Rodriguésia 70: e03222017. 2019
http://rodriguesia.jbrj.gov.br
DOI: http://dx.doi.org/10.1590/2175-7860201970032

Original Article / Artigo Original
Tool for validation and import in herbarium database

Luís Alexandre Estevão da Silva[1,2], *Felipe Alves de Oliveira*[1,2], *Rafael Oliveira Lima*[1], *Ernani Bellon*[1], *Rafael da Silva Ribeiro*[1], *Leonardo da Silva Clemente*[1], *Erika von Sohsten de Souza Medeiros*[1] & *Ulises Rodrigo Magdalena*[1]

Abstract

Many biological collections databases feature data quality problems. On the existing computational resources, we present an import tool and data validation. The program applies filters to data submitted through a spreadsheet at the time of data import, streamlining the error-checking process. The validations presented were divided into three categories according to the taxonomic, geographical and general specimen collection data. Its implementation eliminated the errors in the data entry of new vouchers in the Herbarium of the Botanical Garden of Rio de Janeiro.
Key words: biodiversity database, data quality, herbarium database, herbarium management system.

Resumo

Muitos bancos de dados de coleções biológicas apresentam problemas de qualidade de dados. De acordo com os recursos computacionais existentes, apresentamos uma ferramenta de importação e validação de dados. O programa aplica filtros aos dados enviados através de uma planilha no momento da importação dos dados, agilizando o processo de verificação de erros. As validações apresentadas foram divididas em três categorias de acordo com os dados taxonômicos, geográficos e gerais das coletas de espécimes. Sua implementação eliminou os erros na entrada de dados de novos vouchers no Herbário do Jardim Botânico do Rio de Janeiro.
Palavras-chave: banco de dados de biodiversidade, qualidade de dados, banco de dados de herbário, sistema de gerenciamento de herbário.

Introduction

On the challenges of the protection of natural resources, the generation of knowledge from large databases of biodiversity has attracted the attention of the whole society. These data are essential for taxonomic research and conservation actions, among others (Donaldson 2009; Lavoie 2013; Wen *et al.* 2015), providing relevant and useful information for preservation policies. Various herbarium management information systems and data portals have been developed in recent years to facilitate access and integration of collections. Stands out among them the Global Biodiversity Information Facility-GBIF (GBIF 2010) that serves as an aggregator of data of herbaria around the world. Despite the significant volume of data and considering only the flora data available at GBIF, even in a superficial analysis, one can see that there is a big difference between the number of flora and fauna records and quality of spatial data that needs to be improved, beyond the visible low data quality. Among the possible justifications for the highlighted points, we can consider that the determination of the scientific names of the species in the flora is more complicated, requiring knowledge of taxonomy. In the case of geographical coordinates, should be considered that a significant number of collections is old and the collector did not have sophisticated equipment, such as GPS.

In addition to the problem of data quality (Chapman 2005b), the handling of large volumes of data also has been the subject of studies (Howe *et al.* 2008). Such difficulties led to the search for alternative methodologies, such as data mining, one of the stages of the process of Knowledge

[1] Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, R. Pacheco Leão 915, 22460-030, Jardim Botânico, Rio de Janeiro, RJ, Brasil.
[2] Authors for correspondence: estevao@jbrj.gov.br; felipealves@jbrj.gov.br

Discovery in Databases (KDD) (Fayyad *et al.* 1996), which aims to discover patterns in databases. The KDD process has been used in various areas such as marketing, medicine, economics, engineering, management, agriculture, social networks, geography, and the Earth Sciences (Han *et al*. 2011).

The costs of the development of these databases is high, for example, the financial value in organizing field trips, the storage of plant specimens in the herbarium, the computational expense of equipment, specialized staff for the development and support of information systems. So, we conclude that the values are significant and justify actions, both in the pursuit of improved quality of data from these collections, like more efficient forms of access. We highlight three points for discussion: 1) Does the quality of the data accompanying the increasing amount of data available in flora databases? 2) Information systems used in herbaria are preventing the input of new data with errors and saving time in the necessary corrections? 3) Is it possible to check the entry of new mistakes or what can be done to reduce the number of errors in the data?

