing in employees focused on collection management, and how large sampling projects can contribute to expanding the collection.

METHODS

The IVRs were constructed using the Annelida collection (MNRJP) data from its inauguration in 1999 until March 2023. At the time of analyses, the Annelida database was a single spreadsheet made in the Excel program (Microsoft Office), but originally there were two different spreadsheets, one for each subcollection, IBUFRJ (3257 lots) and MNRJ (4231 lots). Both subcollections were merged into one collection in March 2023, when the lots with IBUFRJ catalog numbers received new catalog numbers with the acronym MNRJP. The curatorial team for the collection and its associated database consists of two people, one curator and one technical staff. But the people who perform these functions, with two other people as curators and six other people as technical staff, have changed, replacing each other over the years.

The IVRs were built using the Altair library in Python 3 and the methodology proposed by Medeiros and Sá et al., 2022. The process of developing the IVRs involved the participation of the curatorial team of the Annelida collection (MNRJP), expressing the challenges and demands of the curators and the designer team. Constant meetings were held over almost 2 years to verify that the IVRs were fulfilling their specific purpose of helping to inspect anomalous data. Once the necessary IVRs had been defined and generated, the data was inspected by the curatorial team. The following information about the cataloged specimens was included and inspected using the IVRs: catalog number, catalog year, taxonomic identification at the family level, determiner name, collector name, sampling date, locality name, and sampling geospatial coordinates and depth. The color palette used in all the legends considers the taxonomic proximity of the families according to Rouse et al. (2022). In all IVRs that included information on taxonomic identification, cataloged lots without identification below order level were grouped as *non-identified* in the legend or axis. Lots cataloged without information on the year of cataloging or the year of sampling were grouped as *not available* (N/A) on the axis representing these data. All IVRs displaying data on sampling coordinates or depth excluded the lots lacking these data. In the case of sampling by dragging over a depth interval, we decided to include only the minimum depth in the IVR for visualization purposes. The depth interval is indicated in the tooltip of each plotted point. The interactive versions (Figures S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, Supplementary material, Messias et al., 2023) allow selecting one family or type status at a time by clicking on the name in the legend. In addition, in the interactive versions, additional information regarding the registers represented at a given x-y position (such as the catalog number, the number of lots, the year, or the minimum and maximum depths) is displayed in drop tags called “tooltips” when hovering the cursor over the points on the graph. In some IVRs, in addition to the tooltip, auxiliary graphs detail the point chosen in the IVR, which is essential to eliminate the ambiguity of overlapping information. The tool used allows the y-axis to be sorted alphabetically and in temporal order.

All data inspections were conducted by observing the patterns yield by one variable and/or the combination of multiple variables in the IVRs by the two people on the curatorial team (Table 1). For example, errors in geospatial coordinates were identified by observing the distribution patterns of one family at a time and the information contained in the tooltip of each record. In this way, we were able to identify lots with divergent locality names and geospatial coordinates. Data causing unexpected patterns were: 1. identified as anomalous data using IVRs; 2. examined further by consulting documentation, publications related to the deposited specimens, and by inquiring former collection staff, collectors, and determiners; 3. classified as a lack of information, an error, or an unexpected event; and 4. filled in or corrected, when necessary, using the same sources of information as in step 2 (Table 1). Once the data had been filled in or corrected, new IVRs were generated and rechecked until all the identified errors had been corrected (Figure 1).

Despite the use of the IVRs for data inspections, the static versions of the visualizations are presented along with the text to illustrate the features available in the IVRs in a static file, which is incompatible with the interactive versions. All the IVR files are hosted in the Zenodo repository (Messias et al., 2023).

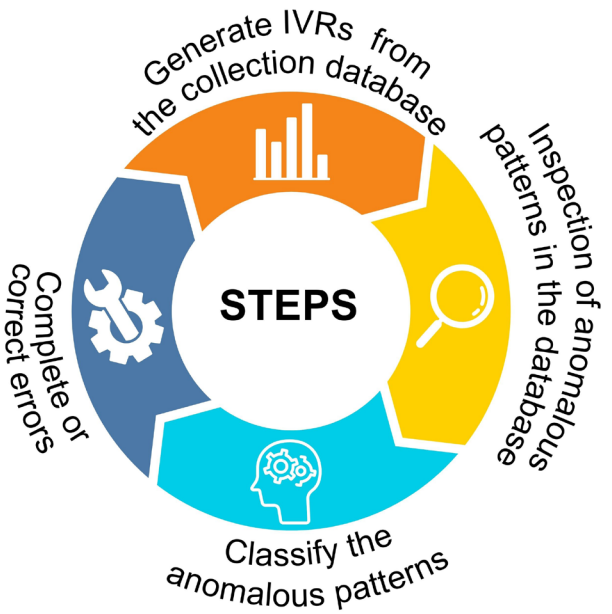

Figure 1. The steps of the method consisted of: 1. to generate interactive visual representations (IVRs) using the collection database; 2. to analyze the IVRs and to identify anomalous patterns in the database; 3. classify the anomalous patterns as a lack of information, an error or a rare event; and 4. complete lack information or correct errors, when needed.

Table 1. Errors identified analyzing the interactive visual representations (IVRs) created with the Annelida collection database (MNRJP) and correction approaches.

