Challenges in data cleaning and normalization of natural history collections

Fred Stauffer, Curator - Herbarium of phanerogamy (Conservatoire et Jardin botaniques de Genève)

Hélène Gabioud-Duinat, Head of Inventories (Musée de la nature, Sion)

1. Why is data cleaning so important?
2. How does data cleaning works?
3. Different approaches for data cleaning
4. Who oversees it and what skills are needed?
5. Exchange of experiences

Conservatoire et Jardin botaniques Genève

Musée de la nature du Valais
Naturmuseum Wallis

scnat
akademie der naturwissens

Why is data cleaning so important?

The increase in online and openly accessible biodiversity databases provides a **vast and invaluable resource** to support research and policy. However, without scrutiny, **errors** in primary species occurrence data can lead to erroneous results and **misleading information** (Ribeiro et al. 2022)

Data cleaning **is a necessary first step in any analysis** that involves data from integrated biodiversity databases. The goal of data cleaning is to **detect inaccurate, unreasonable, or incomplete data** and try to correct them (García-Rosello et al., 2014).

The **importance of primary species occurrence data for many biodiversity applications is evident**, yet they have **limitations**, and their **quality can vary substantially** (Meyer et al., 2016).

Data are fundamental for research and practices in biodiversity conservation. However, **data quality issues associated with biodiversity data have to be addressed before** we can use them with confidence (Jin & yang, 2020)

Issues related to **difficulty standardizing data from different sources** (Kissling et al., 2018), discrepancies and errors in taxonomic and nomenclatural data (e.g. Mesibov, 2013; Nic Lughadha et al., 2019), and **errors** and inaccuracies in geographical and temporal information of primary species occurrence data (e.g. Meyer et al., 2016) **can lead to erroneous results and misleading information** (Maldonado et al., 2015; Nic Lughadha et al., 2019; Zizka et al., 2020).

Significant **challenges remain, especially when assembling large and heterogeneous databases** from online aggregators (Chapman, 2005b; Kissling et al., 2018).

Different approaches for data cleaning

Manual data cleaning: time-consuming, error prone, difficult to reproduce and limited to known taxonomic groups and geographical areas, making it impractical for datasets with numerous records.

Automized data cleaning: Less time-consuming, error free, etc……but are we able to start now with this?


APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro | Santiago José Elías Velazco | Karlo Guidoni-Martins | Geiziane Tessarolo | Lucas Jardim | Steven P. Bachman | Rafael Loyola