Material and Methods

The use of spreadsheets for the inclusion of data from specimens in herbaria databases is a ubiquitous option for many botanists. Inserting one record per line, separating the different attributes on columns, is a mapping like the format of books used in herbaria for registration of collections. The tabular structure of the spreadsheets is so typical that some software has layouts that refer to them, for example, Brahms (<http://herbaria.plants.ox.ac.uk/bol>). Therefore, maintaining the user-friendliness for the end user was the main objective and, thus, the data import model was maintained with the use of spreadsheets in the development of the information system. Besides, the use of editing options, such as copy-and-paste, drag, replacement values throughout the document, allow the user to type of data faster and efficiently. In addition to the above, and considering that a field expedition obtains dozens of plant specimens, the inclusion of records through a form is a tiresome activity and can be performed in a more agile way with a spreadsheet.

This article presents a tool whose primary objective is to analyze data from new collections

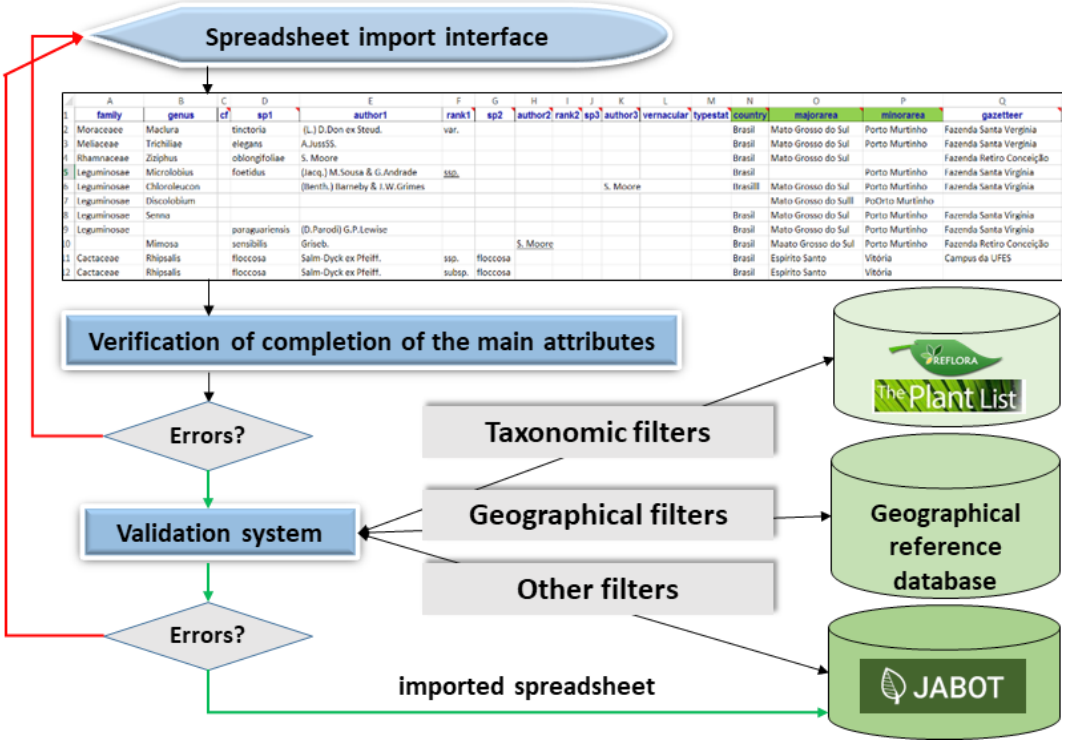

Figure 1 – System macro vision of importing spreadsheets - the green lines indicate that the result of the validation was correct and in red show the flow of errors, noting that a new test is necessary

(Chapman 2005a). This software is part of the management system of scientific collections know as *Jabot* (Silva *et al*. 2017). As a premise that currently there are computational resources sufficient to eliminate new entry errors. The import tool presented in a macro way in Figure 1 receives data in spreadsheets.

During the import process, the tool applies 81 filters to identify errors in the primary occurrence data. We divided the validations into three categories according to a study conducted to evaluate the major types of errors encountered in data of scientific collections of flora. The validations for each class are described in more detail next.