Interactive Visual Representation (IVR) used	Anomalous Patterns Identified	Verification/Correction Approach
Geographic distribution of species cataloged and dynamic version geographic distributions x sampling year (Figs. 2 and S1). Types per Family x Cataloged Year (Fig. 3). Author x Cataloged Year (Fig. 3). Depth x Family, observing details in tooltips (Fig. 4B). Collector x Sampling Year (Fig. 5). Identified by x Cataloged Year (Fig. 6). Family x Cataloged Year (Fig. 7).	Dates (i) Catalog dates prior to the opening date of the collection. (ii) Use of the same value for the fields collection, determination, and catalog dates. (iii) Sampling dates of expeditions differing from scientific documents. (iv) Determination, cataloged or sampling dates in a period inconsistent with the researcher's age or the moment of his scientific production.	Search for the original information in academic papers and reports regarding the deposited specimen. Correlation with the collection's opening year. Inquiry of former employees, collectors, and researchers. Correlation with the period formers employees were responsible for the collection. Correlation with the collection date of other specimens from the same study or field trip.
All IVRs were useful because their elements such as legends, tooltips and words on the x and y axes could be analyzed.	Spelling (i) Misspelling in the taxonomic classification. (ii) Different ways to write the name of the same researcher, collector, employee, or trip. (iii) Typos in general.	Consultation of species databases for the correct spelling of taxon names. Standardizing the way of writing people and trip names.
Geographic distribution of species cataloged interactive version with sampling year (Figs. 2, S1, S2 and S3). Family x Depth interactive version with minimum and maximum depth (Fig. 4 and S4).	Spatial (i) Geospatial coordinates or name of the location do not correspond to where the specimen was collected. (ii) Minimum and maximum depths do not cover expedition information or previously documented distribution for the taxon.	Search for the original information in academic papers and reports regarding the deposited specimen. Standardizing the way of estimating geographic coordinates, when only the name of the occurrence locality is available, and to include observation about this type of approach in the database.
All IVRs had their elements such as legends, tooltips and words on the x and y axes analyzed.	Lack of information (i) Incomplete record in fields of the database.	Search for the original information in academic papers and reports regarding the deposited specimen. Inquiry of former employees, collectors, and researchers. Verification of specimen labels in the collection.

RESULTS AND DISCUSSION

Identifying and correcting errors in the database

The interactive features of the IVRs allow subsets of data to be selected and lots with anomalous data to be precisely identified, which can be classified into four categories and corrected using different approaches (Figures 2-7, S1, S2, S3, S4, S5, S6, S7, S8, S9, S10 and S11, and Table 1). Among the errors are dates (time), and geospatial coordinates (spatial), two of the main determining variables of analytical and inference methods that can be applied to biodiversity databases and for which there are several qualifying automated methods of assessment and correction such as CoordinateCleaner (Zizka et al., 2019), BDcleaner (Jin and Yang, 2020) and BDC toolkit (Ribeiro et al., 2022) (Meyer et al., 2016). The identification of anomalous data using IVRs adds to these automated methods by including the human factor to interpret patterns and recognize unpredicted errors (Keim, 2002; Medeiros e Sá et al., 2022), as in typos in date values recognizable by other variables, such as the time frame of the sampling expedition, coordinates inconsistent with the distribution of taxa, names of researchers, which can only be detected by an expert who knows the history of the collections and the distribution of specimens globally.