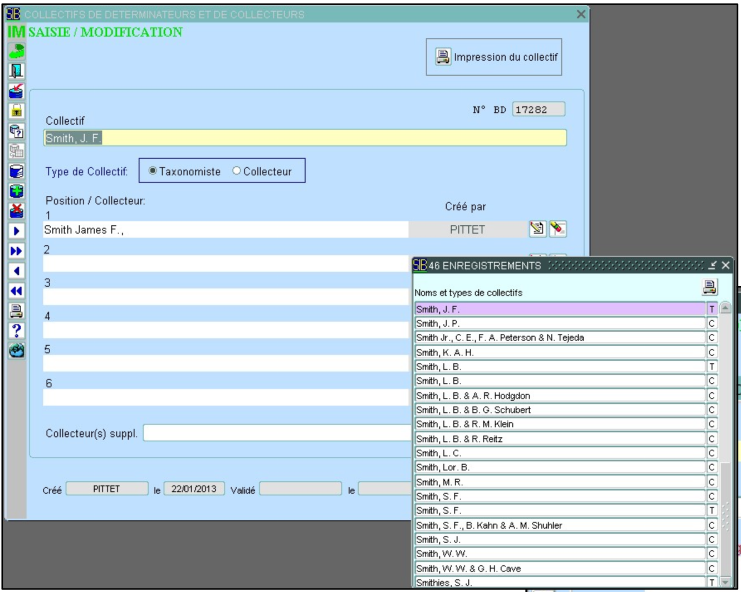
Global Ecology and Conservation

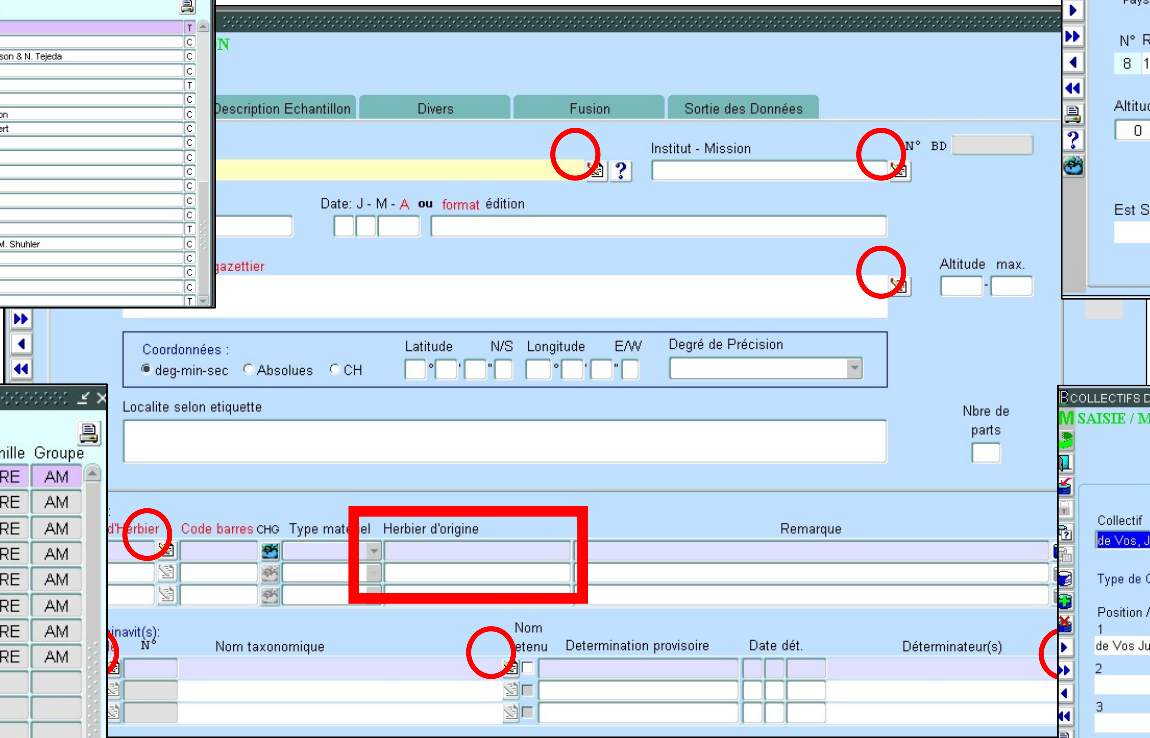
Original Research Article

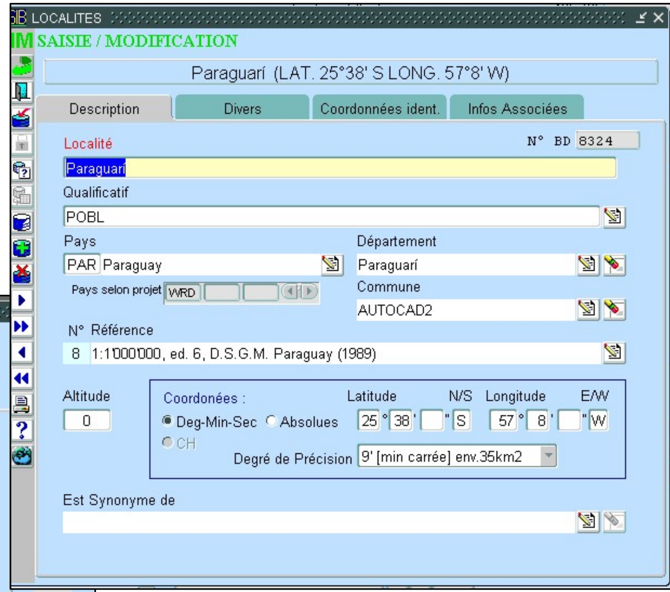
BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

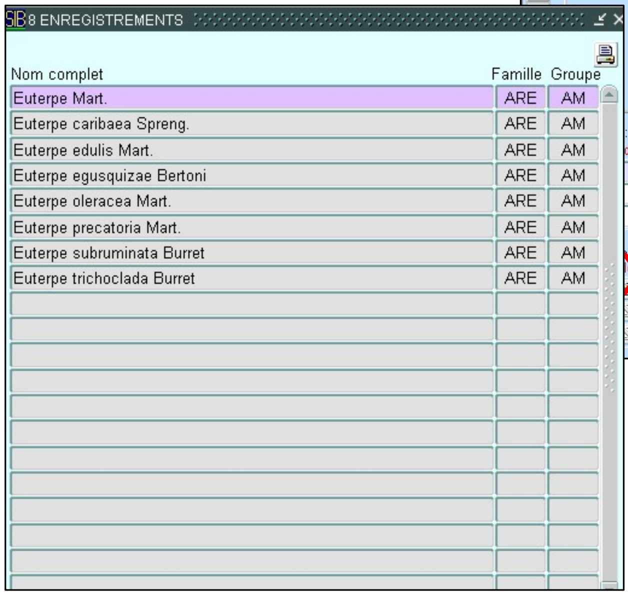
Jing Jin, Jun Yang

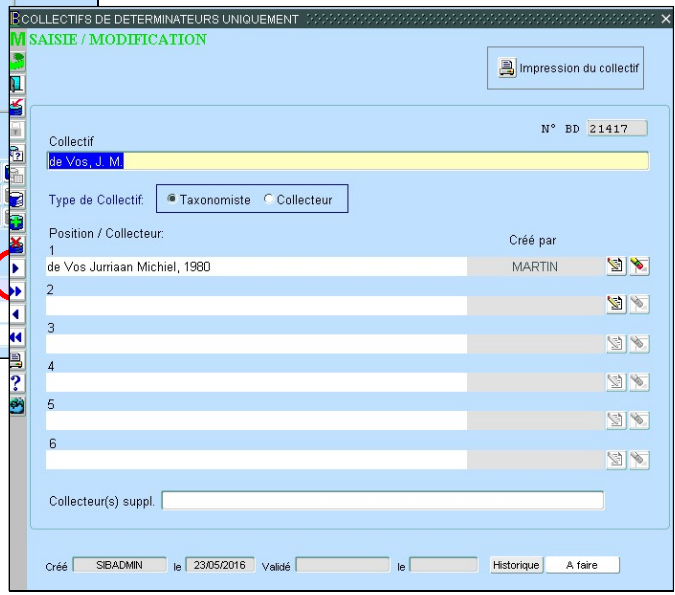
Thesaurus and gazetteers as a good starting point



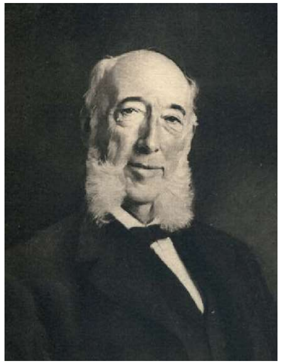


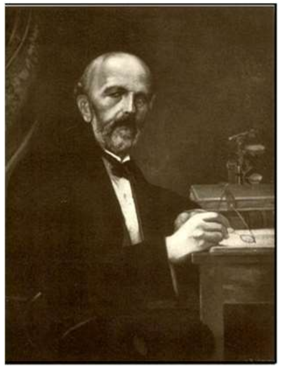






Keeping track of the history of our collections





Plantes provenant

de l'**HERBIER** du **MUSÉE D'HISTOIRE NATURELLE DES GRISONS (Coire),**

données au Conservatoire botanique en 1995,
et intercalées dans la Collection générale dès 1996.

Séries contenant de nombreux échantillons récoltés par Alexander Moritzi (1806-1850, élève d'A.-P. de Candolle), en particulier dans les anciens *Jardin botanique de Soleure et Jardin botanique de Genève (Bastions).*

HERBIER
du Musée d'histoire naturelle des Grisons (Coire)

Donné au Conservatoire botanique de la Ville de Genève en 1995
et intercalé dès 1996.

Plantes provenant de l'herbier du Professeur
ROBERT CHODAT

Intercalées dans l'herbier général du Conservatoire botanique de la Ville de Genève en 1970.

HERBIER BARBEY-BOISSIER

Constitué par William Barbey, après la mort de son beau-père Edmond Boissier (1885); donné en 1918, par les enfants de W. Barbey, à l'Université de Genève où il fut augmenté par différents collaborateurs; transféré en 1944 au Conservatoire botanique.

Conservatoire botanique, Genève

Herbier **BOISSIER**, séries n'ayant pas servi à la rédaction du

Flora Orientalis

HERBARIUM GENAVENSE, G

HERBIER MARCEL GUSTAV BAUMANN-BODENHEIM

Plantes provenant de la collection personnelle de M. G. Baumann-Bodenheim, léguée aux Conservatoire et Jardin botaniques de Genève en 1997.

Intercalé dans l'herbier général en ....

HERBIER D[r] A. HUBER-MORATH
(1901-1990)

Légué par l'auteur au Conservatoire botanique et intercalé dans la collection générale dès 1991.

HERBIER EMIL HASSLER
Plantae Paraguarienses

Herbier personnel du Dr Emil Hassler (1864-1937), constitué de plantes récoltées entre 1885 et 1919 au Paraguay et dans les régions adjacentes de l'Argentine, du Brésil et de la Bolivie. Il a été déposé aux Conservatoire et Jardin botaniques de la Ville de Genève en 1919 et intercalé dans la Collection générale à partir de 1955.

