Applied Geography 96 (2018) 11–15

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

A new methodology for the retrieval and evaluation of geographic coordinates within databases of scientific plant collections

Ulises Rodrigo Magdalena[a,*], Luís Alexandre Estevão Silva[a], Rafael Oliveira Lima[a], Ernani Bellon[a], Rafael Ribeiro[a], Felipe Alves Oliveira[a], Marinez Ferreira Siqueira[b], Rafaela Campostrini Forzza[b]

[a] *Núcleo de Computação Científica e Geoprocessamento, Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro, RJ, Brazil*
[b] *Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão, 915, Rio de Janeiro, RJ, CEP 22460-030, Brazil*

ARTICLE INFO

Keywords:
Biodiversity database
Herbarium database
Data quality
Collection
Herbarium management system

ABSTRACT

Several types of research can be conducted with data about plant collections stored in databases. Although geographic coordinates are on many specimens, the accuracy of this information is often questionable due to several factors. This paper introduces a new method to analyze the quality of spatial data and, when necessary, a way to search for coordinates based on data in other existing fields in a herbarium database. We propose implementing a tool to assist curators in evaluating and improving the quality of geographic data that is made available to the public and used for scientific research, establishing public policies, and environmental management. To validate our approach and its efficiency, a case study was conducted. The present work demonstrates that it is possible to develop and apply a methodology to streamline the process of improving data quality in databases of herbarium collections.

1. Introduction

Biodiversity databases are extremely important for species inventories, conservation assessments, other scientific research, and establishing public policies and environmental management strategies (Costello & Wieczorek, 2014). To further these objectives, the Rio de Janeiro Botanical Garden plant collection management system – Jabot (Silva et al., 2017) is a tool that provides access to data of specimens stored in herbaria. As a system for the dissemination of data about the Brazilian flora, the goal of Jabot (Silva et al., 2017) is to publish, maintain, and improve the data quality of mounted specimens and related collections in the RB herbarium (Forzza et al., 2016). This is the largest herbarium in Brazil (Gasper & Vieira, 2015) and includes mounted specimens (RB – 750,000, with 7500 nomenclatural types and around 3000 paratypes), wood (RBw – ca. 10,300 specimens), fruits (RBcarpo – ca. 8000 specimens), DNA bank (RBdna – ca. 5700 specimens), spirit (RBspirit – ca. 2500 specimens), seed bank (RBsem – ca. 2700 specimens) and ethnobotany (RBetno – ca. 200 specimens) (Lanna et al., 2018). The overall collection (physical and/or digital) has an enhanced importance due to the significant amount of stored material and the annual average number of new samples (20,000–30,000 per year) (Lanna et al., 2018). However, the age of a large part of the samples in this collection precedes the advent of GPS (Forzza et al., 2008) and there is a great deal of unregistered information concerning geographic coordinates of collections (approximately 83% of the collection) (Lanna et al., 2018). Thus, there is an urgent need to georeference collections to accurately conduct research related to historical series, predictive modeling of species distributions, and conservation status assessments of taxa throughout the country (Kamino et al., 2012). A good system should provide reliable information and strengthen public policies for plant conservation. However, to solve these problems, new methodologies are necessary to construct standard processes for retrieving and correcting geographic positioning mistakes that are currently in databases (Chapman & Wieczorek, 2006; Neufeld, Guralnick, Glaubitz, & Allen, 2003; Robertson, Visser, & Hui, 2016; Sua, Mateus, & Vargas, 2005).