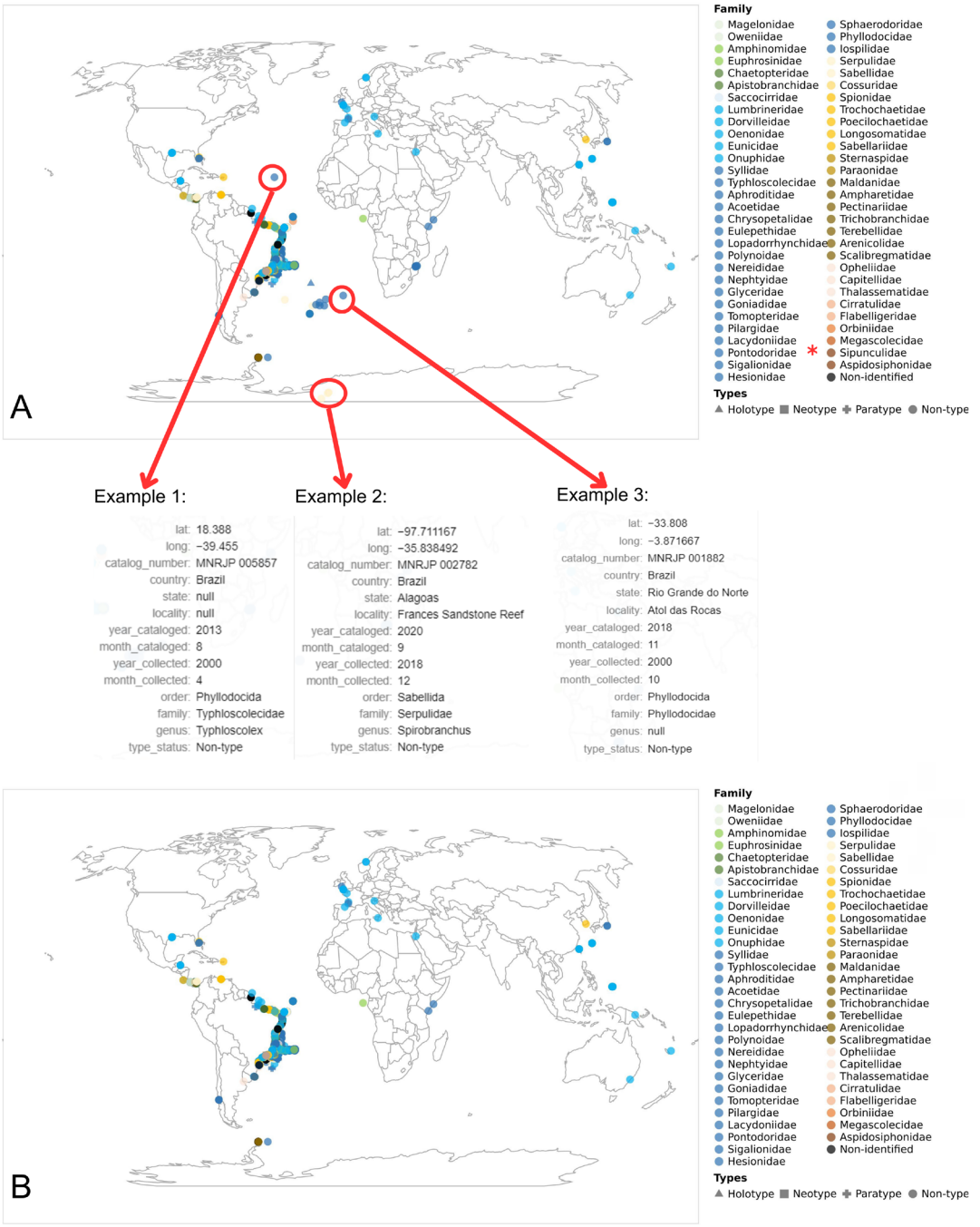

Figure 2. Geographic distribution of species cataloged in the Annelida collection (MNRJP) throughout the globe. A. All lots, examples of tooltips present in the interactive version and of errors detected by analyzing them, the red asterisk indicates an error in the family identification in the database. In this graph, we highlight three errors that are easy to identify based on the analysis of the points and their tooltips: example 1 is a decimal point position error, example 2 is a minus sign error, and example 3 is the exchange of latitude and longitude information. In all these cases, the error was identified due to the inconsistency with the location name. With ever increasing databases, the visualization tools can help in the identification of errors. B. All lots with previously detected errors corrected. In addition, the dynamic version of the graph makes it possible to click on the names in the legends to reduce the data plotted and allows analyses by family or by type status. Cataloged lots without geospecial coordinates data were excluded.