HERBIER MICHEL DESFAYES
(Saillon - VS, Suisse)
(1927 -)

Plantes aquatiques et palustres de Suisse, de l'Europe méridionale et de l'hémisphère sud - Collection cécidologique

Donné au Conservatoire botanique de la Ville de Genève en 2020 et intercalé dans l'Herbier général dès 2021.

HERBIER DE BERTRAM V.D. POST

Herbier de Bertram V.D. Post (1871-1960), donné au Conservatoire botanique de Genève en 1956.

HERBIER PHILIPPE DE PALEZIEUX
(1871-1957)

Herbier personnel, légué en 1957, à l'Institut de botanique générale de l'Université de Genève. Transféré aux Herbiers de la Ville de Genève, il fut intercalé dans la collection générale dès 1966.

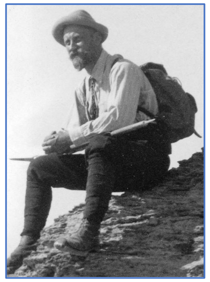
HERBIER HENRY CORREVON (1854-1939)

Collections données par la famille Correvon aux Conservatoire et Jardin botaniques de la Ville de Genève en 1983 et intercalées dès 1984

HENRI PABOT

Plantes de la Syrie et du Liban récoltées de 1952 à 1958.

Collection originale acquise par le Conservatoire botanique de la Ville de Genève en 1972 et intercalée dans l'Herbier général dès 1974.



Messy – heterogeneous data: 1677 records to be standardized

	HERBIER ORIGINE	HERBIER	Nb Echant.	Herbier d'origine: nouvelle valeur	Année d'intercalation
1052	Herb. Reg. Berolinense	PHANERO	1		
1057	herb. Reuter-Barbey	PHANERO	9	Reuter-Barbey	
1089	Herb. Thuillier	PHANERO	1	Thuillier	
1175	Herbario Nacional Colombiano	PHANERO	2		
1177	Herbario Willdenow	PHANERO	1		
1263	Herbier Dr Huber-Morath	PHANERO	1	Huber-Morath	
[illegible]	Herbier Barbey-Boissier	PHANERO	[illegible]	Barbey-Boissier	
1267	Herbier Boisier (ainsi que pécisé par Reuter dans le prot.)	PHANERO	1	Boissier	
1269	Herbier Boissier	PHANERO	62	Boissier	
1270	Herbier Boissier (Fl. Orient.)	PHANERO	9	Boissier	
1271	Herbier Boissier et Barbey Boissier	PHANERO	1	Boissier et Barbey-Boissier	
1272	Herbier Boissier et Barbey-Boissier	PHANERO	6	Boissier et Barbey-Boissier	
1273	Herbier Boissier Fl. Orient.	PHANERO	1	Boissier	
1274	Herbier Bossier	PHANERO	1	Boissier	
1275	Herbier Burnat	PHANERO	6	Burnat	
1277	Herbier Chodat	PHANERO	4	Chodat	
1280	Herbier De Candolle	PHANERO	28914	De Candolle	
1282	Herbier de Moise-Etienne Moricand	PHANERO	12	Moricand	
1283	Herbier de Moïse-Etienne Moricand	PHANERO	2	Moricand	
1284	Herbier de Ventenat	PHANERO	2	Ventenat	
1285	Herbier Delesert	PHANERO	1	Delessert	
1286	Herbier Delessert	PHANERO	140	Delessert	
1287	Herbier Delessert 1931	PHANERO	1	Delessert	1931
1288	Herbier Delessert 1936	PHANERO	1	Delessert	1936
1289	Herbier Dellesert	PHANERO	1	Delessert	
1290	Herbier Dr A. Huber-Morath	PHANERO	1	Huber-Morath	
1291	Herbier Dr Huber-Morath	PHANERO	105	Huber-Morath	
1292	Herbier Dr Huber--Morath	PHANERO	1	Huber-Morath	
1293	Herbier Dr. A. Huber-Morath	PHANERO	1	Huber-Morath	
1294	Herbier Dr. Huber-Morath	PHANERO	4	Huber-Morath	
1295	Herbier Dr. Hubert-Morath	PHANERO	1	Huber-Morath	
1302	Herbier G.	PHANERO	3		
1303	Herbier général 1978	PHANERO	1		1978
1305	Herbier Hassler	PHANERO	1	Hassler	
1310	Herbier Huber-Morath	PHANERO	4	Huber-Morath	
1317	Herbier M.-E. Moricand	PHANERO	6	Moricand	
1318	Herbier Moise-Etienne Moricand	PHANERO	1	Moricand	

What informatic tools are nowadays available?

Global Ecology and Conservation 21 (2020) e00852

Contents lists available at ScienceDirect

Global Ecology and Conservation

journal homepage: http://www.elsevier.com/locate/gecco

Original Research Article

BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

Jing Jin [a], Jun Yang [a, b, *]

[a] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China
[b] Joint Center for Global Change Studies, Beijing, 100875, China

ARTICLE INFO

Article history:
Received 4 October 2019
Received in revised form 18 November 2019
Accepted 18 November 2019

Keywords:
Data quality
Biodiversity
Data cleaning
Tree species
Conservation

ABSTRACT

High-quality data are indispensable for research and management in biodiversity conservation. Nevertheless, errors in biodiversity data must be removed before they can be used with confidence. In this study, we have developed a workflow for cleaning occurrence data archived in various biodiversity databases. The workflow allows researchers and practitioners to identify taxonomic and geographic errors in millions of records in an automatic, reproducible, and transparent manner. It also allows users to correct several types of taxonomic and geographic errors. We applied the workflow to clean global tree occurrence records. The results showed that among the 30,242,556 occurrence records of 58,034 species extracted from eight databases, only 8,624,319 (28.5%) records of 22,766 (39.2%) species were classified as high quality after running through the workflow. Inaccurate and non-standard taxon names appeared as a more severe problem than geographical errors that people are most familiar with. The workflow developed in this study can be easily adapted to clean occurrence records of other taxonomic groups, which allows researchers and practitioners to reduce uncertainties in their findings.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Biodiversity is critical for the sus
to unravel the pattern of biodiversit
answer important questions such as
(Sutherland et al., 2013). Biodiversi
Besides, biodiversity data are indis
these reasons, the demand for high

The exponential growth of data c
herbaria, and field surveys) availab
Organizations and individuals have
occurrence records. Global Biodiver

* Corresponding author. 5721 Mengmingw
E-mail addresses: jinj15@mails.tsinghua

https://doi.org/10.1016/j.gecco.2019.e00852
2351-9894/© 2019 The Authors. Published b
0/).

Jing, J. & J. Yang; 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation* 21: 1-11.

