

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 Directly download data to the app	11
2.2 Upload data from a local file	11
2.3 Map view	11
2.4 Table view	11
3 Data cleaning configuration	15
3.1 Option 1: a questionnaire	17
3.2 Option 2: choose data checks	17
4 Flaging and cleaning	19
4.1 Data flags	19
4.2 Perform the cleaning	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

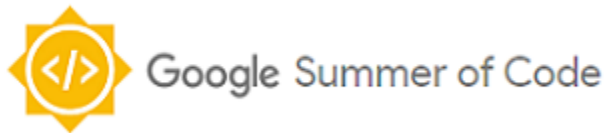
bdclean is a user-friendly data cleaning Shiny app for the inexperienced R user. It provides features to manage complete workflow for biodiversity data cleaning, from uploading the data; gathering input from the user, in order to adjust cleaning procedures; perform the cleaning; and finally, generating 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's concept

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



- [The GSoC project idea page](#)

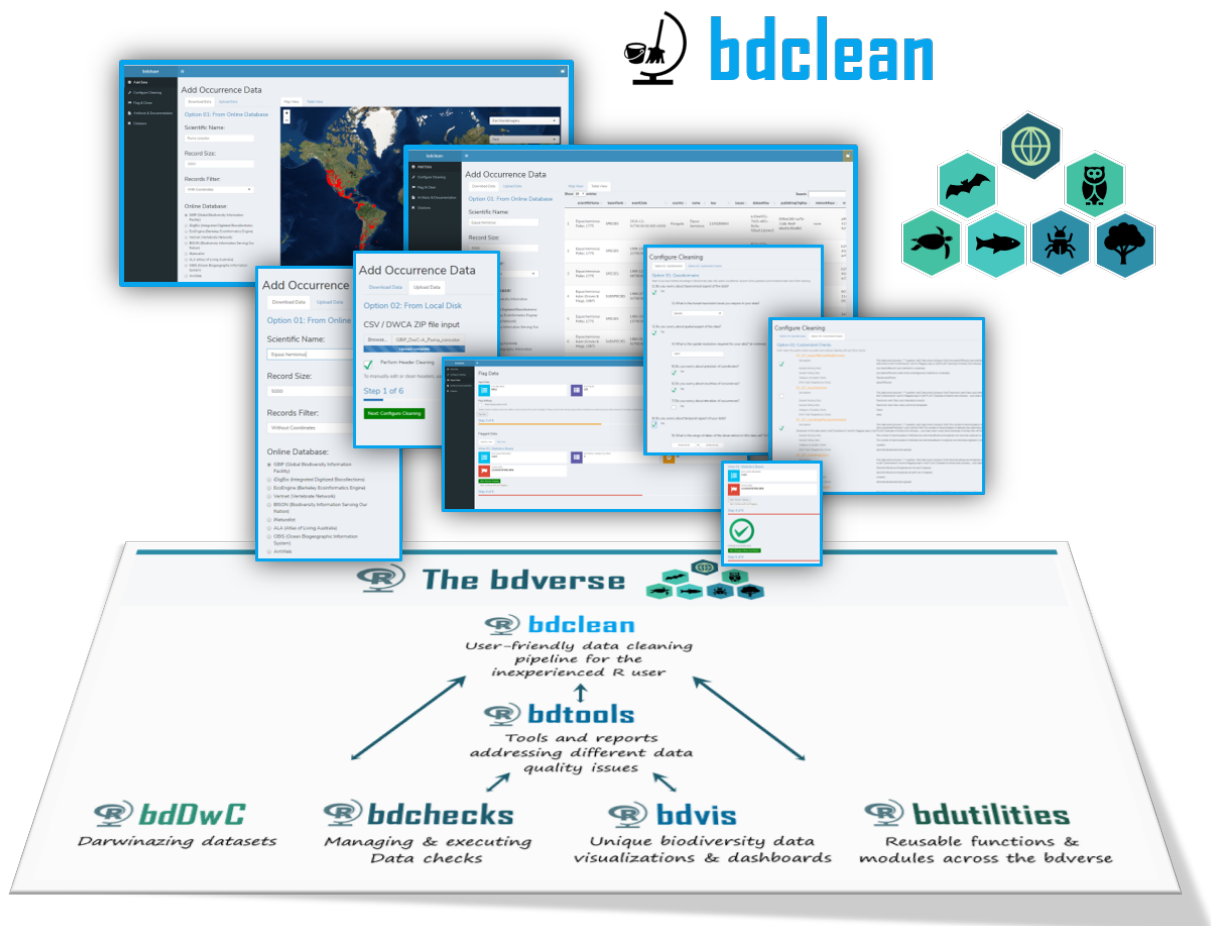


Figure 1: bdclean in the bdverse