Taxonomic

The taxonomic taxa informed are verified with official lists (Kennedy *et al*. 2005) present in systems such as Flora 2020 (<http://floradobrasil.jbrj.gov.br>), for species that occur in Brazil and, The Plant List (<http://www.theplantlist.org/>) for non Brazilian species. The main errors in this category are typos, caused by lack of training in the area and the difficulty of reading old identification tags in vouchers, among other reasons, justifying the use of lists like dictionaries. In this feature, the tool makes a comparison with each part of the scientific name with the taxa of Flora 2020, as well as the full name. The system allows when configured for Jabot, such automatic replacement and use of the name of the author of the taxon by official scientific name. This category is the most advanced currently with exciting works supported by taxonomic tools like *Taxamatch* (Rees 2014), a system for approximate string matching in taxonomy, but not used in this project, having as justification the fact of the names be compared directly with the Flora 2020.

Geographical

The georeferencing is very important for a variety of researchers, and there is a quest for accuracy in the geographical location in collections (García-Roselló *et al.* 2015). In the process of validating the quality of spatial data, before the validation itself, the tool checks if the values entered for degrees, minutes and seconds are at their valid intervals. Once checked the tool converts the values to decimals and compared with the raster of districts limits contained in the vector basis BC250-IBGE2014 (Azevedo & Neto 2011), for it makes use of the features available at the extension PostGIS (<www.postgis.net>) of the Postgresql (<www.postgresql.org>) database. The values of the reported coordinates are then compared with the values of the coordinates of the official geopolitical units. In case of error, the import tool displays an alert to the user indicating the localization of the point. If the point is in Brazil, the second part of the validation uses the values previously validated for the geopolitical unit. If the geopolitical unit (state or district) informed by the user is different from that shown in the map of Brazilian Institute of Geography and Statistics - IBGE (<www.ibge.gov.br>), the system suggests the correct name. If necessary, the user can see a map showing all collection points of that excursion facilitating the review of the collection points.

Miscellaneous

This is the general errors found on specimen data, this category contains filters to identify those caused by the lack of standardization in the names of the collectors, errors in dates of collection and incorrectly filled fields, for example, altitude values containing the unit of measure in different formats. The identification of the collector and collection number are essential for finding duplicates in other

Table 1 – Fields and validations used in tool. The table displays the name of the primary fields used, are also presented the field definition, and validation carried out. In some cases, more than one validation is necessary. Finally, the system shows if the field is required or not.

Field	Description and validation	Mandatory
family	name of family or *indeterminate* if unknown / verifies that the name is on the official lists taxonomic and in the Jabot	yes
genus	name of genus / verifies that the name is on the official lists taxonomic and in the Jabot	family name
species	name of the species / verifies that the name is on the official lists taxonomic and in the Jabot	family and genus name

Field	Description and validation	Mandatory
author	verifies that the name is on the official lists taxonomic and in the Jabot / for the epithet informed, in case of divergence the informed, displays the name of the author listed on Flora2020 / the user does not have to inform the author because the system populates the value from the Flora2020	family, genus, and species
rank1	in addition to the individual cited checks, checks the complete taxonomic lists taxon / verifies whether the field is in the dictionary as a variety or form	family, genus, species, and author
country	checks for the existence of the name of the country in the Jabot	yes to new samples
majorarea	name of the State or federal unit / checks for the existence of the name of the major area in the Jabot	yes to new samples
minorarea	name of the city or municipality / checks for the existence of the name of the country in the Jabot	yes to new samples
lat_grau, lat_min, lat_seg, ns, long_grau, long_min, long_seg, ew	latitude in degrees, latitude in minutes, latitude in seconds and north or south / longitude in degrees, longitude in minutes, longitude in seconds and east or west / checks whether the coordinates informed match the coordinates present in the geographical reference base for the geopolitical unit informed / checks whether lat_grau is less than or equal to 90 / checks whether long_grau is less than or equal to 180 / checks whether lat_min and long_min are less than or equal to 60 / checks whether lat_seg and long_seg are less than or equal to 60 / checks valid values for the ns and ew fields / remove symbols of the coordinate fields (degree, minute, and second), if they have been informed	
nrdups	number of duplicates / checks if the field is numeric	
collyy	year of determination / checks if the field is numeric / checks if the field has 4 digits / verifies that the year of testimony is less than or equal to the year of collection / checks whether the year of testimony is less than or equal to the current year	yes to new samples
collmm	month of testimony / checks if the field is numeric / checks if the field has 2 digits / verifies that the month of testimony is less than or equal to 12 / automatically converts from roman to Arabic	yes to new samples
colldd	day of determination / checks if the field is numeric / checks if the field has 2 digits / verifies that the day of testimony is less than or equal to 31	yes to new samples
detyy	year of determination / checks if the field is numeric / checks if the field has 4 digits / verifies that the year of THE specimen is less than or equal to the year of collection / checks whether the year of testimony is less than or equal to the current year	
detmm	the month of determination / checks if the field is numeric / checks if the field has 2 digits / automatically converts from roman to Arabic	
detdd	day of determination / checks if the field is numeric / checks if the field has 2 digits / check that the determination date has been reported without the collection date	
collector	name of the collector / verifies that the name of the main collector appears as additional collector	yes to new samples
number	number of the sample / checks whether the collection number already exists for the collector in specific collection / checks if the collection number already exists for the collector in your own worksheet	yes to new samples
altprof	this field stores the unit of altitude or depth minimum of the sample / / checks if the field is numeric	