Received: 7 March 2022 | Accepted: 30 March 2022
DOI: 10.1111/2041-210X.13868

Methods in Ecology and Evolution — BRITISH ECOLOGICAL SOCIETY

APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro[1] | Santiago José Elías Velazco[2,3,4] | Karlo Guidoni-Martins[1] | Geiziane Tessarolo[5] | Lucas Jardim[6] | Steven P. Bachman[7] | Rafael Loyola[8,9]

[1]Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Brazil; [2]Department of Botany and Plant Sciences, University of California—Riverside, Riverside, CA, USA; [3]Instituto de Biologia Subtropical, Universidad Nacional de Misiones – CONICET, Puerto Iguazú, Argentina; [4]Programa de Pós-Graduação em Biodiversidade Neotropical, Universidade Federal da Integração Latino-Americana (UNILA), Foz do Iguaçu, Brazil; [5]Universidade Estadual de Goiás, UEG, Campus de Ciências Exatas e Tecnológicas – CCET, Anápolis, Brazil; [6]Instituto Nacional de Ciência e Tecnologia em Ecologia, Evolução e Conservação da Biodiversidade, Universidade Federal de Goiás, Goiânia, Brazil; [7]Royal Botanic Gardens, Kew, Richmond, UK; [8]Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Brazil and [9]International Institute for Sustainability, Rio de Janeiro, Brazil

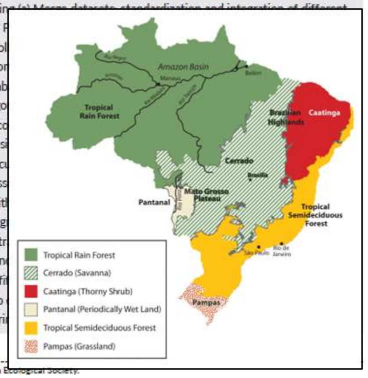
Correspondence
Bruno R. Ribeiro
Email: ribeiro.brr@gmail.com

Funding information
Argentine National Council of Scientific and Technological Research; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 465610/2014-5, 201810267000023 and 306694/2018-2; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 001; National Science Foundation, Grant/Award Number: 1853697

Handling Editor: Samantha Price

Abstract

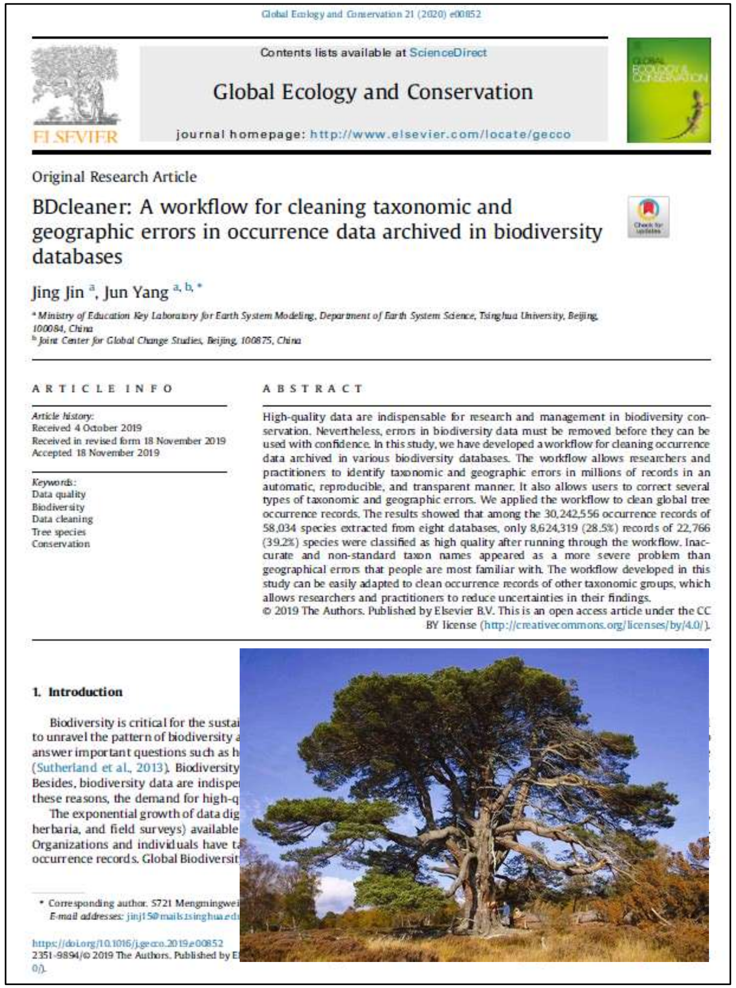
1. The increase in online and openly accessible biodiversity databases provides a vast and invaluable resource to support research and policy. However, without scrutiny, errors in primary species occurrence data can lead to erroneous results and misleading information.
2. Here, we introduce the Biodiversity Data Cleaning (*bdc*), an R package to address quality issues and improve the fitness-for-use of biodiversity datasets. The *bdc* package brings together several aspects of biodiversity data cleaning in one place. It is organized in thematic modules related to different biodiversity dimensions, includin
datasets; (b) P
formation, fol
harmonization
onomic datab
matching algo
geographic co
tion of inconsi
visualize, docu
quality assess
functions wit
can also integ
3. We demonstr
lion occurren
around one-fi
4. Compared to
are that it bri



© 2022 The Authors. Methods in Ecology and Evolution © 2022 British Ecological Society.

Ribeiro, B. R., Velazco, S. J., 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: 1421–1428.

BDcleaner



Global Ecology and Conservation 21 (2020) e00852

Contents lists available at ScienceDirect

Global Ecology and Conservation

journal homepage: http://www.elsevier.com/locate/gecco

ELSEVIER

Original Research Article

BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

Jing Jin [a], Jun Yang [a, b, *]

[a] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China

[b] Joint Center for Global Change Studies, Beijing, 100875, China

A R T I C L E I N F O

Article history:
Received 4 October 2019
Received in revised form 18 November 2019
Accepted 18 November 2019

Keywords:
Data quality
Biodiversity
Data cleaning
Tree species
Conservation

A B S T R A C T

High-quality data are indispensable for research and management in biodiversity conservation. Nevertheless, errors in biodiversity data must be removed before they can be used with confidence. In this study, we have developed a workflow for cleaning occurrence data archived in various biodiversity databases. The workflow allows researchers and practitioners to identify taxonomic and geographic errors in millions of records in an automatic, reproducible, and transparent manner. It also allows users to correct several types of taxonomic and geographic errors. We applied the workflow to clean global tree occurrence records. The results showed that among the 30,242,556 occurrence records of 58,034 species extracted from eight databases, only 8,624,319 (28.5%) records of 22,766 (39.2%) species were classified as high quality after running through the workflow. Inaccurate and non-standard taxon names appeared as a more severe problem than geographical errors that people are most familiar with. The workflow developed in this study can be easily adapted to clean occurrence records of other taxonomic groups, which allows researchers and practitioners to reduce uncertainties in their findings.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Biodiversity is critical for the sustai
to unravel the pattern of biodiversity a
answer important questions such as h
(Sutherland et al., 2013). Biodiversity
Besides, biodiversity data are indispe
these reasons, the demand for high-q

The exponential growth of data dig
herbaria, and field surveys) available
Organizations and individuals have ta
occurrence records. Global Biodiversit

* Corresponding author. 5721 Mengmingwei
E-mail addresses: jinj15@mails.tsinghua.ed

https://doi.org/10.1016/j.gecco.2019.e00852
2351-9894/© 2019 The Authors. Published by E
0/).