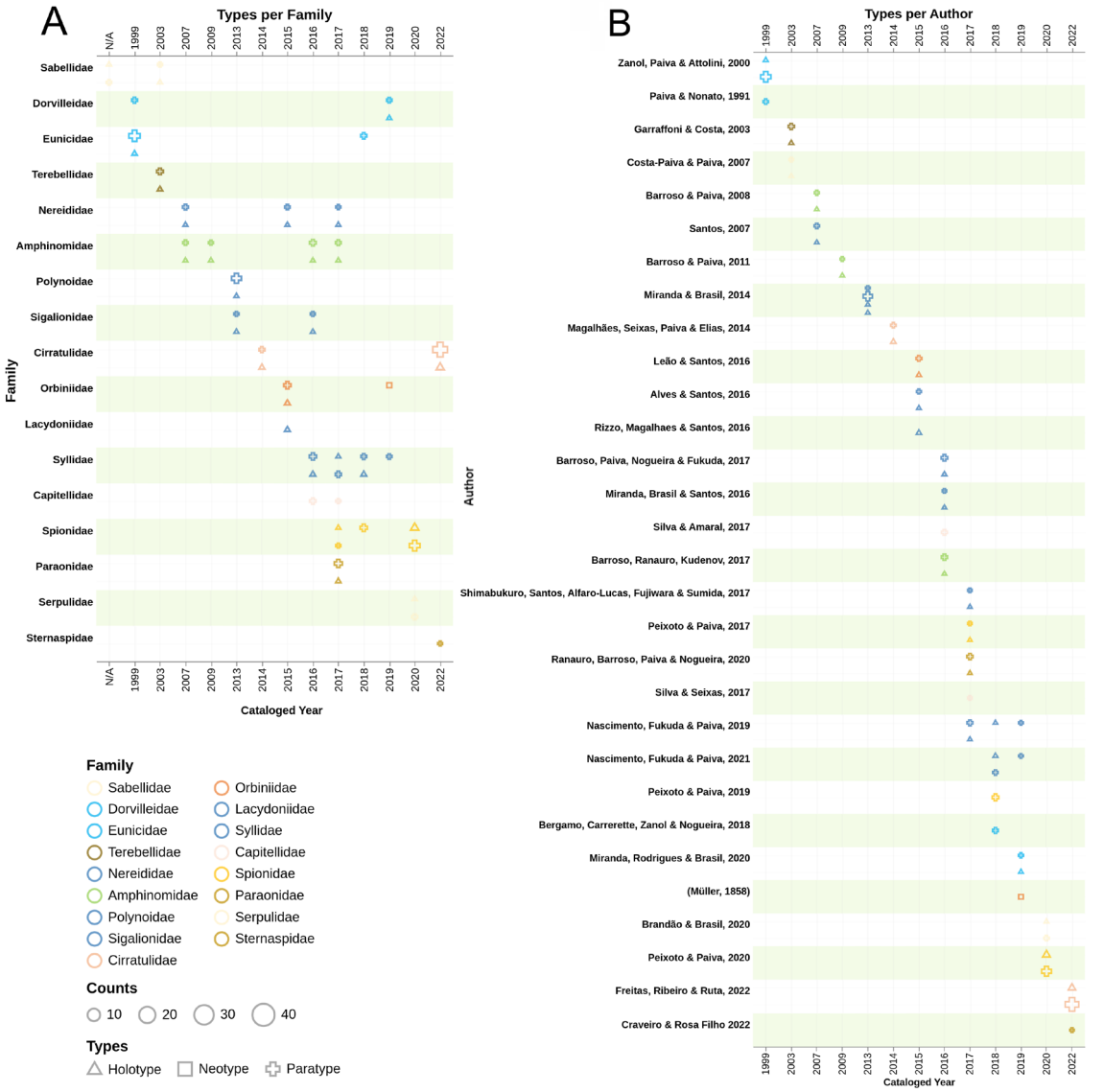

Figure 3. Type material in the in the Annelida collection (MNRJP). A. Count of type material per family cataloged over the years in Annelida collection (MNRJP) organized from the oldest to the most recent. B. Contribution of species determiners to the specimens cataloged in the type series over the years. These graphs (A and B) lead to an understanding of which type specimens are being deposited in the collection at each time. This information shows the relevance of the collection to scientific research. In addition, the dynamic version of the graph makes it possible to click on the names in the legends to reduce the data plotted and allows analyses by family or by different category in the type series. N/A = not available, cataloged lots without cataloged year information.

Most of the anomalous geospatial coordinates were errors caused by switched negative and positive signs on latitude and longitude values, or by the decimal position (Figures 2, S1, S2 and S3). In these cases, the knowledge of the curatorial team of the collection's history and the places where specimens occur were essential to identify points plotted in the wrong places, but mainly to understand whether the error was in the coordinate or in the name of the locality, corroborating the human-in-the-loop paradigm (Shneiderman, 1996; Liu et al., 2018; Medeiros e Sá et al., 2022). The scientific literature related to the specimens also facilitated correction and access to extra information, especially when it included the associated catalog numbers.

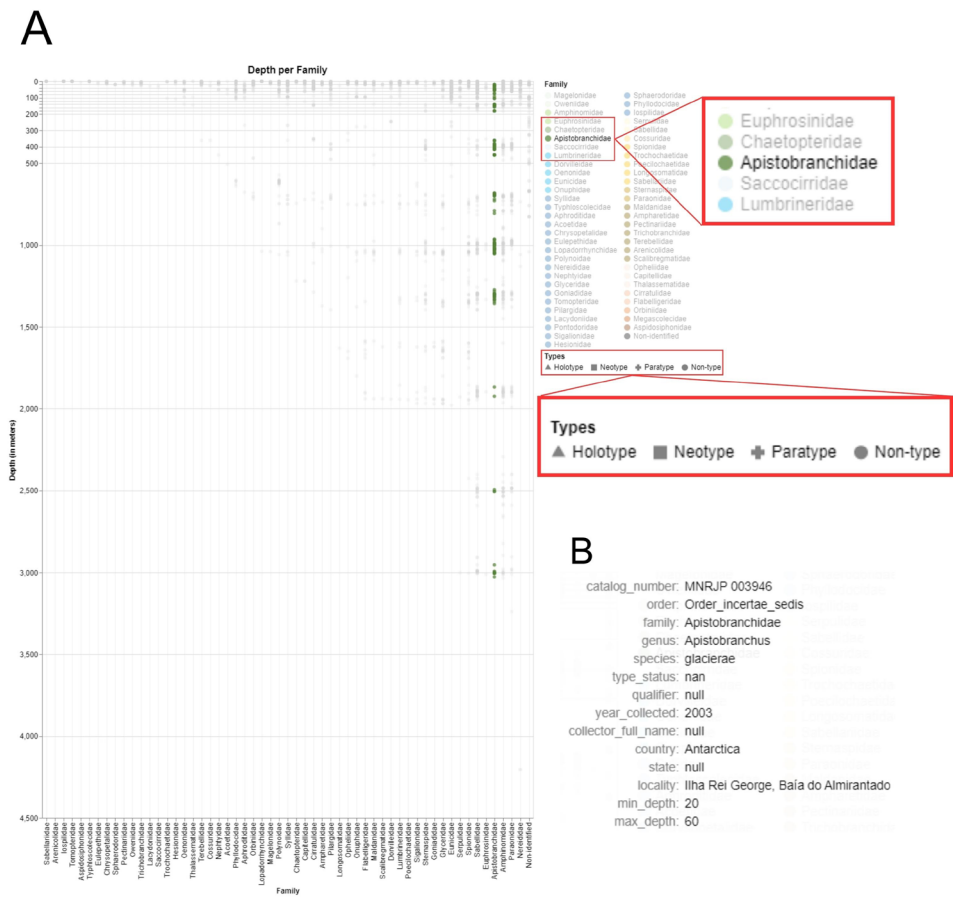

Figure 4. Minimum sampling depths of lots by families deposited in the Annelida collection (MNRJP). A. Interactive function of clicking the family name on the legend, example for the family Apistobranchidae, helping to visualize its distribution at different depths. The same interactive function occurs when clicking on the names of the type status, reducing the plotted data, and facilitating the consultation of information sets. B. Tooltip present in the interactive version showing the depth range of the collection site of one of the lots and additional information that allows each point on the graph to be associated with its data in the database. This graph can be used by collection consultants to generally understand whether there is the presence of specimens of the taxa of interest deposited in the collection in a depth range that is the target of their research. The detailed search for the metadata of the cataloged specimens, after the perception of interest, is carried out by querying the database. Cataloged lots without depth data were excluded.

