

bdclean User Guide

Authors: Tomer Gueta and Thiloshon Nagarajah

2018-09-05

Contents

Introduction	5
1 Installing <code>bdclean</code>	9
1.1 Development version from GitHub	9
1.2 Very soon: a stable version from CRAN	9
1.3 Possible problems & solutions	9
2 Add data	11
2.1 Load package	11
2.2 Darwinizing a dataset	11
2.3 Updating the Darwin Cloud dictionary	11
3 Data cleaning configuration	13
3.1 Launching the app	13
3.2 App overview	13
3.3 Data upload	13
3.4 Dictionaries	16
3.5 Darwinizing your dataset	17
3.6 Darwinizer results	17
3.7 Download your Darwinized data	18
3.8 Closing the app	18
3.9 References	18
4 Flag and clean the data	19
4.1 Load package	19
4.2 Darwinizing a dataset	19
4.3 Updating the Darwin Cloud dictionary	19
5 Artifacts and reports	21
6 Getting your feedback	23
6.1 Report a bug	23
6.2 Contribute	23
7 <code>bdclean</code> citation	25
8 Learn more about data cleaning	27

Introduction

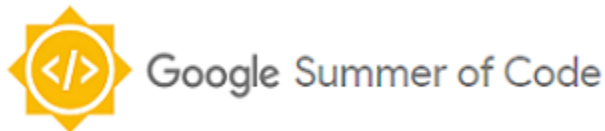
bdclean is a user-friendly data cleaning Shiny app for the inexperienced R user. It provides features to manage complete workflow for biodiversity data cleaning, from uploading the data; gathering input from the user, in order to adjust cleaning procedures; perform the cleaning; and finally, generating various reports and several versions of the data. **bdclean** is part of The bdverse – a collection of tools, that form a general framework for facilitating biodiversity science in R.

bdclean overview

bdclean workflow is comprised of three distinct mechanisms, user input, data cleaning and outputs. In most R packages this basic workflow (i.e. input; processing; output) operates via an R function. Functions are fundamental building blocks of R, and usually focus on very specific task. Users must understand and supply the function with its mandatory arguments (e.g. data in the specified format, setting of various function variables). Thus, in order to create a specific workflow, users must write an R script, which requires reasonable programming skills. bdclean avoids all that by creating a user-friendly Shiny app with questionnaire that collects the necessary user input.