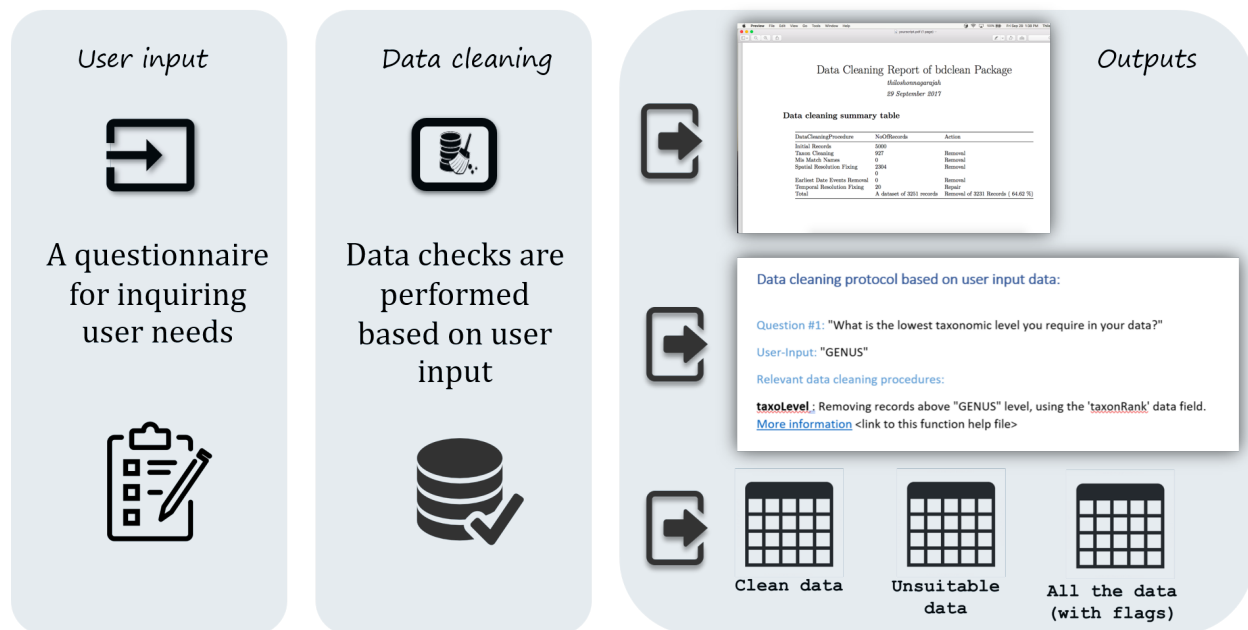


Figure 2: The main idea behind bdclean

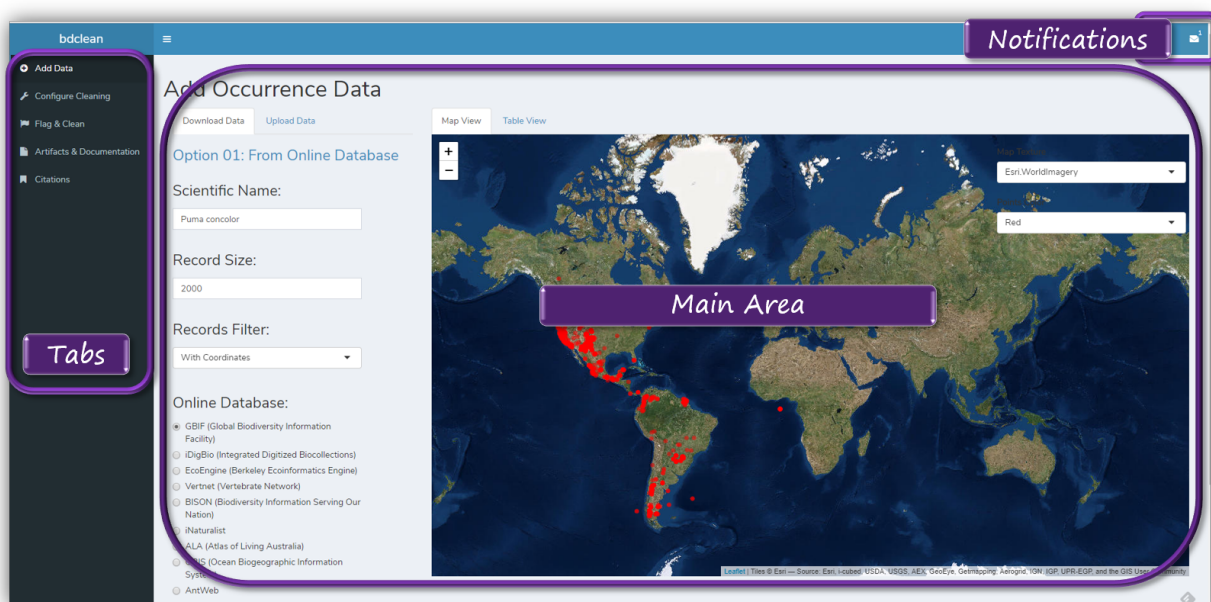


Figure 3: bdclean overview



This work is supported by the
Israel Science Foundation (ISF)
Grant No. 127/16

Figure 4:

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 Directly download data to the app

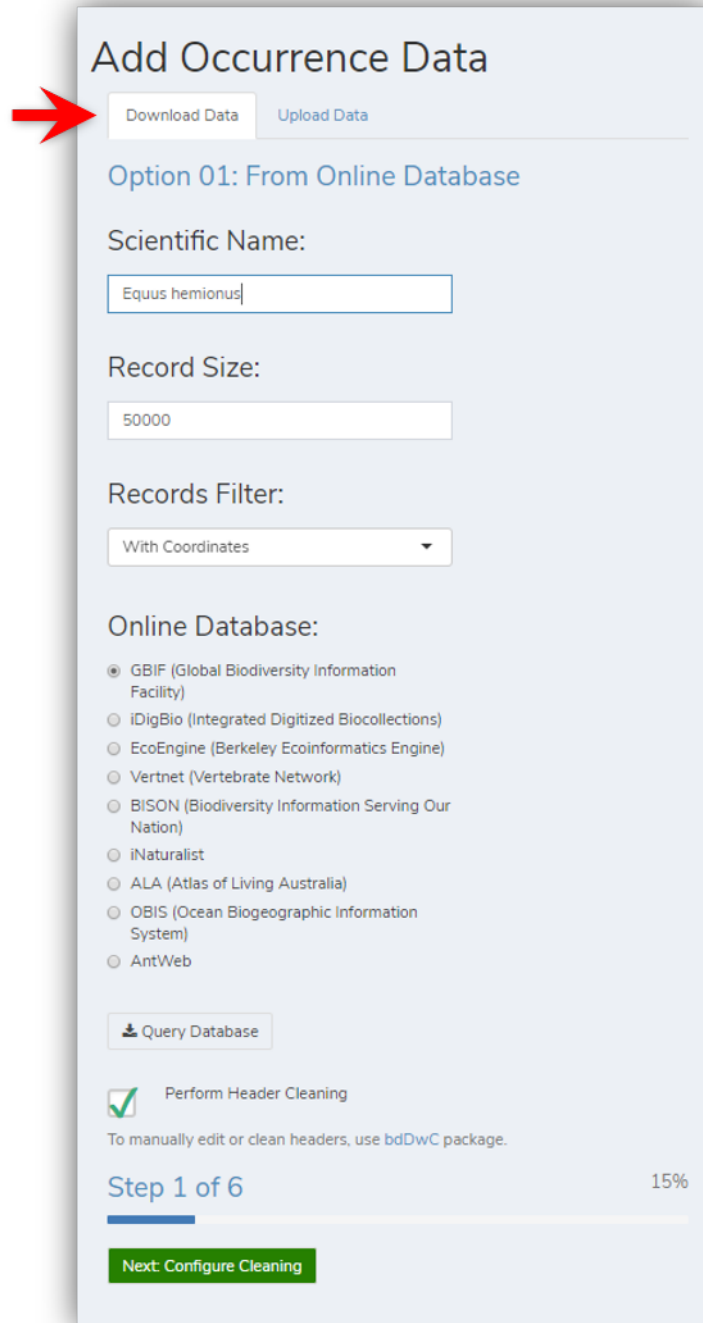
OR

2.2 Upload data from a local file

The app supports CSV files and DwC-A zip files (Darwin Core Archive)