Many of the anomalous data identified as missing were due to information not requested at the time of registration. It is fundamental that specimen entry forms have all the informative fields filled in and checked by specialists before incorporating the data into the database. This facilitates the construction of databases with as much information as possible, allowing scientific knowledge to be multiplied and avoiding the waste of time and money to obtain information that has already been worked on previously (He et al., 2021; Wilson et al., 2021). The lack of information in some cases was due to the practice of reserving catalog numbers, and for many of the numbers the specimen was never deposited despite the use of the numbers in publications. The practice of reserving catalog numbers can lead to irreversible errors in the long term and does not allow the principle of scientific reproducibility to be verified (Graham et al., 2004; Wilson et al., 2021). The deposited specimens, quantity, and integrity must be verified before catalog numbers are assigned, except in extraordinary situations that make this impossible, such as health or climate catastrophes (e.g., the COVID-19 pandemic).

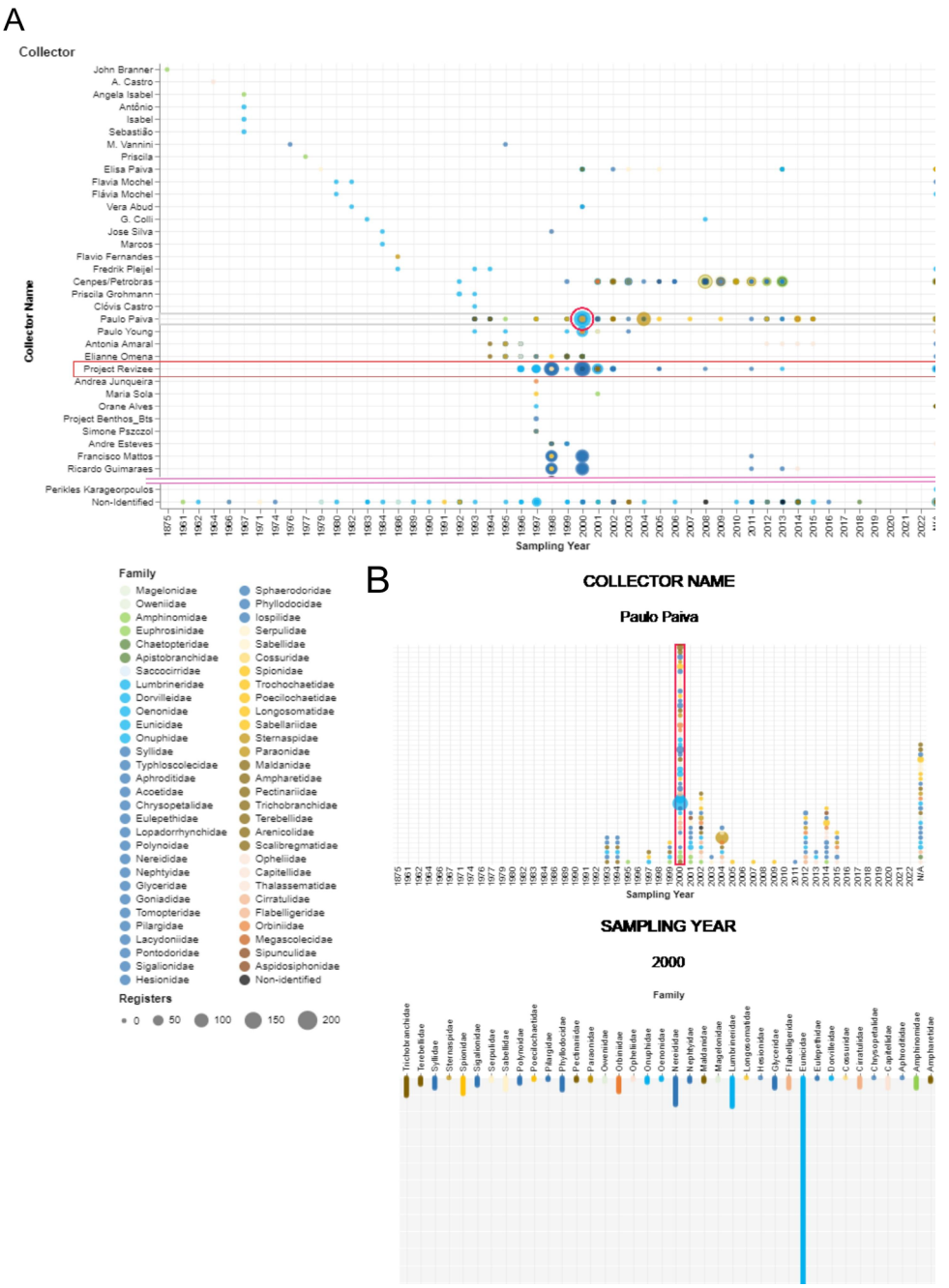

Figure 5. Collectors and sampling years of the lots in the Annelida collection (MNRJP) by collectors. A. Partial collector list with y-axis in temporal order. The Project Revizee SCORE Central record is highlighted as an example of an expedition with documented sampling years from 1996 and 2002, but the graph shows that there are records outside this period, indicating the need to correct the database. B. Auxiliary graphs present in the interactive version showing in detail the contribution of one collector over the years, example choosing Paulo Paiva. Furthermore, the tool used allows the axis to be sorted in alphabetical order, facilitating the search for duplicated names due to differences in abbreviation or typo. Pink lines indicate omitted data that can be verified in the complete interactive version. Cataloged lots not identified down to the family level were excluded. N/A = not available, cataloged lots without sampling year information.