App overview

Fundings



See the GSoC project idea page

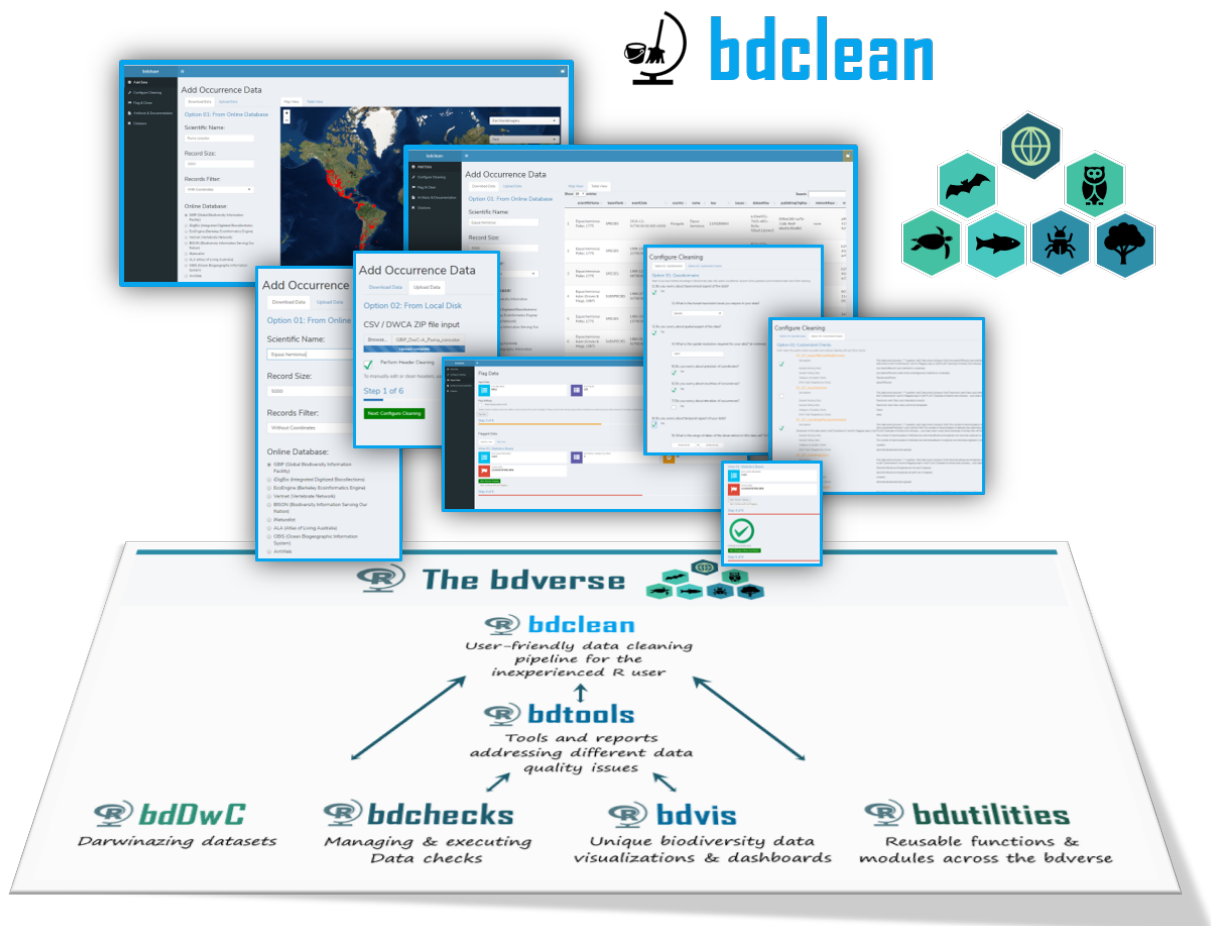


Figure 1: bdclean in the bdverse

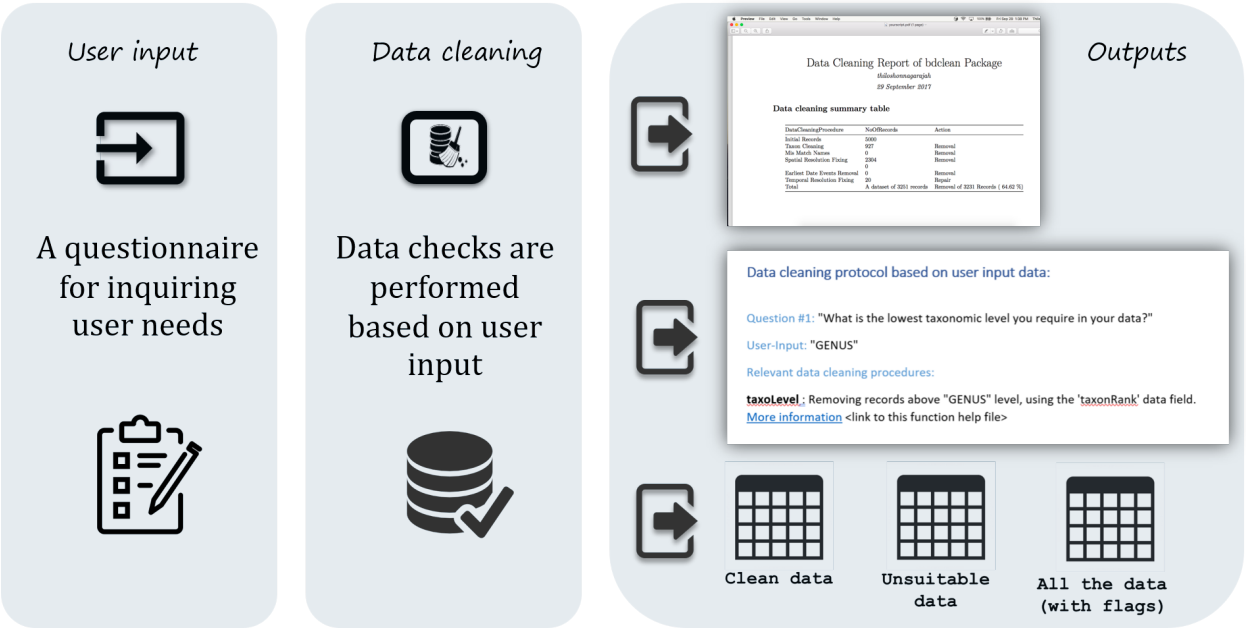


Figure 2: bdclean overview



Figure 3:

Chapter 1

Installing bdclean

1.1 Development version from GitHub

Windows users install Rtools first.

```
install.packages("devtools")
devtools::install_github("bd-R/bdclean")
# And also
devtools::install_github("bd-R/bdchecks")
```

To open the Shiny app, simply run:

```
run_bdclean()
```

1.2 Very soon: a stable version from CRAN

```
install.packages("bdDwC")
```

1.3 Possible problems & solutions

[TBA]

1.3.1 ???

