`bdchecks` User Guide

Authors: Tomer Gueta and Povilas Gibas

built on 2018-10-16

Contents

Introduction

`bdchecks` supplies a Shiny app and a set of functions to perform and manage various data checks for biodiversity data. `bdchecks` is part of the `bdverse`– a collection of tools, that form a general framework for facilitating biodiversity science in R.

What are biodiversity data checks?

Data checks can include format checks, completeness checks, reasonableness checks, limit checks, etc. These processes usually result in flagging, documenting, and subsequent correcting or eliminating of suspect records. The checks must be specifically tailored around the structure of the data at hand, in our case, the Darwin Core standard. Ideally, a data check needs to hold its functionality and relevant metadata.

What `bdchecks` can do for you?

`bdchecks` offers various features for various R users:

- Using the Shiny app **inexperienced R users** can easily perform all data check and can easily filter the data accordingly. See The shiny app section.
- **Experienced R users** can perform all data checks by utilizing few R functions form the command line or within an R script. See Command line operations section.
- **Advanced R users** can even edit, add and manage their own collection of data checks, quite easily so. See Data checks YAML file section.

Fundings

Google Summer of Code

- See the GSoC project idea page

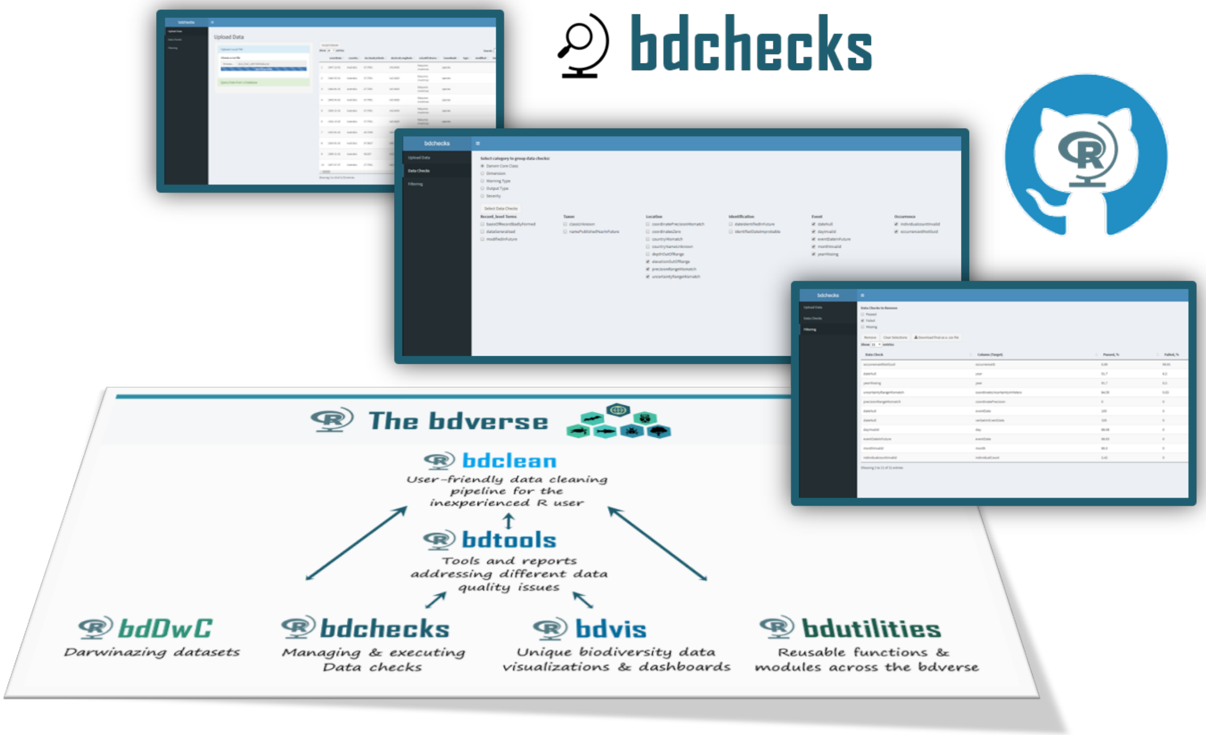

bdchecks
The bdverse
bdclean
User-friendly data cleaning pipeline for the inexperienced R user
bdtools
Tools and reports addressing different data quality issues
bdDwC
Darwinazing datasets
bdchecks
Managing & executing Data checks
bdvis
Unique biodiversity data visualizations & dashboards
bdutilities
Reusable functions & modules across the bdverse

Figure 1: bdchecks in the bdverse

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

Figure 2:

Chapter 1

Installing bdchecks

1.1 Development version from GitHub

Windows users install Rtools first.

```r
install.packages("devtools")
devtools::install_github("bd-R/bdchecks")
```

1.2 Very soon: a stable version from CRAN

```r
install.packages("bdchecks")
```

1.3 Possible installation problems & solutions

[**TBA**]

1.3.1 ???

TBA

1.3.2 ????

TBA***

Chapter 2

The shiny app

2.1 Launching the app

```
library(bdchecks) # Uplaod package library
runbdchecks() # Launch the app
```

2.2 Data upload

2.2.1 From a local file

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

2.2.2 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.