Field	Description and validation	Mandatory
altprofmax	this field stores the unit of altitude or depth maximum of the sample / in case of informing altprofmax and not informing altprof, in this case the system replaces one field with the other automatically / checks if the field is numeric	
unidmedalprof	/ verifies if value is in the list of units of measures allowed	yes if the altprof field was informed
unidmedaltura	the field stores the height of the specimen	yes if the height field was informed

systems, for example. The tool also identifies and prevents duplicate collection entry, whereas this occurs mainly with large amounts of data. Table 1 gives the primary fields, their description, the validations and if the attribute is required or not.

The import and validation tool is available to the public as a way to promote the improvement of data quality, through the link Jabot: <http://jabot.jbrj.gov.br/v2/validarplanilha_externo.php>. Figure 2 presents the result of a parsed by the spreadsheet tool. The mechanism indicates the line and the type of error, to speed up the process of review of the errors encountered. In the case of the name of the author, the system suggests the name found on Flora 2020.

Results and Discussion

Information systems often generate unexpected difficulties for its users, one of the leading complaints is that related to the interface of the system. Many long forms require more time for the user to enter data into the system. So, one of the perceived advantages of spreadsheet data entry is that directly related to the speed of data entry. Considering the cost of hiring labor for the typing of samples, this can lead to a considerable reduction in the values of the project (Gonzalez 2009). The automatic check with the official lists saves a lot of time the researcher who could only do this comparison individually, *i.e.*, name by name.

The experience of the use of the tool on the system Jabot (<http://jabot.jbrj.gov.br>) of Rio de Janeiro Botanical Garden Research Institute has shown that users understand the process and consider the use of the tool as a resource to speed up the work of inclusion of data in the information system and subsequent printing of labels.

Regarding the elimination or reduction of the amount of data entry errors in the system, the tool proved to be very efficient. Users used import tool

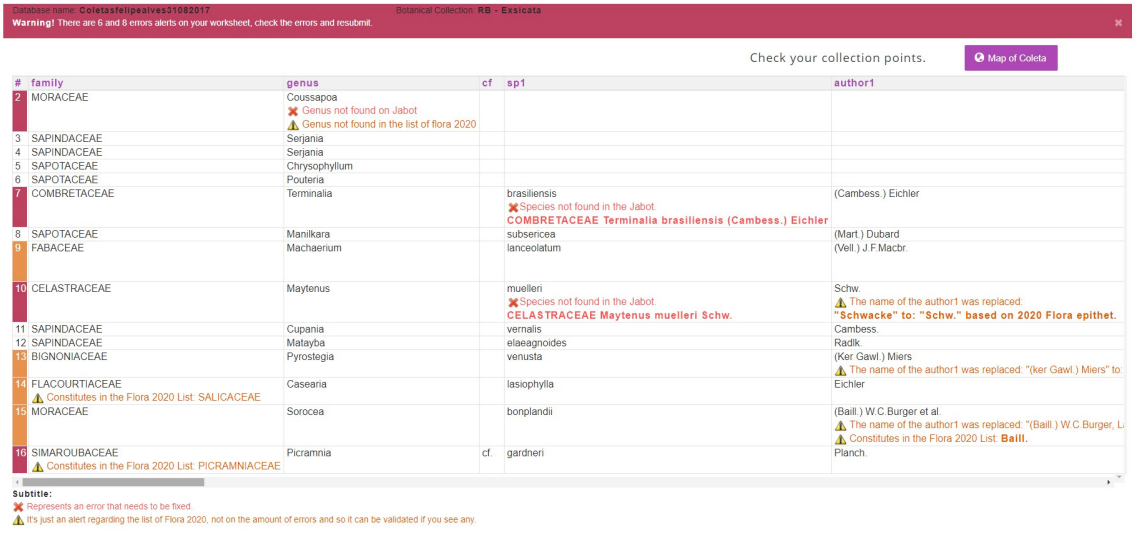

Figure 2 – Results of the validation screen of a spreadsheet collection.