Figure 6. Determiners of the identification of specimens cataloged over the years in the Annelida collection (MNRJP). A. Complete list with y-axis in alphabetical order. B. Complete list with y-axis in temporal order. C. Auxiliary graphs present in the interactive version showing in detail the contribution of one determiner over the years, example choosing Paulo Lana. These graphs allow the understanding of the contribution of researchers to taxonomic identifications. Cataloged lots not identified down to the family level were excluded. Pink lines indicate omitted data that can be verified in the complete interactive version. N/A = not available, cataloged lots without cataloged year information.

Figure 7. Count of families cataloged over the years in Annelida collection (MNRJP) organized from the most to the least frequently cataloged. Families that are most cataloged are at the top of the y-axis. This graph helps us to manage the physical space of the collection, as it shows which families are cataloged most over time. In addition, the dynamic version of the graph makes it possible to click on the names in the legends to reduce the data plotted and allows analyses by family. N/A = not available, cataloged lots without cataloged year information.

Alternative representations of the name of a person, who contributed to the collection as collector or determiner, and of taxa, as well as incorrect dates were identified in all IVRs that included these information (Figures 2, 3, 4, 5, 6, S1, S2, S3, S4, S5, S6, S7, S8, S9 and S10). The alphabetical ordering of the y-axis with the names of collectors and determiners makes it easier to identify errors and duplicated names (Figures 5, 6, S7 and S9). Graphs with the names of major expeditions or research projects and sampling dates (Figures 5, S7 and S8) also helped to identify incoherent dates. The information was confirmed by reports and articles generated about these expeditions and projects or by the curatorial team's experience with specimens from a specific sampling efforts. The Revizee Project, for example, had the SCORE Central expeditions from 1996 to 2002,

but the graph showed that there were records outside this period in our database (Figures 5A and S7). Errors in the records of sampling depth could also be recognized by graphs that include these data (Figures 4 and S6). In part, these errors are due to typos, the lack of controlled vocabulary, and the turnover of collection staff. This turnover leads to different logics of cataloging the database that often cannot be clarified due to the lack of records of what motivated each decision.

The use of digital documents for biological collection management practices is greatly encouraged, as it facilitates the identification of errors, avoids ambiguous interpretations in writing, and favors access to information (Hedrick et al., 2020). In this direction, the cataloging and movements in the collection that were historically carried out in notebooks and physical forms are being replaced by born-digital databases and forms. A great advantage of born-digital databases is the use of software workflows (e.g. OpenRefine, BDC toolkit, Bdcleaner) to indicate inconsistencies in data such as taxonomy identification, coordinates, and names of locations and dates to be verified "and corrected by the team (Zizka et al., 2019; Jin and Yang, 2020; Ribeiro et al., 2022). The generation of IVRs presented here complements the workflow, improving the identification, by the curatorial team's knowledge, of unexpected errors and also of global patterns in the databases that add scientific value to new hypotheses and data papers (Fayyad et al., 2001; Wang et al., 2015; Medeiros e Sá et al., 2022). The human factor is both the strength and the weakness of the suggested data inspection and correction method. The strength comes from the possibility of analyzing anomalous patterns that arise from combining data and for which there is no global pattern or facets, thus making the construction of automated software workflows a challenge. The weakness comes from the dependence on a knowledgeable team to observe the IVRs, recognize anomalous patterns, and search for the information to understand the reason for that pattern and correct it if necessary, which is time-consuming, in addition to the dependence on a design team to create IVRs that meet curatorial needs. This is evident when examining the study conducted by Medeiros and Sá et al. (2022), wherein IVRs generated from the database of the Carcinology collection (MN/UFRJ) were presented alongside additional visualization strategies chosen by its curatorial team and developed in collaboration with a design team. It is worth noting that curatorial teams that do not have the opportunity to collaborate with a design team to create customized IVRs can use SVR tools built on the metadata similar to that employed in this study (Table 1). Despite their inherent limitations, these SVR tools can offer initial support for data inspection.

KNOWLEDGE AND MANAGEMENT OF THE COLLECTION

Analyses of the IVRs allowed the extraction of patterns that help optimizing the interpretations of the database, such as geographic and bathymetric distribution, the collection contributors, and the management of cataloged specimens, such as the physical organization of lots. IVRs also help to broaden the scientific community's knowledge of the collection (Figures 2, 3, 4, S2, S4, S5 and S6), which can increase consultation of the database and cataloged.