2.3 Map view

2.4 Table view



Add Occurrence Data

Download Data Upload Data

Option 01: From Online Database


Scientific Name:

Record Size:

Records Filter:

Online Database:

- ☒ GBIF (Global Biodiversity Information Facility)
- ☐ iDigBio (Integrated Digitized Biocollections)
- ☐ EcoEngine (Berkeley Ecoinformatics Engine)
- ☐ Vertnet (Vertebrate Network)
- ☐ BISON (Biodiversity Information Serving Our Nation)
- ☐ iNaturalist
- ☐ ALA (Atlas of Living Australia)
- ☐ OBIS (Ocean Biogeographic Information System)
- ☐ AntWeb

 Query Database

☒ Perform Header Cleaning
To manually edit or clean headers, use `bdDwC` package.

Step 1 of 6 15%

Next: Configure Cleaning

Figure 2.1: Downloading data from online biodiversity databases

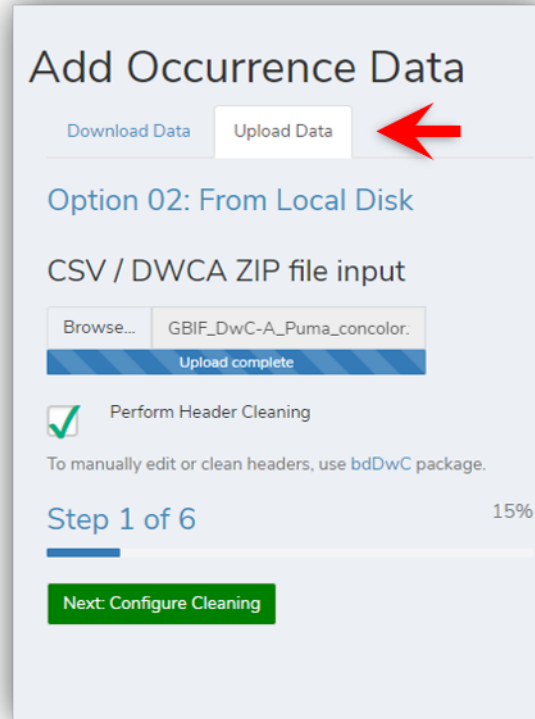


Figure 2.2: Upload file from a disk

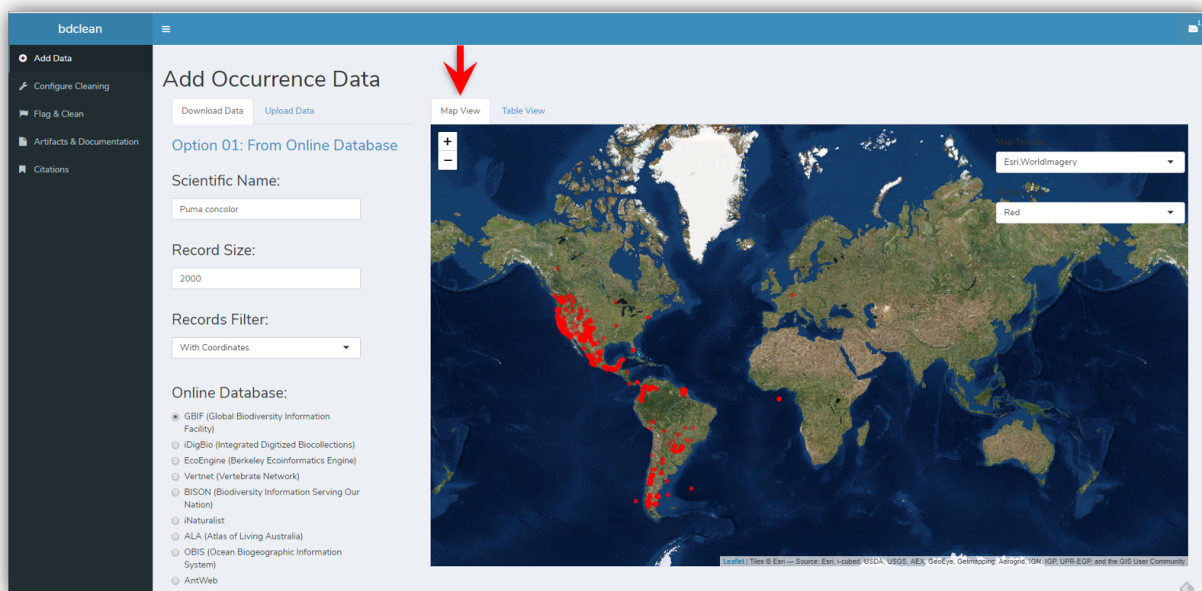


Figure 2.3: View records on a map

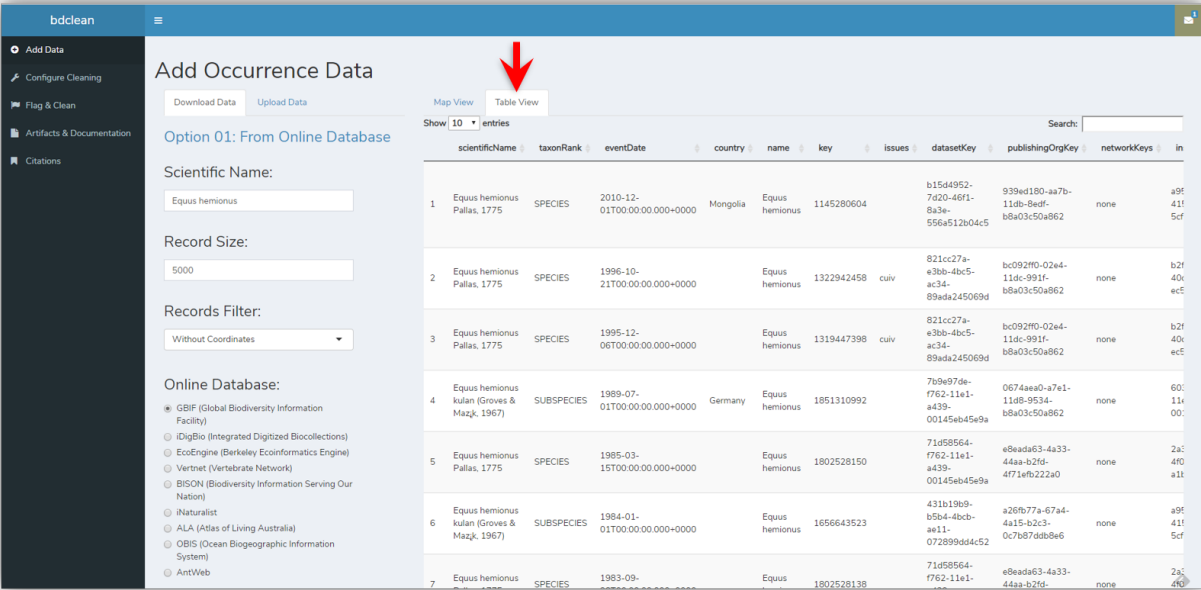
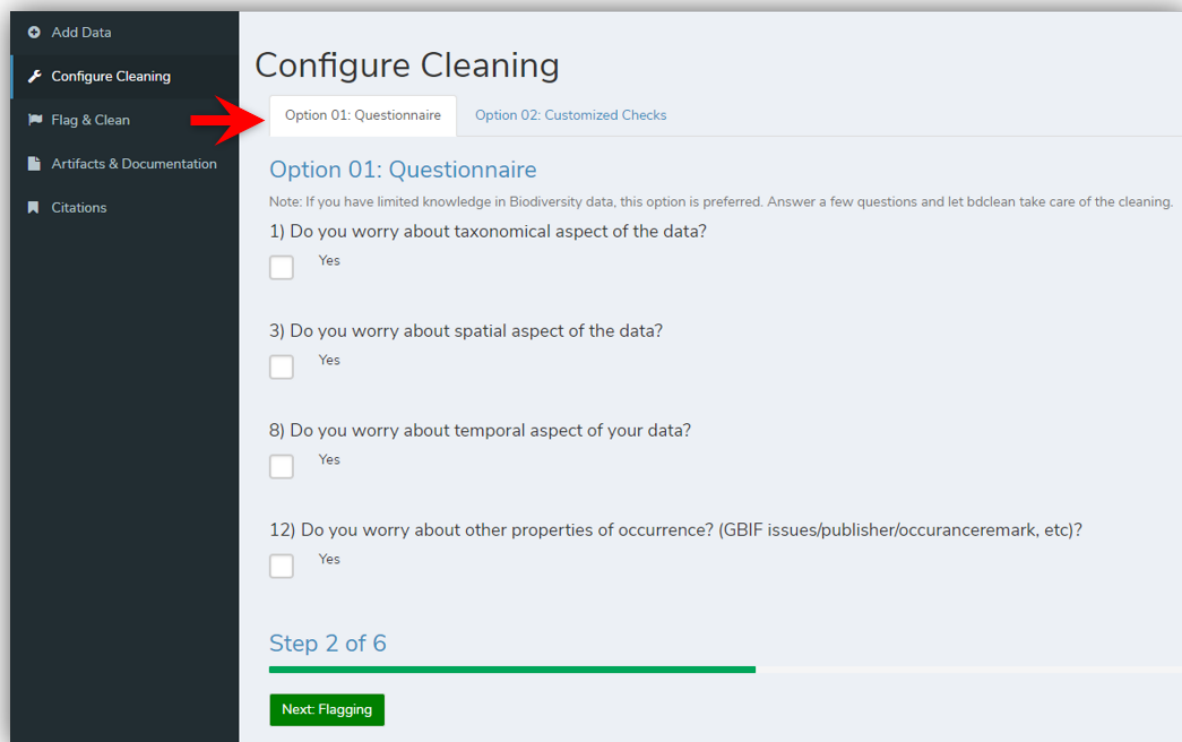