Development of a workflow for cleaning occurrence data archived in various biodiversity databases.

The workflow allows researchers and practitioners to identify taxonomic and geographic errors in millions of records in an automatic, reproducible, and transparent manner.

Case study: Study on global tree occurrence records (30.242.556 occurrence records of 58.034 species extracted from eight databases,

R code for this study is available via the Mendeley Data Repository https://doi.org/10.17632/pghkfm5sm9.1 (Jin and Yang, 2019).

Jing, J. & J. Yang; 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation* 21: 1-11.

Main functions of BD cleaner:

1.- To **integrate** multi-source species occurrence datasets,

2.- To **identify errors in the taxonomic and spatial dimensions** of the data,

3.- To **correct taxonomic and geographic errors** in records instead of simply discarding them.

Most important databases storing occurrence records associated to biological entities:

1.- GBIF: **By mid-2019, GBIF had integrated over 1.2 billion species occurrence records** and **data retrieved yesterday shows that more than 2.6 species occurrence records** are now available from various sources (http://gbif.org/GBIF.org, 2024).

2.- Botanical Information and Ecology Network (BIEN): https://bien.nceas.ucsb.edu/bien/

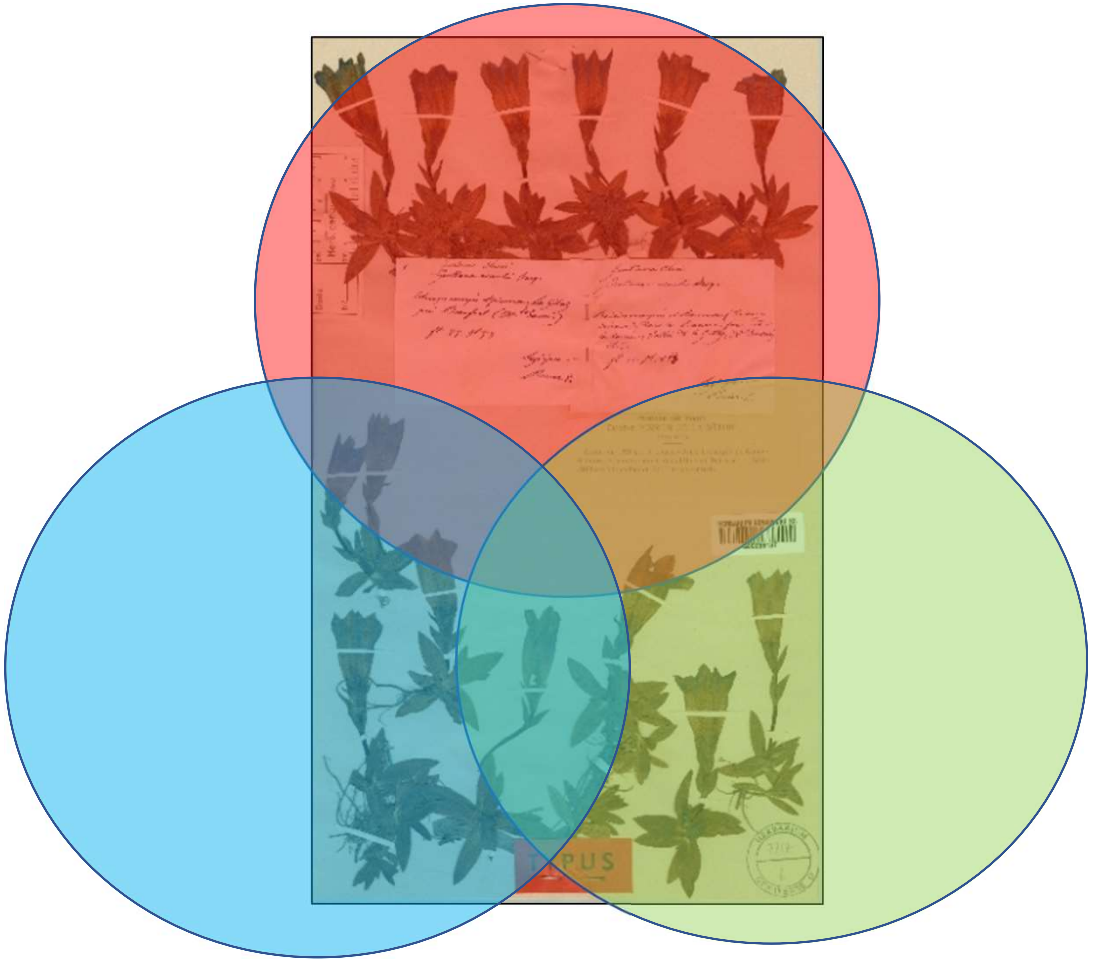
3.- SpeciesLink: https://specieslink.net/

4.- Ebird: https://ebird.org/home

5.- iDigBio: https://www.idigbio.org/

The three dimensions of scientific data of occurrence records : taxonomy, space, time

.



The three dimensions of scientific data of occurrence records : taxonomy, space, time

In the taxonomic dimension, misleading and outdated taxonomy of occurrence data poses a significant challenge to users (Tessarolo et al., 2017). For example, Meier and Dikow (2004) estimated that the rate of specimen misidentification ranges from 5% to nearly 60% in Zoological Record Database.

In the spatial dimension, geographical errors in occurrence records are the most common data quality issue (Otegui et al., 2013; Topel et al., 2017; Yesson et al., 2007), which affects applications such as species distribution modeling significantly.

In the temporal dimension, data collected in an early time period typically have lower quality than data collected in recent times due to the loss of metadata and improvements made in data collection methods as time goes by (Tessarolo et al., 2017).

Uncertainties and biases in each dimension can significantly impact their applications! (Goodwin et al., 2015; Robertson et al., 2016).

Four steps are proposed for a workflow:

Step 1: Integrating occurrence data. Occurrence data from eight datasets were merged: GBIF, BIEN, the Atlas of Living Australia (ALA, 2019), BioTime (Dornelas et al.,2018), RAINBIO (Dauby et al., 2016), the Integrated Digitized Biocollections, SpeciesLink, and Biodiversity Information Serving.