The IVRs related to geographic and bathymetric distribution of the Annelida collection lots show that the specimens represent all continents, totalizing 25 countries and Antarctica, and diverse depths, from the intertidal region to the deep sea (Figures 2, 4, S2, S3 and S6). The interactive filter in the IVRs enables researchers to obtain a comprehensive overview of the taxon's distribution. Subsequently, if interested, they can delve deeper into the database. In this way, the biodiversity information present in the collection's database can contribute to broader analyses aimed at understanding and modelling the taxa distribution (Guisan and Thuiller, 2005).

The history of sampling dates and the locality of specimens in the collection can be traced on the map's IVRs, which can show only the lots sampled in a certain year (Figure S1). This highlights the presence of temporal series from the same region in the collection and the historical importance of lots, which may have been sampled before the collection was opened. In the case of the Annelida collection (MN/UFRJ), the oldest lot was sampled in 1875

during the Geological Commission of the Empire in Brazil (1875–1877).

Researchers, collectors, projects, and expeditions have played an essential role in collection growth over time. Despite this, little credit is given to their contribution to the increment of their respective collections (Rouhan et al., 2017; Hedrick et al., 2020). IVRs can emphasize the contribution of different people and the taxa impacted by their work, as authors of new species (Figures 3, S4 and S5), collectors (Figures 5, S7 and S8), or as taxonomists who identified specimens (Figures 6, S9 and S10), and allow verification of the correct spelling of their names. The arrangement of the y-axis from the first collectors or determinants to the most recent allowed us to see their contributions to the collection over time (Figures 5, 6B, S8 and S10), which can be clearly perceived by distinguishing overlapping points in the auxiliary graphs (Figures 5B, 6C, S7, S8, S9 and S10). The use of this type of graph helps to give credibility to the contribution of these researchers in data aggregators, such as GBIF, and in data papers of biological collections databases that can be referenced in funding projects and in the career progressions of the researchers involved.

The graphs also helped to understand the importance of depositing the specimens collected, even without taxonomic identification at species level. Making specimens available to the scientific community allows for their analyses, which may result in new records or the description of new species for science, without the need of sampling efforts in the field (Figures 3, S4 and S5). Information associated with specimens in biological collections has the potential to corroborate scientific studies and to enable new hypotheses (Keim, 2002; Shiravi et al., 2012; Wilson et al., 2021; Medeiros e Sá et al., 2022).

Projecting the growth of biological collections is one of the challenges in management, as it implies the possibility of expansion and the need for maintenance resources. There are mathematical approaches that help with this estimation, but they require familiarity with complex calculations and depend on many variables (Xu et al., 2007; Ariño, 2010; Comoglio et al., 2013). On the other hand, the graph of family x year cataloged with the families organized from the most to the least frequently cataloged allows us to understand the pattern of families entering the collection each year (Figures 7 and S11). This pattern, along with knowledge of the size of specimens, helps to plan the space necessary for future growth of the entire collection, as well as to specific taxa. It also allows us to understand the relative abundance of families, revealing the taxonomic strengths of the collection and where it is interesting to increase taxonomic representation.

CONCLUSION

Biological collections have been positively impacted by technological advances. The interpretation and creation of IVRs with the extraction of their potential and information content is not disconnected from the human factor. The specialists in curating collections generate demands and needs, as well as recognize incoherent patterns that can mean errors or rare valid facts. The important challenges of the approach presented are: 1. curatorial teams need to be able to generate the data visualizations so that these can be included in the work routine, and 2. collections need to have specimen entry forms as complete as possible and in digital format to enable data manipulation and generation of IVRs.

The analyses of these graphs made it possible to identify inconsistent information, which could be corrected or supplemented in the event of missing records. They also made it possible to identify global patterns of spatial and bathymetric distribution of the specimens deposited over time, as well as the growth rate of the collection, thus in projection of future growth and of solutions for the physical organization of the vials. A qualified team, funding efforts, institutional policies for making data available, and smart tools are essential for maintaining biological collections as a source of scientific knowledge.

ACKNOWLEDGMENTS

We are grateful to Paulo Lana for all his great contributions to science and, in particular,

to Brazilian science and scientists, challenging us to think outside the box. Paulo was a role model for his great enthusiasm, excellence in science and guidance, even for those of us who did not work with him directly. For all of these, we would like to dedicate this article to him. We would also like to thank Cristiana Serejo for initiating the partnership between the A.M.S. and the curatorial teams of the National Museum (MN/UFRJ), forming a task force for optimizing curatorship, digitizing collections, and making data available.

AUTHOR CONTRIBUTIONS

C.S.M.A.M.: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing – original draft; Writing – review & editing.

C.C.O.F.: Methodology; Software; Visualization; Writing – review & editing.

M.C.S.: Data curation; Writing – review & editing.

A.M.S.: Methodology; Software; Visualization; Supervision; Writing – review & editing.

J.Z.: Conceptualization; Supervision; Resources; Formal analysis; Project Administration; Writing – review & editing.

REFERENCES

Ariño, A. H. 2010. Approaches to estimating the universe of natural history collections data. Biodiversity Informatics, 7(2), 81-92.

Beaman, R. & Cellinese, N. 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, 209, 7–17.

Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T. & Smith, V. 2012. No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys*, 209, 133–146.

Cook, J. A., Edwards, S. V., Lacey, E. A., Guralnick, R. P., Soltis, P. S., Soltis, D. E., Welch, C. K., Bell, K. C., Galbreath, K. E., Himes, C., Allen, J. M., Heath, T. A., Carnaval, A. C., Cooper, K. L., Liu, M., Hanken, J. & Ickert-Bond, S. 2014. Natural history collections as emerging resources for innovative education. *BioScience*, 64(8), 725-734.