Figure 2.4: View records in a table

Chapter 3

Data cleaning configuration



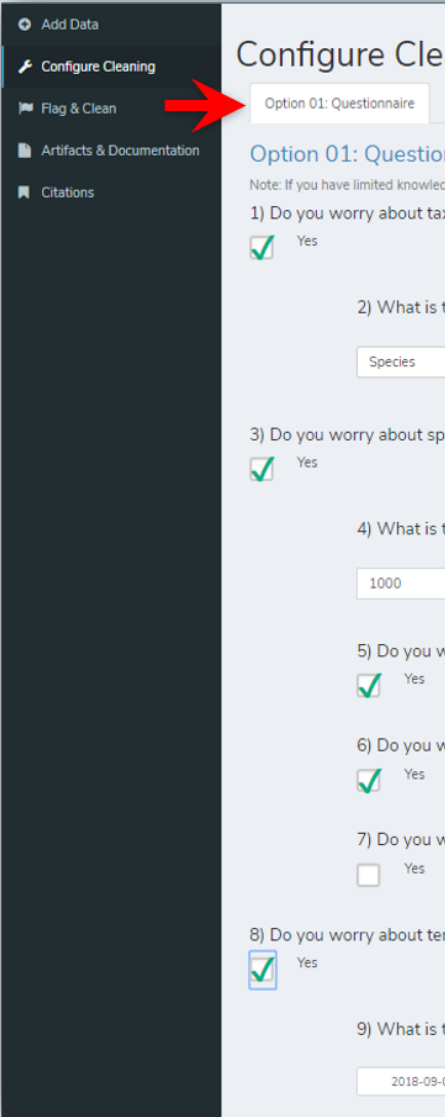
The screenshot displays a web application interface for configuring data cleaning. On the left is a dark sidebar with navigation links: 'Add Data', 'Configure Cleaning' (highlighted with a red arrow), 'Flag & Clean', 'Artifacts & Documentation', and 'Citations'. The main content area is titled 'Configure Cleaning' and contains two tabs: 'Option 01: Questionnaire' (active) and 'Option 02: Customized Checks'. Under the active tab, there is a note: 'Note: If you have limited knowledge in Biodiversity data, this option is preferred. Answer a few questions and let bdclean take care of the cleaning.' Below the note are four questions, each with a 'Yes' checkbox:

- 1) Do you worry about taxonomical aspect of the data?
☐ Yes
- 3) Do you worry about spatial aspect of the data?
☐ Yes
- 8) Do you worry about temporal aspect of your data?
☐ Yes
- 12) Do you worry about other properties of occurrence? (GBIF issues/publisher/occurrenceremark, etc)?
☐ Yes

At the bottom of the main area, it says 'Step 2 of 6' with a green progress bar. A green button labeled 'Next: Flagging' is located at the bottom left of the main content area.