The accuracy of record retrieval depends directly on raw location data written on the specimen label. Therefore, some measures for error estimation or confidence must be considered based on the accuracy of the information reported by the collectors (Chapman, 2005b; Murphey, Guralnick, Neufeld, & Ryan, 2004). Improving these data would lead to considerable gains in spatial analyses for species distribution modeling, in which errors and their implications on results can be estimated (Velásquez-Tibatá, Graham, & Munch, 2016). In this paper, we discuss

* Corresponding author.
E-mail addresses: ulisesrodrigo@id.uff.br (U.R. Magdalena), estevao@jbrj.gov.br (L.A.E. Silva), rafael@jbrj.gov.br (R.O. Lima), ebellon@jbrj.gov.br (E. Bellon), ribeiro@jbrj.gov.br (R. Ribeiro), felipealves@jbrj.gov.br (F.A. Oliveira), marinez@jbrj.gov.br (M.F. Siqueira), rafaela@jbrj.gov.br (R.C. Forzza).

https://doi.org/10.1016/j.apgeog.2018.05.002
Received 14 September 2017; Received in revised form 2 May 2018; Accepted 6 May 2018
0143-6228/ © 2018 Elsevier Ltd. All rights reserved.

the development of a new methodology for retrieving geographic coordinates of records of specimens in the RB herbarium, in order to assist users of Jabot in the evaluation process and improvement of plant collection spatial data.

2. Methodology

The analysis, retrieval, and certification of the coordinates in this work correspond to applying routine procedures that allow recognizing mistakes, correcting them, and acquiring the precise values of a specific location. In general, our methodology proposal is organized in five phases: (1) definition of geographic database; (2) application of data standardization routines; (3) identification of the records with geographic coordinates; (4) identification of the records without geographic coordinates; and (5) location search.

2.1. Definition of the geographic database

In order to implement the methodology, the first step is to define a specific geographic reference base. Jabot uses the Continuous Cartographic Base of Brazil (BC250, 2015), which is provided by IBGE (Instituto Brasileiro de Geografia e Estatística). This cartographic base includes all 5570 municipalities in the 26 states and one federal district that constitute the formal divisions of Brazil, which makes it possible to standardize and synchronize the names and geographic limits of the divisions. Thus, Jabot can update information about changes in territorial delimitations over time while retaining the records of locations described on specimen labels.

2.2. Data cleansing and standardization

Before describing the steps that must be taken for the evaluation, the quality of the data must be evaluated (Chapman, 2005a) using a group of functions and routines that clean and standardize the records. During this process, the name of the country, state or federal district, and municipalities are standardized and then assessed according to BC250. In addition to the names of the geographic units, other important attributes are evaluated, such as collector names. These steps are needed because collections come from many different collectors over time, so the writing and information on specimen labels varies, which generates gaps in the database. Author names that are randomly written and the lack of collection distinction (marine or terrestrial) are standardization examples.

2.3. Selecting records with geographic coordinates

Using BC250 allows records with coordinates to be validated through a PostGis (Spatial and Geographic Objects for PostgreSQL) function of correlation (Obe & Hsu, 2011) between values of the collection coordinates and values obtained from the reference base. The function verifies if the geographic coordinates reported on the specimen correspond to the geographic limits of the geopolitical unit in Fig. 1. To validate a record, the coordinates must be converted to decimal format. If they match, then the data are marked as correct. However, if there are inconsistencies, then three common types of problems related to geographic data from botanical collections are analyzed: (1) error caused by the inversion of latitude and longitude values; (2) the absence of a cardinal direction that identifies the N – S and E – W hemispheres; and (3) the confirmation that a marine collection corresponds to an aquatic taxon. In the first case, the solution is to invert the coordinates based on the location details. In the second case, it is necessary to denominate the cardinal direction according to the locality in the record. For the third case, an analysis of the taxon and location must be conducted to confirm if the material is of aquatic origin.

As a rule, in historical herbaria such as RB, records with geographic coordinates are in the minority. Therefore, incompatible data can only be individually analyzed after performing the processes previously described. This is due to errors, such as the following: typing coordinates in an incompatible way, confusing them with respect to international (longitude and latitude) and national (latitude and longitude) standards; typing municipalities that are homonyms but belong to different states (e.g., Bonito, Mato Grosso do Sul and Bonito, Pernambuco); and other cases described in Table 1. In the case of reviewed records, newly recovered coordinates are classified and stored in the database as “suggested coordinates.” The suggested values are stored in another field, which preserves the original values. Therefore, both values are available for queries and can serve as additional information for users.