Step 2: Cleaning taxonomic errors. Taxonomic errors in the records were identified and corrected.

Matching the scientific name of occurrence records to the names in The Plant List. For records that did not match with names in TPL, we tried to identify and correct the possible spelling errors in the string of scientific names. The influence of particular punctuation in strings was revolved by automatically removing the punctuation from the strings

Exemple:
***Bauhinia pes-caprae* Cav**. Versus *Bauhinia pes_caprae* Cav., *Bauhinia pescaprae* Cav., *Bauhinia pes=caprae* Cav.

Our results showed that inaccurate and non-standard taxon names of tree occurrence data were the most troubling problem!. Only 66.0% of the occurrence data could match names in The Plant List.

Different plant name databases!
But which one can be really considered as the «backbone»?

IPNI

TROPICOS

WORLD FLORA ON LINE

PLANT OF THE WORLD ON LINE (POWO)

THE PLANT LIST

Step 3: Cleaning geographical errors. Geographic errors in the records were identified and corrected.

For these records with coordinates, we followed Meyer et al. (2016) to use the decimal digits as a proxy to judge the precision of locations. Then we identified several common location errors based on the coordinates and the country code information:

Records whose latitude and/or longitude have zero values were removed. Coordinates in oceans and spatial mismatches between national boundaries and coordinates were considered as location errors.

Step 4: Quality labeling. Ech occurrence record was assigned quality levels in the taxonomic and geographical dimensions.

Different studies have different requirements for data quality. Even species occurrence data with low quality are useful in some situations. For example, the continent-level precision may be sufficient to fit the SDM model for global studies (Zizka et al., 2019).

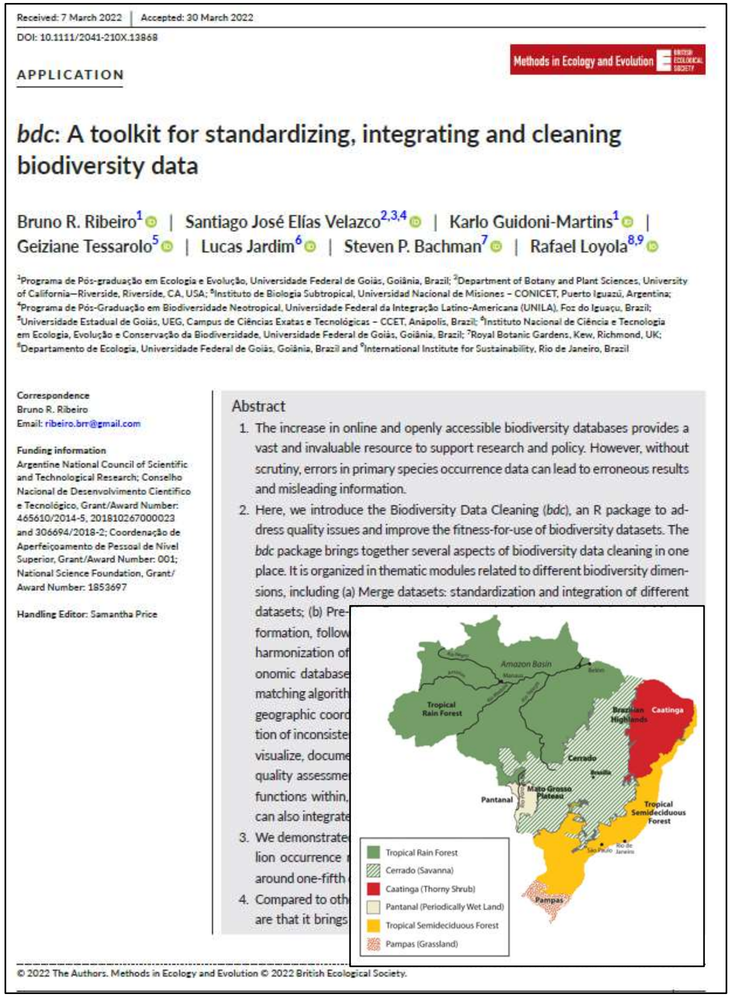
Common geographic errors examined in the workflow and the data sets used for identifying them According to Jin & Yang (2020)

ERROR → **CAUSE** → **SOLUTION**

	Geographic errors	Potential causes	Data used for identifying errors
1	Latitude and/or Longitude is zero	No available coordinates in original record.	No
2	Points are located in oceans	Low precision of the original coordinate; Incorrect sign or position of the x- and y-coordinate.	World map in *maptools* R package (Bivand and Lewin-Koh, 2018)
3	Points are located on land but outside of the boundary of the country where they are reported	Same as 2	World map in *maptools* R package (Bivand and Lewin-Koh, 2018); Natural earth world map (https://www.naturalearthdata.com/downloads/10m-cultural-vectors/); GADM world political boundary (https://gadm.org/data.html); ISO-3166 country code list (https://www.iso.org/obp/ui/#search/code/).
4	Centroids of administrative areas used as coordinates of records	Missing locations of original data when digitization.	GeoNames global cities (https://www.geonames.org/) Global cities and province distribution (Zizka et al., 2019) GADM world political boundary (https://gadm.org/data.html)
5	Coordinates of institutions used as coordinates of records	Same as 4	Global database of biodiversity institutions (Zizka et al., 2019)
6	Botanical Garden	Same as 4	BGCI Garden Search (https://tools.bgci.org/garden_search.php)

Criteria for assigning quality levels - Geographical and taxonomic data (According to Jin & Yang, 2020)

Level	Criteria
Geographical	
High	The number of decimal digits of coordinates≥3, no geographic errors.
Medium	The number of decimal digits of coordinates = 2, no geographic errors.
Low	The number of decimal digits of coordinates = 1, no geographic errors.
Other	No geo-referenced information or with geographic errors that cannot be rectified.
Taxonomic	
High	Species name matches a species name in TPL with a high confidence level.
Medium	Species name matches a species name in TPL with a medium confidence level.
Low	Species name matches a species name in TPL with a low confidence level.
Other	Species name does not match records in TPL or with taxonomic errors that cannot be rectified.
Geographical and Taxonomic	
High	High quality in both taxonomic and geographical dimensions.
Medium	Quality in taxonomic and geographical dimensions are at or above “Medium” and at least one of them is “Medium”.
Low	Quality in taxonomic and geographical dimensions are at or above “Low” and at least one of them is “Low”.
Other	Records that belong to the “Other” category in either spatial or taxonomic dimension.