TBA

1.3.2 ????

TBA

Chapter 2

Add data

2.1 Load package

Load the bdDwC package

```
library(bdDwC)
```

2.2 Darwinizing a dataset

bdDwC contains Indian Reptile dataset `bdDwC:::dataReptiles`.

The function to Darwinize a dataset is `darwinizeNames` (replace `bdDwC:::dataReptiles` with wanted dataset):

```
result <- darwinizeNames(dataUser = bdDwC:::dataReptiles,  
                        dataDWC   = bdDwC:::dataDarwinCloud$data)
```

You can replace `bdDwC:::dataReptiles` with your dataset

Rename your dataset field names to Darwinized names using `renameUserData`:

```
renameUserData(bdDwC:::dataReptiles, result)
```

2.3 Updating the Darwin Cloud dictionary

To get newest version of Darwin Cloud Data run:

```
downloadCloudData()
```

which will download data from the remote repository and extract field and standard names.

Chapter 3

Data cleaning configuration

3.1 Launching the app

```
library(bdDwC) # Upload package library  
runDwC() # Launch the app
```

3.2 App overview

In the first screen, you'll need to upload or download your biodiversity data; choose dictionary and run the Darwinizer.

3.3 Data upload

3.3.1 From a local file

A CSV file or a Darwin Core Archive (DwC-A) zip file can be uploaded.

3.3.2 From an online database

Also, data can be retrieved directly from various online biodiversity databases. You need only to:

- Select the database
- Specify the desired scientific name.
- Specify the number of records (upper limit of 50,000).
- Check the box if records must have coordinates.
- Wait for data to be downloaded.