2.4. Identifying records without geographic coordinates

Upon establishing records that lack geographic coordinates, a filter is employed to differentiate collections with and without a location description. For records without a description, there is nothing that can be done with respect to georeferencing; although, it does not reduce their importance for studies of a historical series, taxonomy, and the possibility of inferring a general location based on Jabot's integration with Flora of Brasil 2020 (www.floradobrasil.jbrj.gov.br) For the specimen labels that include a location description, analyses are performed to ensure the greatest accuracy possible when inferring the geographic coordinates. For these data, the user is informed that the stated coordinates are estimated in accordance with the description on the specimen label. Table 1 provides descriptions of the types of errors found and their possible causes.

2.5. Location search

It is possible to infer the collection location for a record without coordinates if the specimen has a location description and/or some of the following information on the label: country, state, municipality, gazetteer, name of protected area (if informed), and physical toponyms.

Georeferencing is conducted using toponym lists from IBGE and other federal institutes with database services, such as Serviço Geológico do Brasil,[1] Instituto Chico Mendes de Conservação da Biodiversidade,[2] Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis,[3] and Ministério do Meio Ambiente - MMA.[4] This is the slowest phase because it is usually carried out record by record. In order to optimize the location, searches of all the collections from the same location are grouped and, subsequently, the coordinates are inferred for the group.

3. Results

To evaluate the efficiency of the described methodology, this study considered 5779 samples that had inconsistencies between the geographic information from BC250 and the coordinates and locations from the sample labels in the Jabot database (Fig. 2).

While making the corrections, the types of errors and their possible causes were cataloged (Table 1). As a result, 5421 records were assessed and fixed, which were from the following biomes and locations: Atlantic Rainforest (2,467), Amazon (865), Coastal and Marine Zones (724), Cerrado (697), Caatinga (517), Pantanal (11), Pampa (4), other locations in South America (135), Central America (1), and Antarctica on King George Island (1) (Fig. 3). Of the total number of samples, 19% are categorized according to Flora of Brasil 2020 (www.floradobrasil.jbrj.gov.br) as endemic species of Brazil (Table 2), 2% are categorized according to the National Centre for Plant Conservation (www.cncflora.

[1] http://geobank.cprm.gov.br/.

[2] http://www.icmbio.gov.br/portal/.

[3] http://www.ibama.gov.br/.

[4] http://www.mma.gov.br/.

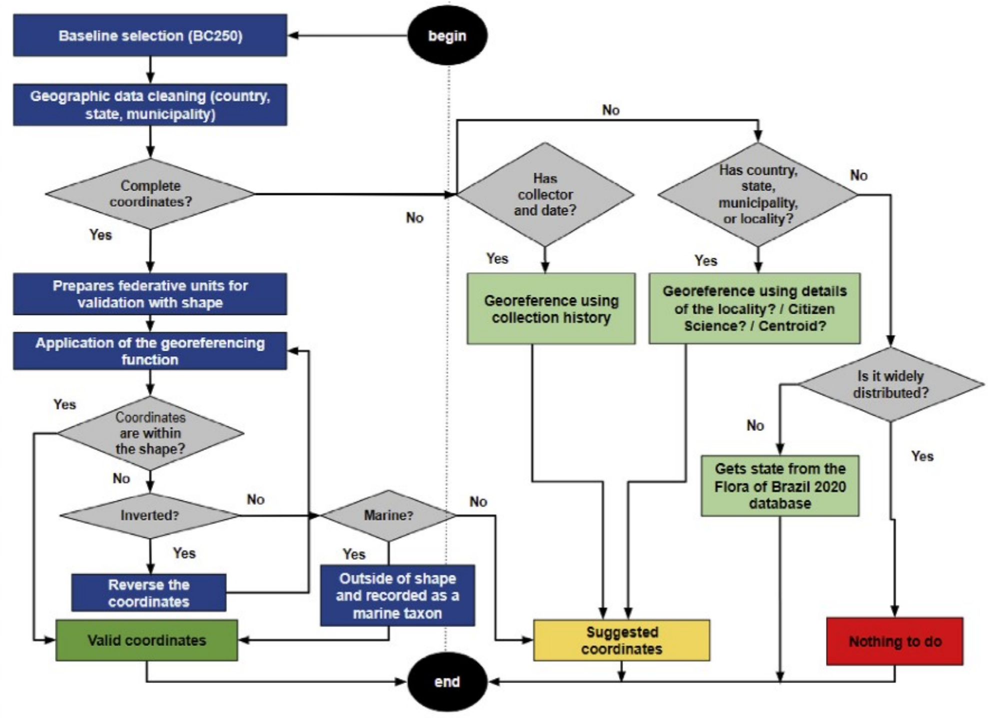
begin
Baseline selection (BC250)
Geographic data cleaning (country, state, municipality)
Complete coordinates?
Yes
No
Prepares federative units for validation with shape
Application of the georeferencing function
Coordinates are within the shape?
Yes
No
Inverted?
Yes
No
Reverse the coordinates
Marine?
Yes
No
Outside of shape and recorded as a marine taxon
Valid coordinates
Has collector and date?
Yes
No
Georeference using collection history
Has country, state, municipality, or locality?
Yes
No
Georeference using details of the locality? / Citizen Science? / Centroid?
Is it widely distributed?
No
Yes
Gets state from the Flora of Brazil 2020 database
Suggested coordinates
Nothing to do
end