Comoglio, F., Fracchia, L. & Rinaldi, M. 2013. Bayesian Inference from Count Data Using Discrete Uniform Priors. *PLoS ONE*, 8(10), e74388.

Fayyad, U., Grinstein, G. G. & Wierse, A. 2001. Information visualization in data mining and knowledge discovery. Burlington, Morgan Kaufmann Publishers

Graham, C., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9), 497–503.

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9), 993–1009.

He, P., Chen, J., Kong, H., Cai, L. & Qiao, G. 2021. Important Supporting Role of Biological Specimen in Biodiversity Conservation and Research. *Bulletin of Chinese Academy of Sciences*, 38(12), 11.

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V. & Davis, C. C. 2020. Digitization and the Future of Natural History Collections. *BioScience*, 70(3), 243–251.

Hutchings, P. 1998. Biodiversity and functioning of polychaetes in benthic sediments. *Biodiversity and Conservation*, 7(9), 1133–1145.

Jin, J. & Yang, J. 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, 21, e00852.

Johnson, K. R., Owens, I. F. P. & The Global Collection Group. 2023. A global approach for natural history museum collections. *Science*, 379(6638), 1192–1194.

Keim, D. A. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8.

Krishtalka, L. & Humphrey, P. S. 2000. Can Natural History Museums Capture the Future? *BioScience*, 50(7), 611 -617.

Lana, P. C. & Bernardinho, A. F. (ed.). 2018. Brazilian Estuaries. Cham: Springer International Publishing.

Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., Wang, Y. S. & Hong, S. 2018. Steering data quality with visual analytics: The complexity challenge. *Visual Informatics*, 2(4), 191–197.

Liu, S., Cui, W., Wu, Y. & Liu, M. 2014. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12), 1373–1393.

Medeiros e Sá, A., Oliveira, F. A., Schneider, B., Echavarria, K. R. & Serejo, C. S. 2022. Visually Overviewing Biodiversity Open Data Digital Collections. *In*: *Proceedings of the Symposium on Open Data and Knowledge for a Post-Pandemic Era ODAK22*, UK.

Messias, C. S. M. A., Fonseca, C., Santos, M., Sá E Medeiros, A. & Zanol, J. 2023. New perspectives of Annelida collection (National Museum/UFRJ) database: using data visualization to analyze and manage biological collections. *Ocean and Coastal Research*. https://doi.org/10.5281/zenodo.8092072

Meyer, C., Weigelt, P. & Kreft, H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 19(8), 992–1006.

Miller, M. & Vielfaure, N. 2022. OpenRefine: An Approachable Open Tool to Clean Research Data. *Bulletin - Association of Canadian Map Libraries and Archives*), (170), 2-8.

National Academies of Sciences, Engineering and Medicine. 2020. *Biological Collections: Ensuring Critical Research and Education for the 21st Century*. Washington, DC, National Academies Press.

Page, L. M., Macfadden, B. J., Fortes, J. A., Soltis, P. S. & Riccardi, G. 2015. Digitization of Biodiversity Collections Reveals Biggest Data on Biodiversity. *BioScience*, 65(9), 841–842.

Peterson, A. T., Navarro-Sigüenza, A. G. & Pereira, R. S. 2004. Detecting errors in biodiversity data based on

collectors' itineraries. *Bulletin of the British Ornithologists Club*, 124, 143–151.

Ribeiro, B. R., Velazco, S. J. E., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P. & Loyola, R. 2022. bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, 13(7), 1421–1428.

Rouhan, G., Dorr, L. J., Gautier, L., Clerc, P., Muller, S. & Gaudeul, M. 2017. The time has come for Natural History Collections to claim co-authorship of research articles. *TAXON*, 66(5), 1014–1016.

Scott, B., Baker, E., Woodburn, M., Vincent, S., Hardy, H. & Smith, V. S. 2019. The Natural History Museum Data Portal. *Database*, 2019, baz038.

Shiravi, H., Shiravi, A. & Ghorbani, A. A. 2012. A Survey of Visualization Systems for Network Security. *IEEE Transactions on Visualization and Computer Graphics*, 18(8), 1313–1329.

Shnneiderman, B. 1996. The eyes have it: a task by data type taxonomy for information visualizations. *In*: *Proceedings IEEE Symposium on Visual Languages* (pp. 336–343). Boulder: IEEE Computer Society Press.

Suarez, A. V. & Tsutsui, N. D. 2004. The Value of Museum Collections for Research and Society. *BioScience*, 54(1), 66-74.

Wang, R., Perez-Riverol, Y., Hermjakob, H. & Vizcaíno, J. A. 2015. Open source libraries and frameworks for biological data visualisation: A guide for developers. *PROTEOMICS*, 15(8), 1356–1374.

Wilson, S. L., Way, G. P., Bittremieux, W., Armache, J., Haendel, M. A. & Hoffman, M. M. 2021. Sharing biological data: why, when, and how. *FEBS Letters*, 595(7), 847–863.

Xu, J., Wu, S. & Li, X. 2007. Estimating Collection Size with Logistic Regression. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 789-790). New York, ACM.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte, C. R., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V. & Antonelli, A. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods in Ecology and *Evolution*, 10(5), 744–751.