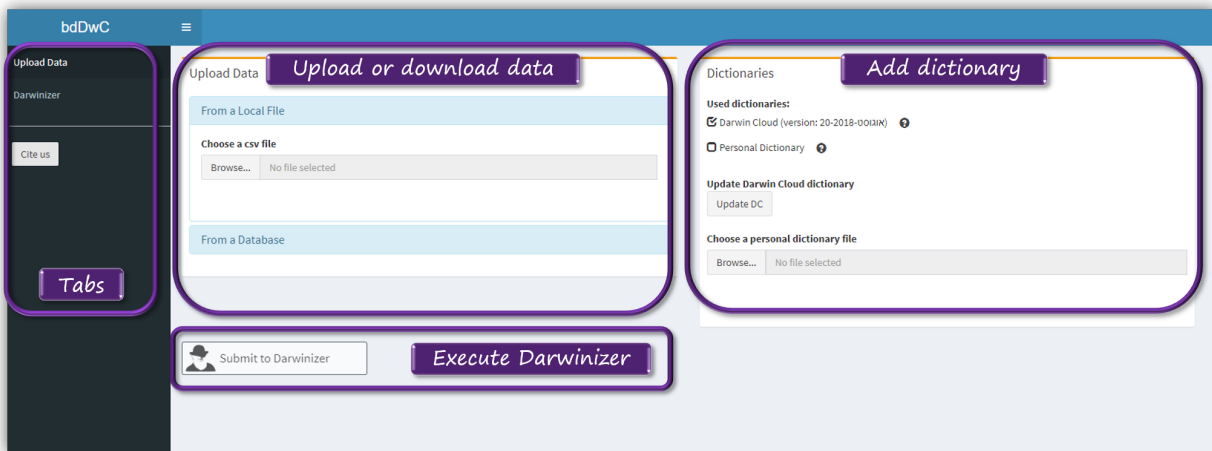


Figure 3.1: bdDwC App Overview

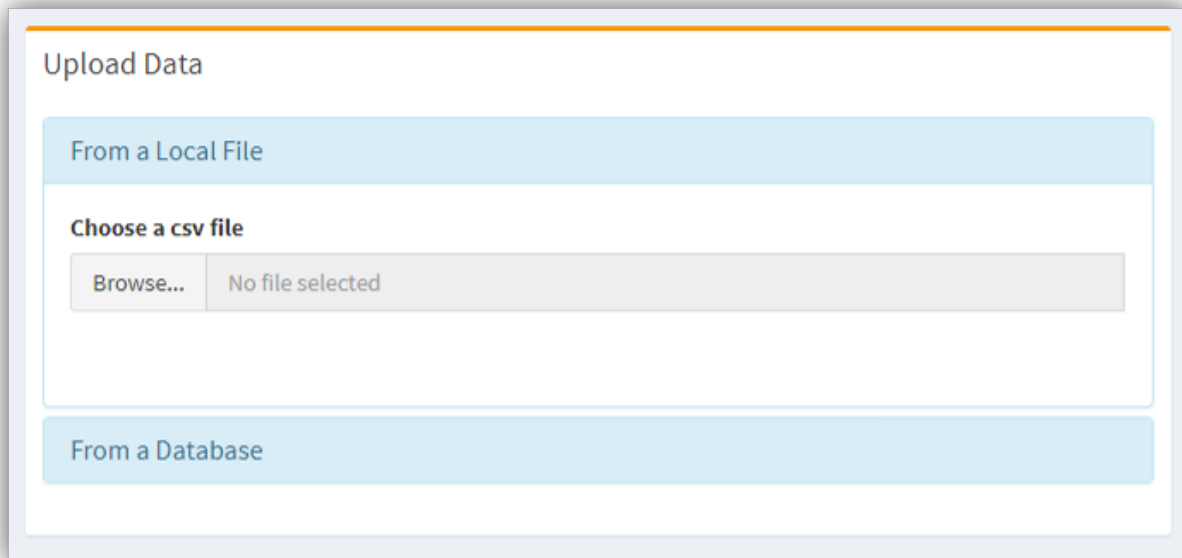


Figure 3.2: Data upload from a local file

Upload Data

From a Local File

From a Database

Scientific Name:

Record Size:

500

50,000

05,00010,00015,00020,00025,00030,00035,00040,00045,00050,000

Online Database:

☒ GBIF

☐ Bison

☐ Inat

☐ eBird

☐ Ecoengine

☐ Vertnet


 Query Database

Figure 3.3: Data upload from online biodiversity databases

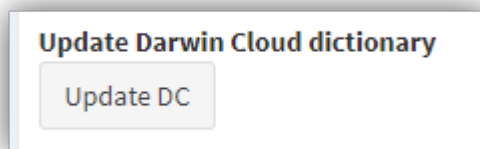


Figure 3.4: Updating the Darwin Cloud

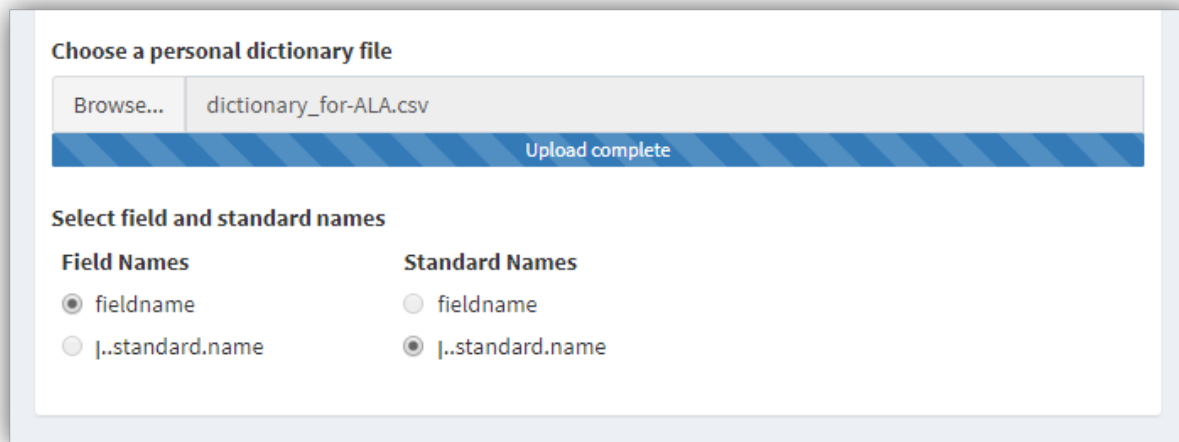


Figure 3.5: Uploading your own dictionary

3.4 Dictionaries

A dictionary is a key component when Darwinizing a dataset. It's basically a lookup table that lists a possible variation of field name and its corresponding DwC name.