Fig. 1. General overview of the proposed methodology – the proposed methodology operates with two different groups of data: with coordinates (left) and without coordinates (right). The quality of the data is evaluated according to existing data, in the first scenario, and an attempt is made to recover the coordinates from the existing data.

Table 1
Description of errors and their causes.

Error	Causes
a) Collection's coordinates with values equal to zero ("Lat/ Long = zero")	**1)** The collection is old; **2)** the coordinates were not included in the database; **3)** the coordinates are not listed on the specimen label, even though the collection is recent.
b) Marine locality for terrestrial species or land locality for marine species	**1)** Latitude or longitude was mistakenly entered in the database; **2)** the location description was not clearly written; **3)** the coordinate was wrongly written on the specimen label; **4)** the coordinates are inverted (i.e., where the latitude and longitude values are written as longitude and latitude values).
c) Collection made on the border of a region (municipality, state or country) or land-sea border	**1)** Imprecise coordinates; **2)** unclear location when determining the coordinates (this kind of problem occurs in regions where there is no GPS signal).
d) The coordinates are incomplete, insufficient or wrong	**1)** Latitude and longitude incorrectly entered in the database or in the collection description on the specimen label.
e) Name of the country, state or municipality does not correspond to the coordinates on the specimen label	**1)** The coordinates were mistakenly entered; **2)** updated database due to new state or municipality; **3)** country, state and/or municipality wrongly entered in the database.

jbrj.gov.br) as endangered species, and 1% as both endemic and endangered species.

The relative percentage of recovered records (Table 2) indicates that the Caatinga and Cerrado biomes (with 32% and 17% endemic species, respectively) are relatively rich and biologically more fragile due to the smallest number of integrated protected areas and conservation units coupled with a strong landscape transformation process caused by human occupation (MMA[4]). The compilation of endemic and endangered species, as another possible tool for the method, can help environmental management by identifying hotspots of biodiversity and endemism. The high values for the Atlantic Rainforest are related to the historical collection and location of the RB herbarium.

The other retrieved collections (358) did not have sufficient information on their sample labels (i.e., they lacked coordinates and a location description). When it comes to percentages regarding the samples used, the methodology identified the following: 34% with correct coordinates; 14% marine material; 14% with inverted coordinates; 7% with coordinates indicating another location; 1% with coordinates on a border; 28% with incomplete coordinates; and 2% without coordinates.

4. Discussion

Analyses of the geographic coordinate retrieval process in Jabot highlight the fact that control mechanisms and routines are greatly needed to maintain and verify the quality of data. The proposed methodology managed to decrease the impacts caused by the huge amount of uncorrected raw data that exists in herbarium databases. It also favors the processes of individually correcting samples since it saves time analyzing errors and recovering geographic coordinates of a specimen.

