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 Cientifica 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

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/.