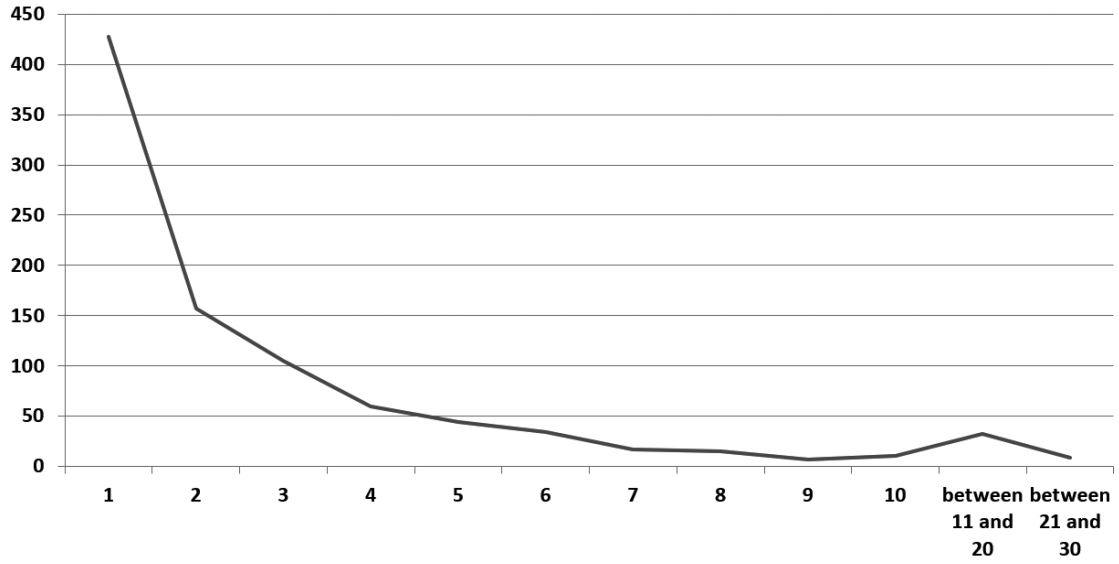

Figure 3 – Distribution of total imported sheets according to the number of attempts required to complete the import process.

934 times and performed 3,548 attempts until data were fully validated, what represents on average 3.7 tries the spreadsheet. Only 428 imports were completed on the first try, representing a rate of 46%. Figure 3 shows a chart with the distribution of the number of imported spreadsheets according to the number of attempts required.

One can consider that as the user was getting acquainted with the system, the number of attempts has been decreasing. In the first two years, the use of the system has doubled, clearly indicating that the resource has become a tool of easy use, allowing you to streamline the work of the collector in the import of samples. The primary factor to achieve this goal was the study and creation of filters to the key fields used in the spreadsheet.

Conclusions

Much attention has been given to the analysis of errors contained in the databases, but the systems do not adequately prevent the entry of new data with low quality. Even with various techniques to assess the quality of data, the time required for cleaning of these data has a high cost to the herbaria and publishers. The user should be aware that other researchers also use the data inputs. Even if the researcher does not need precision in the coordinates in his work, he should keep in mind that this data will be used by other researchers for studies as ecological studies, conservation, predictive modeling and climate change, among others.

The motivating factor for their use is that the tool act as a feature that streamlines the import of the specimens, performing batch checks on spreadsheets, preventing the user from having to search the scientific names individually with official sources and suggesting corrections.

In the current phase of the development of the tool, data mining is being evaluated to identify outliers in collections and standardization of names of collectors (Silva 2016). The association analysis identified suspicious names of collectors. Even though it's a tool that requires a user's extra time for correction of errors in the data, the experience in the first three years of use, leads to the conclusion that the user has a quick adaptation to its application.

Acknowledgements

Thanks to Applied Research Board at Universidade Estácio de Sá for support this work.