The Jabot collection georeferencing process also highlights species distributions on a time scale, which can contribute to public policies of biodiversity conservation related to land use and coverage,

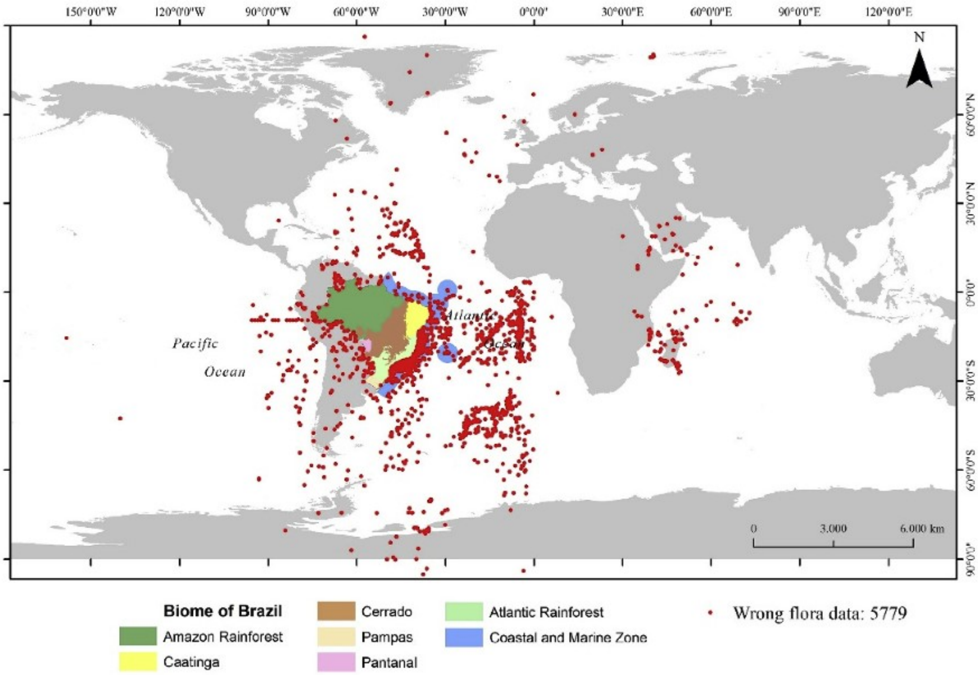

Fig. 2. Map of situation before applying the proposed methodology. Note the numerous dots that are incorrectly marked in the sea.

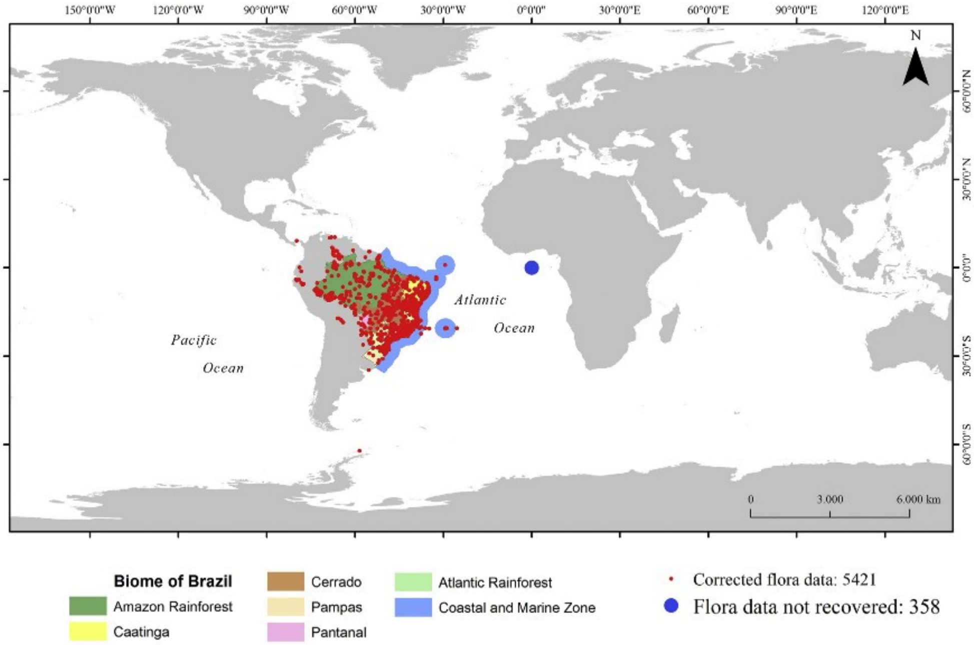

Fig. 3. Map after applying the methodology. As a result, a major retrieval of terrestrial points and confirmation of aquatic points can be observed.