3.4.1 The Darwin Cloud dictionary

The Darwin Cloud dictionary (Wieczorek et al., 2017), is a lookup table that accumulates different variations in DwC field names from different publishers. This valuable and critical dictionary was created and is maintained by the Kurator project (<http://kurator.acis.ufl.edu/kurator-web/>), which provides workflow tools for data quality improvement of biodiversity data, via a user-friendly web interface. The development of bdDwC was inspired by Kurator's own Darwinizer.

Updating the Darwin Cloud

It's recommended to update the Darwin Cloud file. This can be done easily by clicking the **Update DC** button.

3.4.2 Your own dictionary

It's also possible to add your own dictionary by simply creating a CSV file with two columns, one for the Field Names and one for the Standard Names.

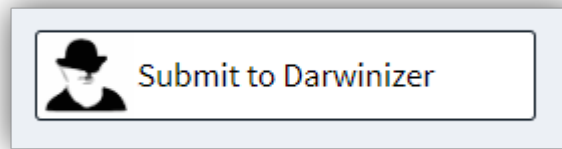


Figure 3.6: Submit to Darwinizer button

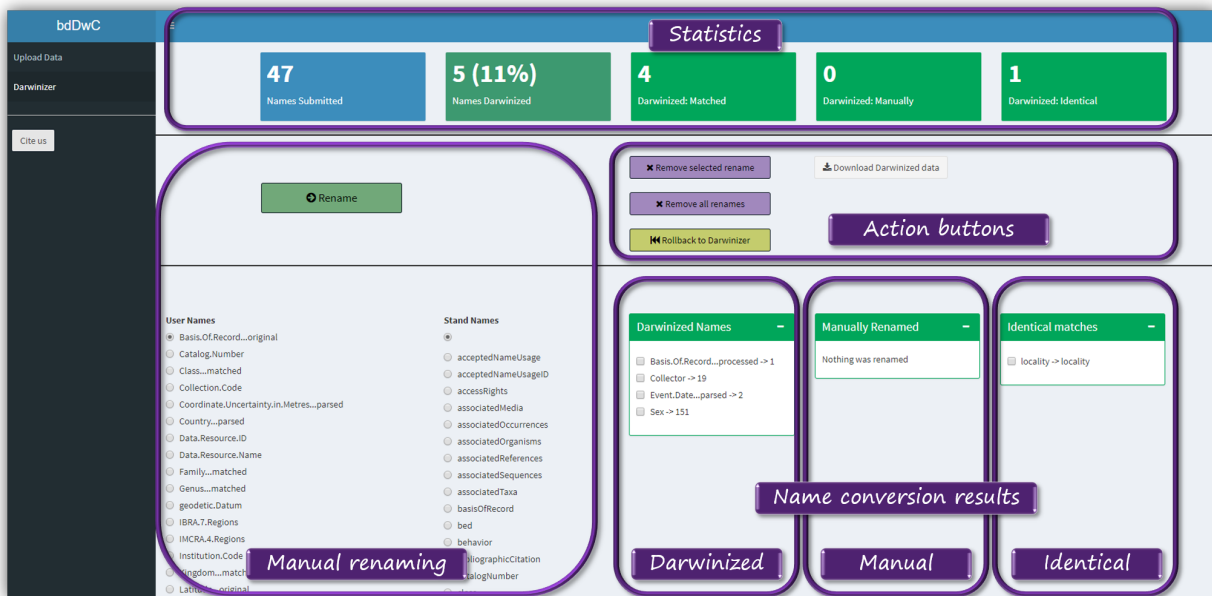


Figure 3.7: Darwinizer results

3.5 Darwinizing your dataset

Once a dataset is uploaded, the ‘Submit to Darwinizer’ button is activated, Clicking it will Darwinize the dataset.

3.6 Darwinizer results

3.6.1 Results page overview

Manually renaming field names can be done very easily, just choose the two corresponding fields and click the Rename button.

Hovering over a DwC standard name will display its description.