Received: 7 March 2022 | Accepted: 30 March 2022

DOI: 10.1111/2041-210X.13868

Methods in Ecology and Evolution

APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro[1] | Santiago José Elías Velazco[2,3,4] | Karlo Guidoni-Martins[1] | Geiziane Tessarolo[5] | Lucas Jardim[6] | Steven P. Bachman[7] | Rafael Loyola[8,9]

[1]Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Brazil; [2]Department of Botany and Plant Sciences, University of California—Riverside, Riverside, CA, USA; [3]Instituto de Biologia Subtropical, Universidad Nacional de Misiones – CONICET, Puerto Iguazú, Argentina; [4]Programa de Pós-Graduação em Biodiversidade Neotropical, Universidade Federal da Integração Latino-Americana (UNILA), Foz do Iguaçu, Brazil; [5]Universidade Estadual de Goiás, UEG, Campus de Ciências Exatas e Tecnológicas – CCET, Anápolis, Brazil; [6]Instituto Nacional de Ciência e Tecnologia em Ecologia, Evolução e Conservação da Biodiversidade, Universidade Federal de Goiás, Goiânia, Brazil; [7]Royal Botanic Gardens, Kew, Richmond, UK; [8]Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Brazil and [9]International Institute for Sustainability, Rio de Janeiro, Brazil

Correspondence
Bruno R. Ribeiro
Email: ribeiro.brr@gmail.com

Funding information
Argentine National Council of Scientific and Technological Research; Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, Grant/Award Number: 465610/2014-5, 201810267000023 and 306694/2018-2; Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior, Grant/Award Number: 001; National Science Foundation, Grant/Award Number: 1853697

Handling Editor: Samantha Price

Abstract

1. The increase in online and openly accessible biodiversity databases provides a vast and invaluable resource to support research and policy. However, without scrutiny, errors in primary species occurrence data can lead to erroneous results and misleading information.
2. Here, we introduce the Biodiversity Data Cleaning (*bdc*), an R package to address quality issues and improve the fitness-for-use of biodiversity datasets. The *bdc* package brings together several aspects of biodiversity data cleaning in one place. It is organized in thematic modules related to different biodiversity dimensions, including (a) Merge datasets: standardization and integration of different datasets; (b) Pre- formation, follow harmonization of onomic database matching algorith geographic coord tion of inconsiste visualize, docume quality assessme functions within, can also integrate
3. We demonstrate lion occurrence around one-fifth
4. Compared to oth are that it brings

© 2022 The Authors. Methods in Ecology and Evolution © 2022 British Ecological Society.

BDC Biodiversity **D**ata **C**leaning (*bdc*)

R package (R Core Team, 2020) **OPEN SOURCE!**

The *bdc* package is a toolkit that offers the means to convert raw data into high-quality information through a suite of core functions used to flag, clean, document and enrich data quality

Main advantage: Compared to other available R packages, the main strengths of the bdc package are that it brings together available tools, and a series of new ones, to assess the quality of different dimensions of biodiversity data into a single and flexible toolkit.

Case study: Brazilian Flora 2020 project - Projeto Flora do Brasil 2020 (https://ipt.jbrj.gov.br/jbrj/resource?r=lista_especies_flora_brasil)

Five thematic modules aiming to address quality issues and improve the fitness-for-use of a dataset. It offers a series of new tests and tools developed for 1) validating, 2) documenting and 3) reporting data quality.

1.- Merge datasets: standardization and integration of different datasets;

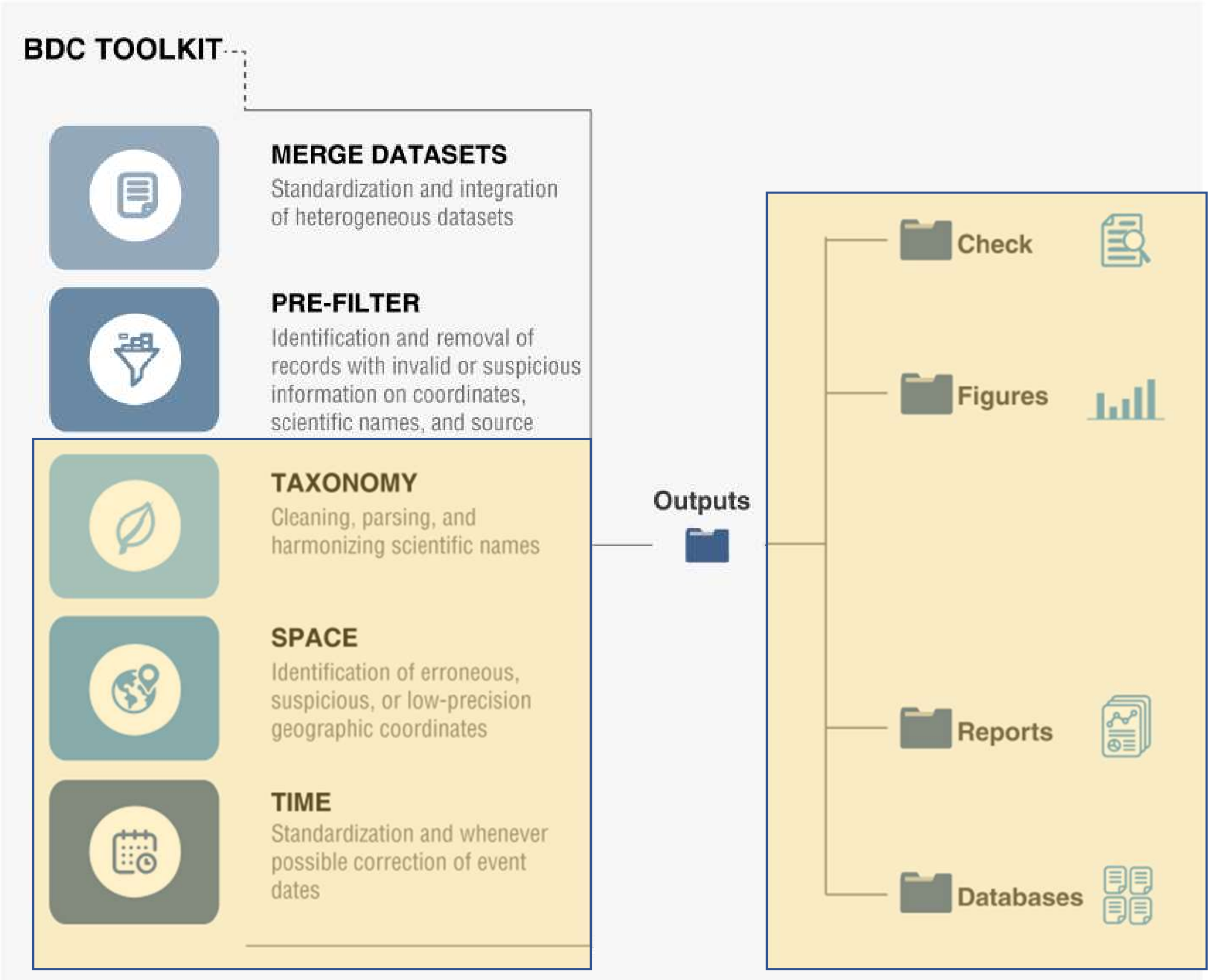
2.- Pre-filter: flagging and removal of invalid or non-interpretable information, followed by data amendments;

3.- Taxonomy: cleaning, parsing and harmonization of scientific names from several taxonomic groups against taxonomic databases locally stored through the application of exact and partial matching algorithms;

4.- Space: flagging of erroneous, suspect and low-precision geographic coordinates;

5.- Time: flagging and, whenever possible, correction of inconsistent collection date.