References

Azevedo JB & Scandar Neto WJ (2011) Índice de nomes geográficos - Base cartográfica contínua do Brasil ao milionésimo - BCIM. Available at <http://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bcim/versao2016/informacoes_tecnicas/documentacao_tecnica/DocTecnica_BCIM_VOL_I_03nov16.pdf>. Access on 31 August 2017.

Chapman AD (2005a) Principles and methods of data cleaning: primary species and species-occurrence data. Available at <http://www2.gbif.org/DataCleaning.pdf>. Access on 31 August 2017.

Chapman AD (2005b) Principles of data quality. Available at < https://assets.contentful.com/uo17ejk9rkwj/2gupj7dJIw62UeOUYiqSsm/0a4bb732bd7fd8cf28f7703dc20a43ba/Data_Quality_-_ENGLISH.pdf >. Access on 31 August 2017.

Donaldson JS (2009) Botanic gardens science for conservation and global change. Trends in plant science, v. 14, n. 11, p. 608–13. Available at <http://www.cell.com/article/S1360138509002076/fulltext>. Access on 31 August 2017.

Fayyad U, Piatetsky-Shapiro G & Smyth P (1996) Knowledge discovery and data mining : towards a unifying framework. *In*: Simoudis E, Han J & Fayyad U (eds.) KDD-96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI, Portland. Pp. 82-88. Available at <http://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf>. Access on 31 August 2017.

García-Roselló E, Guisande C, Manjarrés-Hernández A, González-Dacosta J, Heine J, Pelayo-Villamil P, González-Vilas L, Vari RP, Vaamonde A, Granado-Lorencio C & Lobo JM (2015) Can we derive macroecological patterns from primary global biodiversity information facility data? Global Ecology and Biogeography 24: 335-347.

GBIF - Global Biodiversity Information Facility (2010) Natural History, Scripta Botanica Belgica. Vol. 29, March, p. 1-2. Available at <http://www.gbif.org/>. Access on 31 August 2017.

Gonzalez M (2009) Quantificação de custo e tempo no processo de informatização das coleções biológicas brasileiras: a experiência do herbário do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Rodriguésia 60: 1-11. Available at <http://rodriguesia.jbrj.gov.br/FASCICULOS/rodrig60_3/014-09a.pdf>. Access on 31 August 2017.

Han J, Kamber M & Pei J (2011) Data mining: concepts and techniques. 3ª ed. Morgan Kaufmann, San Francisco. Pp 5-14.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O & Rhee SY (2008) Big data: the future of biocuration. Nature 455: 47-50.

Kennedy J, Kukla R & Paterson T (2005) Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. *In*: Ludäscher B & Raschid L (eds.) Data integration in the life sciences. Vol. 3615. Springer, Berlin, Heidelberg. Pp. 80-95. Available at <http://researchrepository.napier.ac.uk/3027/>. Access on 31 August 2017.

Lavoie C (2013) Biological collections in an ever changing world: herbaria as tools for biogeographical and environmental studies. Perspectives in Plant Ecology, Evolution and Systematics 15: 68-76.

Rees T (2014) Taxamatch, an algorithm for near ('Fuzzy') matching of scientific names in taxonomic databases. PLoS ONE 9: 23.

Silva LAE, Fraga CN, Almeida TMH, Gonzalez M, Lima RO, Rocha MS, Bellon E, Ribeiro RS, Oliveira FA, Clemente LS, Magdalena UR, Medeiros EVS & Forzza RC (2017) Jabot - Botanical Collections Management System: the experience of a decade of development and advances. Rodriguésia 68: 391-410. Available at <http://rodriguesia.jbrj.gov.br/FASCICULOS/rodrig68-2/08-0162-2016.pdf>. Access on 31 August 2017.

Silva LAE (2016) A data mining approach for standardization of collectors names in herbarium database. IEEE Latin America Transactions 14: 805-810.

Wen J, Ickert-Bond SM, Appelhans MS, Dorr LJ & Funk VA (2015) Collections-based systematics: opportunities and outlook for 2050. Journal of Systematics and Evolution 53: 477-488. Available at <http://dx.doi.org/10.1111/jse.12181>. Access on 31 August 2017.

Editor de área: Dr. Marcelo Trovó

Artigo recebido em 31/08/2017. Aceito para publicação em 26/12/2017.