Figure 3.1: Data cleaning questionnaire

3.1 Option 1: a questionnaire



The screenshot displays a software interface for configuring data cleaning. On the left is a dark sidebar with a menu containing: 'Add Data', 'Configure Cleaning' (highlighted with a red arrow), 'Flag & Clean', 'Artifacts & Documentation', and 'Citations'. The main panel is titled 'Configure Cleaning' and shows 'Option 01: Questionnaire' selected. Below the title is a note: 'Note: If you have limited knowledge'. The questionnaire consists of nine items:

- 1) Do you worry about tax: ☒ Yes
- 2) What is t:
- 3) Do you worry about sp: ☒ Yes
- 4) What is t:
- 5) Do you w: ☒ Yes
- 6) Do you w: ☒ Yes
- 7) Do you w: ☐ Yes
- 8) Do you worry about ter: ☒ Yes
- 9) What is t:

The questionnaire is reactive and more questions will be shown based on your input.

3.2 Option 2: choose data checks

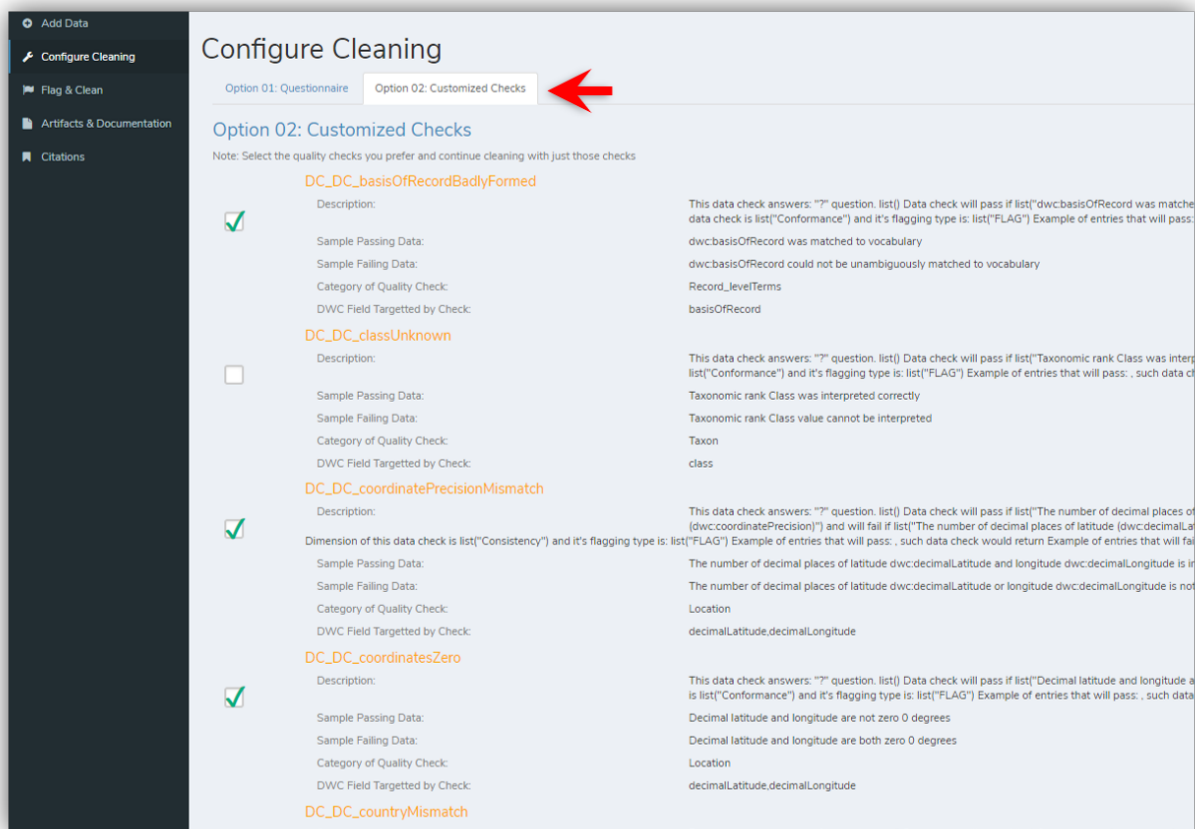


Figure 3.2: Choose your data checks

Chapter 4

Flaging and cleaning