conservation status, and distribution patterns. The proposed and performed methodology also highlighted the performance of the researchers at the Jardim Botânico do Rio de Janeiro, and associate colleagues, in relation to the plant species collections from Brazil and abroad.

Recovered species records categorized as threatened by the CNCFlora were compared to published data in the *Red Book of the Brazilian Flora* (Martinelli & Moraes, 2013). The results show that from a total of 119 records of endangered species, only 23 are in accordance with the areas of occurrence listed in this book and 17 are far outside the known areas of occurrence (supplementary materials). Thus, the results support the expansion or creation of new occurrence areas for

Table 2
Number of endemic species sampled and the percentage of retrieved samples by biome.

Biome	Samples	%
Pampa	0	0
Other	2	1
Amazon	34	4
Coastal and Marine Zone	31	4
Cerrado	116	17
Pantanal	2	18
Atlantic Rainforest	699	28
Caatinga	165	32

many endangered species.

The proposed methodology differs from other works (Chapman & Wieczorek, 2006; Robertson et al., 2016; Sua et al., 2005) due to the goal of correction and recovery optimization of geographic coordinates for the collected and stored data of the plant specimens in the RB herbarium, which minimizes manual data cleaning. The consulted methodologies emphasize the definition of standardization actions in the process of data entry in flora databases. BioGeoR (Robertson et al., 2016) proposes a method that automatically corrects geographic coordinates of historic samples used for modeling species distributions. However, it does not consider important information, such as inferring the regional distribution based on the known distribution of the taxon.

The methodology proposed in the present work is defined by a logical sequence and promotes the recovery of geographic coordinates by identifying errors (Table 1) and showing how to fix them, in addition to using information about the distributions of endemic taxa to infer the general location for records that do not have coordinates. Based on this perspective, the approach decreases the amount of raw data during the process of manual cleaning, retrieves records of historical collections in Jabot, and assists in the demarcation of new areas of occurrence of endangered species.

5. Conclusion

One of the limitations of evaluating geographic data of herbarium specimens is the incompatibility between the extensive amount of data to be analyzed and the reduced number of people available to work on collections. Therefore, any proposed methodology must consider the maximum optimization of the error identification process and the possible retrieval and release of data for new research. Thus, this work shows that designing a routine methodology can optimize the processes of analysis and retrieval of geographic coordinates of Jabot collections. The study also emphasizes the use of filters and control routines for importing new data records. Correlations were created in order to avoid the input of incorrect data into new records and there is a new tool for importing data (http://jabot.jbrj.gov.br/v2/validarplanilha_externo.php). This tool was developed to adjust filters, aiming to preserve the quality of the data added to Jabot. We conclude that the developed methodology is having a positive impact on the quality of flora data. Hence, we suggest using this methodology in management systems of scientific collections, which will increase the number of reviewed georeferenced specimens, with the goal of making more reliable data available for research.

Funding

The authors thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro), SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira), and the Natura Company for supporting the project and providing computational resources and support for systems development. M.F.S. and R.C.F. received a Research Productivity Fellowship and U.R.M received a Research Technology Productivity Fellowship from CNPq.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apgeog.2018.05.002.

References

BC250 (2015). *Continuous cartographic base of Brazil.* ftp://geoftp.ibge.gov.br/mapeamento_sistematico/base_vetorial_continua_escala_250mil.

Chapman, A. D. (2005a). *Principles and methods of data Cleaning: Primary species and species- occurrence data.* http://www2.gbif.org/DataCleaning.pdf.