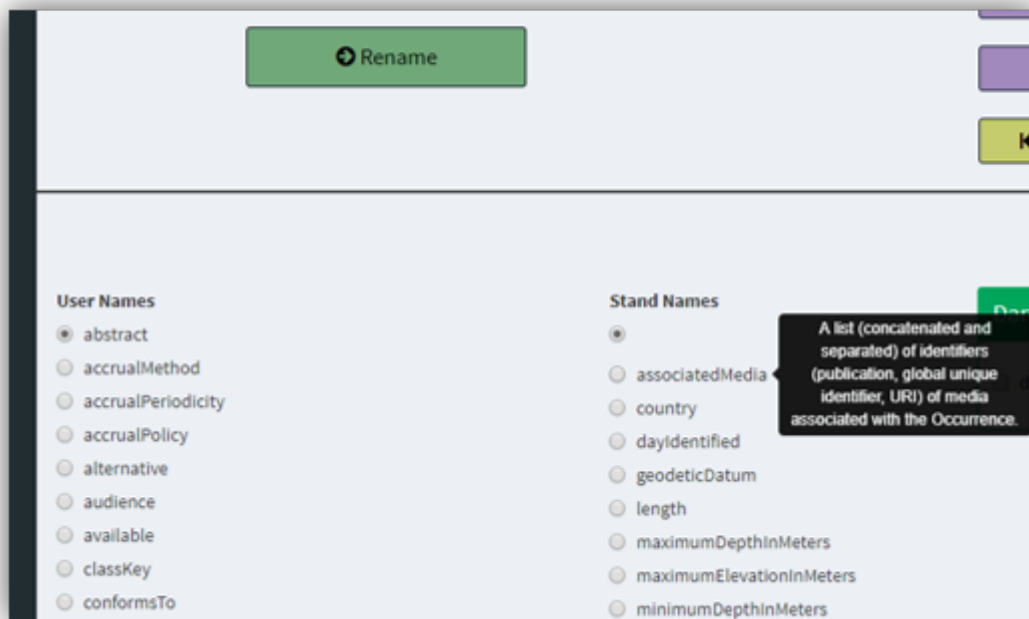


Figure 3.8: Manually renaming fields

3.7 Download your Darwinized data

3.8 Closing the app

Just close the app browser tab, and the R session will be terminated. To reopen it run in the R Console `runDwC()`.

3.9 References

Chapter 4

Flag and clean the data

4.1 Load package

Load the bdDwC package

```
library(bdDwC)
```

4.2 Darwinizing a dataset

bdDwC contains Indian Reptile dataset `bdDwC:::dataReptiles`.

The function to Darwinize a dataset `isdarwinizeNames` (replace `bdDwC:::dataReptiles` with wanted dataset):

```
result <- darwinizeNames(dataUser = bdDwC:::dataReptiles,  
                        dataDWC   = bdDwC:::dataDarwinCloud$data)
```

You can replace `bdDwC:::dataReptiles` with your dataset

Rename your dataset field names to Darwinized names using `renameUserData`:

```
renameUserData(bdDwC:::dataReptiles, result)
```

4.3 Updating the Darwin Cloud dictionary

To get newest version of Darwin Cloud Data run:

```
downloadCloudData()
```

which will download data from the remote repository and extract field and standard names.

Chapter 5

Artifacts and reports

[TBA]

Chapter 6

Getting your feedback

Loading...

6.1 Report a bug

Submit an issue at <https://github.com/bd-R/bdclean/issues>

6.2 Contribute

Contribute: <https://github.com/bd-R/bdclean>

Join: <https://bd-r-group.slack.com>

Chapter 7

bdclean citation

```
citation("bdclean")
```

```
##
## To cite package 'bdclean' in publications use:
##
##   Tomer Gueta, Thiloshon Nagarajah, Vijay Barve, Ashwin Agrawal,
##   Povilas Gibas and Yohay Carmel (2018). bdclean: A user-friendly
##   data cleaning app for the inexperienced R user. R package
##   version 0.1.900. https://github.com/bd-R/bdclean
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {bdclean: A user-friendly data cleaning app for the inexperienced R user},
##     author = {Tomer Gueta and Thiloshon Nagarajah and Vijay Barve and Ashwin Agrawal and Povilas Gibas},
##     year = {2018},
##     note = {R package version 0.1.900},
##     url = {https://github.com/bd-R/bdclean},
##   }
```


Chapter 8

Learn more about data cleaning

-
- [The Darwin Core Questions & Answers Site](#)
 - [Darwin Core Hour webinar series](#)
 - [The Darwin Core Questions & Answers wiki](#)
 - [GBIF: What is Darwin Core, and why does it matter?](#)
 - [Darwin Core: An Evolving Community-Developed Biodiversity Data Standard \(Wieczorek et al., 2012\)](#)

References

Bibliography

- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglaiss, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS one*, 7(1):e29715.
- Wieczorek, J., Morris, P. J., Hanken, J., B. Lowery, D., Ludäscher, B., Macklin, J., McPhillips, T., A. Morris, R., and Zhang, Q. (2017). Darwin cloud: Mapping real-world data to darwin core. *Biodiversity Information Science and Standards*, 1:e20486.