4.1 Data flags

4.2 Perform the cleaning

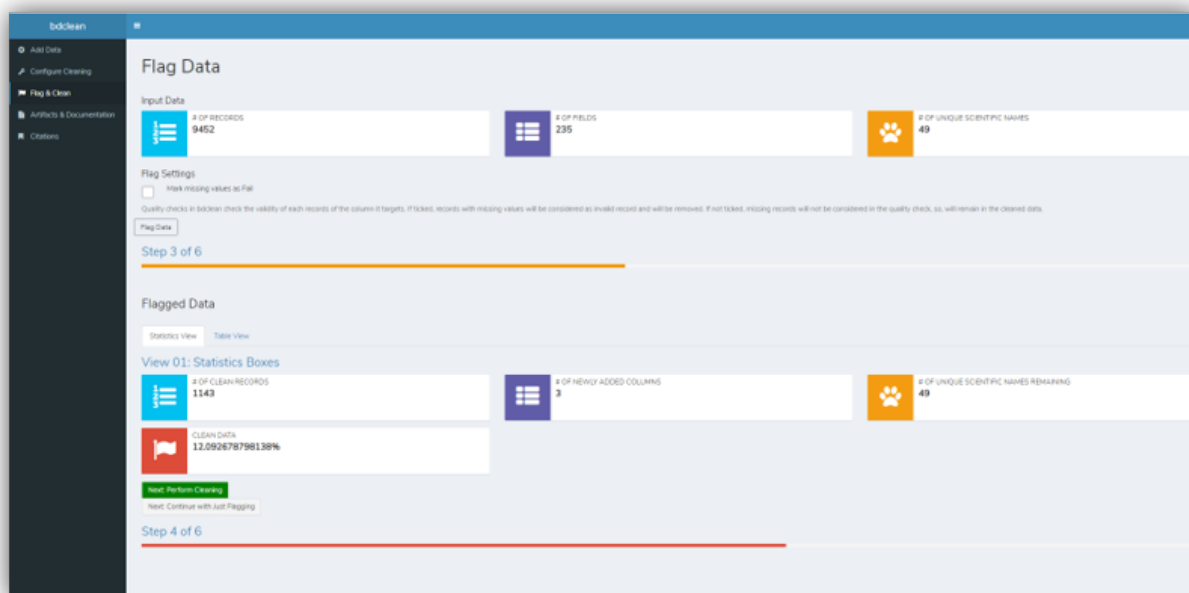


Figure 4.1: View flags

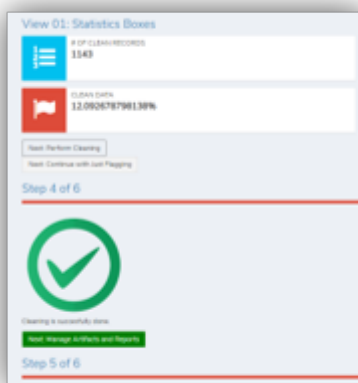


Figure 4.2: Perform the cleaning

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

-
- Biodiversity Informatics Training Curriculum (BITC)
 - BITC: Webinar Series
 - Darwin Core Hour webinar series
 - The Darwin Core Questions & Answers wiki
 - Principles of Data Quality (Chapman, 2005)
 - A conceptual framework for quality assessment and management of biodiversity data (Veiga et al., 2017)
 - Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models (Gueta and Carmel, 2016)

References

Bibliography

- Chapman, A. D. (2005). Principles of data quality, version 1.0. Technical report, GBIF.
- Gueta, T. and Carmel, Y. (2016). Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics*, 34:139–145.
- Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., and Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE*, 12(6):e0178731.