Chapman, A. D. (2005b). *Principles of data quality.* http://imsgbif.gbif.org/CMS_ORC/?doc_id=3135&download=1.

Chapman, A. D., & Wieczorek, J. (2006). Guide to best practices for georeferencing. *Copenhagen Global Biodiversity Information Facility*. http://www.herpnet.org/herpnet/documents/biogeomancerguide.pdf.

Costello, M. J., & Wieczorek, J. R. (2014). Best practice for biodiversity data management and publication. *Biological Conservation, 173*, 68–73. https://doi.org/10.1016/j.biocon.2013.10.018.

Flora do Brasil 2020 Under construction. http://floradobrasil.jbrj.gov.br/.

Forzza, R. C., Carvalho, A., Andrade, C. S., Franco, L., Estevão, L. A., Fonseca-Kruel, V. S., et al. (2016). Coleções Biológicas do Jardim Botânico do Rio de Janeiro à luz das metas da GSPC/CDB: Onde estaremos em 2020? *Revista Museologia & Interdisciplinaridade, 5*(1), 125–141. http://periodicos.unb.br/index.php/museologia/article/view/19234.

Forzza, R. C., Mynssen, C. M., Tamaio, N., Barros, C., Franco, L., Pereira, M. C. A., et al. (2008). *As coleções do herbário. 200 anos do Jardim Botânico do Rio de Janeiro.* Rio de Janeiro: Jardim Botânico do Rio de Janeiro.

Gasper, A. L., & Vieira, A. O. S. (2015). Herbários do Brasil. *66 Congresso Nacional de Botânica: Vol. 4*, (pp. 1–11). Santos: Bioscience.

Kamino, L. H. Y., Stehmann, J. R., Amaral, S., Marco, P., Rangel, T. F., Siqueira, M. F., et al. (2012). Challenges and perspectives for species distribution modelling in the neotropics. *Biology Letters, 8*(3), 324–326. https://doi.org/10.1098/rsbl.2011.0942.

Lanna, J. M., Silva, L. A. E., Morim, M. P., Leitman, P. M., Queiroz, N. O., Filardi, F. L. R., et al. (2018). Herbarium collection of the Rio de Janeiro Botanical Garden (RB), Brazil. *Biodiversity Data Journal, 6*, e22757. https://doi.org/10.3897/bdj.6.e22757.

Martinelli, G., & Moraes, M. A. (2013). *Livro vermelho da flora do Brasil. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro* (1st ed.). Rio de Janeiro: Andrea Jakobsson978 85 88742 58 1.

Murphey, P. C., Guralnick, R. P., Neufeld, D., & Ryan, J. A. (2004). Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the mountain and plains spatio-temporal database- informatics initiative (Mapstedi). *PhyloInformatics, 3*, 1–29. http://doi.org/10.5281/zenodo.59792.

Neufeld, D. L., Guralnick, R. P., Glaubitz, R., & Allen, J. R. (2003). Museum collections data and online mapping applications - a new resource for land managers. *Mountain Research and Development, 23*(4), 334–337 isi:000187317300006.

Obe, R. O., & Hsu, L. S. (2011). PostGIS in action. *Geography*.

Robertson, M. P., Visser, V., & Hui, C. (2016). Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography*. (January), Pre-release https://doi.org/10.1111/oik.02629.

Silva, L. A. E., Fraga, C. N., Almeida, T. M. H., Gonzalez, M., Lima, R. O., Rocha, M. S., et al. (2017). Jabot - botanical collections management system: The experience of a decade of development and advances. *Rodriguésia - Revista do Jardim Botânico do Rio de Janeiro, 68*(2), 391–410. https://doi.org/10.1590/2175-7860201768208.

Sua, S., Mateus, R. D., & Vargas, J. C. (2005). *Georreferenciación de registrosbiológicos y gacetero digital de localidades*.

Velásquez-Tibatá, J., Graham, C. H., & Munch, S. B. (2016). Using measurement error models to account for georeferencing error in species distribution models. *Ecography*. https://doi.org/10.1111/ecog.01205.