FIGURE 1 The Biodiversity Data Cleaning (*bdc*) package contains functionalities for standardizing and integrating data from different sources and implements several tests to flag, document, clean and correct biodiversity data. The *bdc* package is organized in thematic modules (merge datasets, pre-filter, taxonomy, space and time). Several outputs documenting the data cleaning process can be saved, including files needing further inspections, figures and reports



From Ribeiro, B. R., Velazco, S. J., 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: 1421–1428. https://doi.org/10.1111/2041-210X.13868

1.- Merge datasets

The lack of terminology standardization makes the integration of large and heterogeneous datasets a challenge.

The function *bdc_standardize_datasets* specifically handles the standardization of heterogeneous datasets. To do so, users must fill in a configuration table (available as Appendix S2 in the paper) to indicate which field names (i.e. column headers) of each original dataset match a list of Darwin Core standard terms (as defined by Wieczorek et al., 2012).

2.- Pre-filter

This module contains functions to flag and remove (a) records missing species names, (b) records missing partial or complete information on geographic coordinates, (c) out-of-range coordinates, (d) records from doubtful sources (e.g. from drawings, photographs or multimedia objects, among others) and (e) records outside a region of interest (for example, out of Brazil), that is, records in other countries or at an informed distance from the coast (e.g. in the ocean).

Bonus: The pre-filter module also includes functions for data enhancement, such as deriving country names from valid geographic coordinates, standardizing country names,

3.- Taxonomic harmonization

The bdc package includes functions to help the taxonomic name harmonization by comparing scientific names against one of 10 taxonomic databases. The taxonomic harmonization uses taxadb package (Norman et al. 2020).

The goal of taxadb is to provide fast, consistent access to taxonomic data, supporting common tasks such as resolving taxonomic names to identifiers, looking up higher classification ranks of given species, or returning a list of all species below a given rank.

Backdraw: misspelled scientific names—commonly found in biodiversity databases—cannot be resolved by an exact matching algorithm, which may result in many unresolved names.

Bonus: To troubleshoot this problem, bdc developed additional functions!

taxadb abbreviation	name
itis	The Integrated Taxonomic Information System, https://www.itis.gov/
col	The Catalogue of Life
ncbi	The National Center for Biotechnology Information
gbif	The Global Biodiversity Information Facility
tpl	The Plant List
fb	FishBase https://fishbase.org
slb	SeaLifeBase
wd	WikiData, (wikidata.org)
iucn	The IUCN Red List of endangered species status, https://www.iucnredlist.org
ott	Open Tree of Life taxonomy.

4.- Space: identification of errors in geographic coordinates

BDC uses *CoordinateCleaner (Open Source)*, an R package based on geographic gazetteers, to flag potential erroneous coordinates **(Zizka et al., 2019)***

CoordinateCleaner is tailored to problems common in biological and palaeontological databases and can handle datasets with millions of records. The software includes: (a) functions to flag potentially problematic coordinate records based on geographical gazetteers, (b) a global database of 9,691 geo-referenced biodiversity institutions to identify records that are likely from horticulture or captivity, (c) novel algorithms to identify datasets with rasterized data, conversion errors and strong decimal rounding and (d) spatio-temporal tests for fossils.

*Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., 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. https://doi.org/10.1111/2041-210X.13152

"We found that in GBIF more than 3.4 million records (3.7%) are potentially problematic and that 179 of the tested contributing datasets (18.5%) might be biased. In the Paleobiology Database (PDBD), 1205 records (6.3%) are potentially problematic"

5.- Time: Standardization and validation of temporal information

To standardize and validate temporal data, bdc contain a function (bdc_year_from_eventDate) to extract the collection year whenever possible from complete and legitimate date information (Table S2 in the paper).

Records with dubious collection year (e.g. 10/10/12) as well as with illegitimate (e.g. 1450, 2050) or no collection date supplied (e.g. 0 and NA) are flagged and can be subsequently removed (bdc_year_outOfRange function).

Who oversees data cleaning and what skills are needed?


APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro | Santiago José Elías Velazco | Karlo Guidoni-Martins | Geiziane Tessarolo | Lucas Jardim | Steven P. Bachman | Rafael Loyola

Global Ecology and Conservation

Original Research Article

BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

Jing Jin, Jun Yang

DEST powered by CETAF

Digitisation of Natural History collections: From data quality to data cleaning and data publishing (5 days long course)
Natural History Museum of Crete-University of Crete 13-17 November, 2023

TARGETED GROUPS OF THIS COURSE

The course is addressed to everyone who is engaged in biological and geological collections and their data, such as **Curators** and **Collections' managers**, Directors/Senior managers, **Collections' Digitization managers/officers**, **Scientists on bio- or geo informatics**, **Students** (Graduates, Post graduates, MSc, PhD), **Technicians of collections**.

1.- The profile of "the data cleaner person" is complex!

2.- Most Swiss collection lack a "the data cleaner person" as permanent staff and getting a new position will be difficult

3.- Is there a feasible solution to that?

Data quality: ensurin ... (i.e taxonomic data, specimens and materi

Data cleaning: improv ... rch and identify error instances, correct the ... e errors.
OPEN REFINE/QGI

Data visualization: fo

Biodiversity Data Sta ... lefinition, structuring, tagging, transmission, ... nalysed and reused by the wider scientific co

ABCDEFG Standards w

Publishing of data us ... rm for the geological, palaeontological & mi

Types of Data Sets convenient for publication in GeoCASe platform.

Sources of information:

THIS PPT WILL BE COMPLETELY AVAILABLE!

Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species- Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

https://assets.ctfassets.net/uo17ejk9rkwj/46SfGRfOesU0IagMMAOIkk/1c03ea3e21fcd9025cc800d786890e72/Principles_20and_20Methods_20of_20Data_20Cleaning_20-_20ENGLISH.pdf

Jin, J. & J. Yang 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity Databases. Global Ecology and Conservation 21: 1-12

https://www.sciencedirect.com/science/article/pii/S235198941930633X

Ribeiro, B. R., Velazco, S. J., 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: 1421–1428. https://doi.org/10.1111/2041-210X.13868

Ribeiro B, Velazco S, Guidoni-Martins K, Tessarolo G, Jardim L (2023). bdc: Biodiversity Data Cleaning. R package version 1.1.5 https://brunobrr.github.io/bdc/index.html

Exchange of experiences