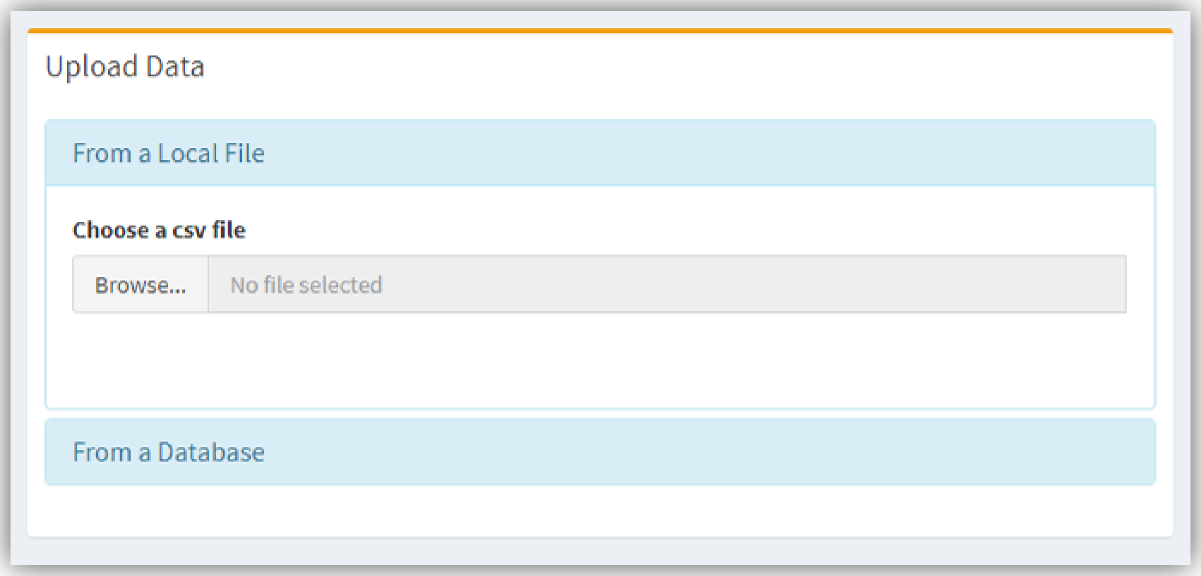


Figure 2.1: Data upload from a local file

2.2.3 Accept dataset

2.3 Choose data checks

2.4 Checks results and data filtering

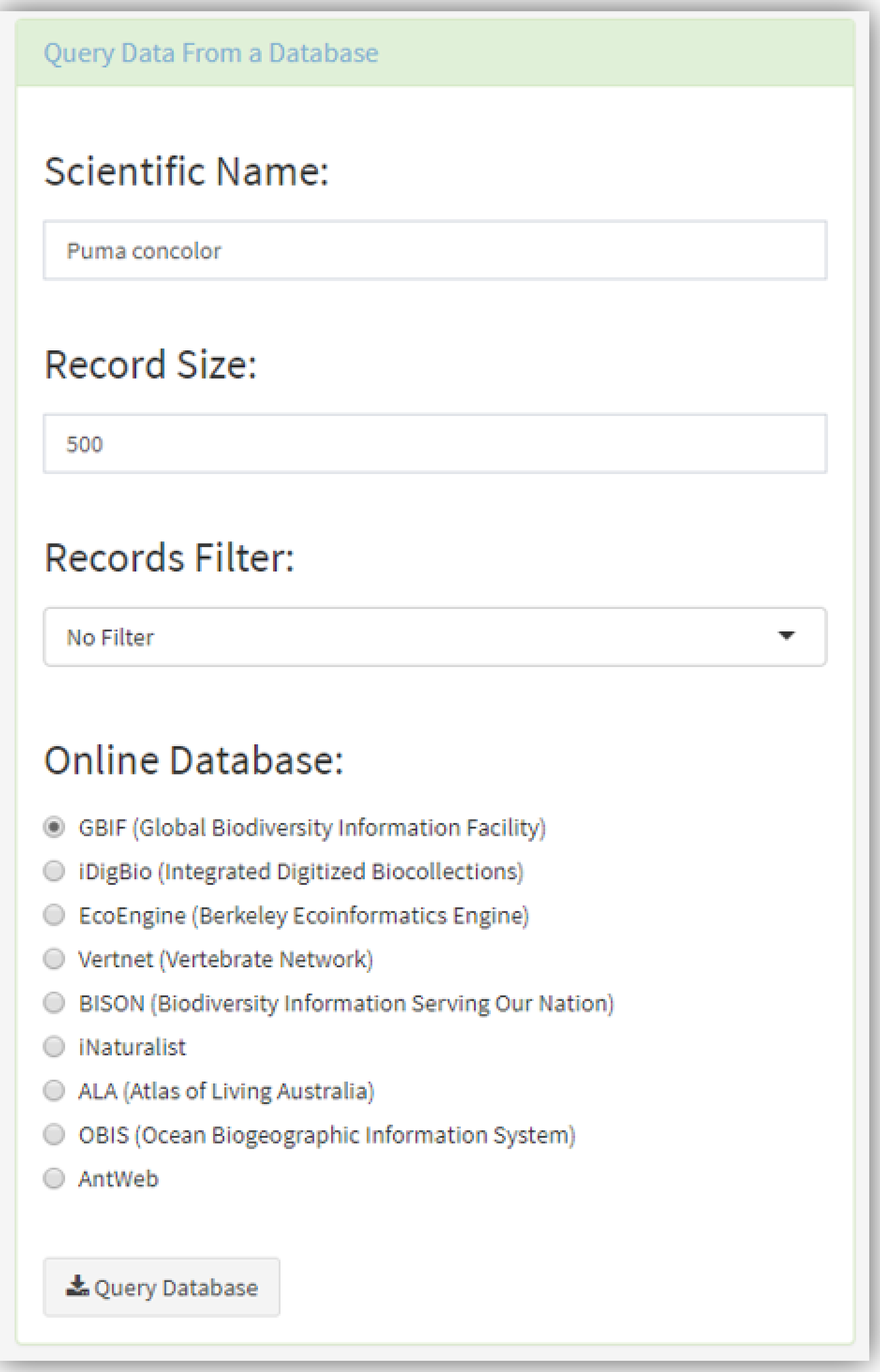
2.4.1 Overwiew

2.4.2 Filtering the data based on the results

2.5 Closing the app

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

2.6 References



Query Data From a Database

Scientific Name:

Puma concolor

Record Size:

500

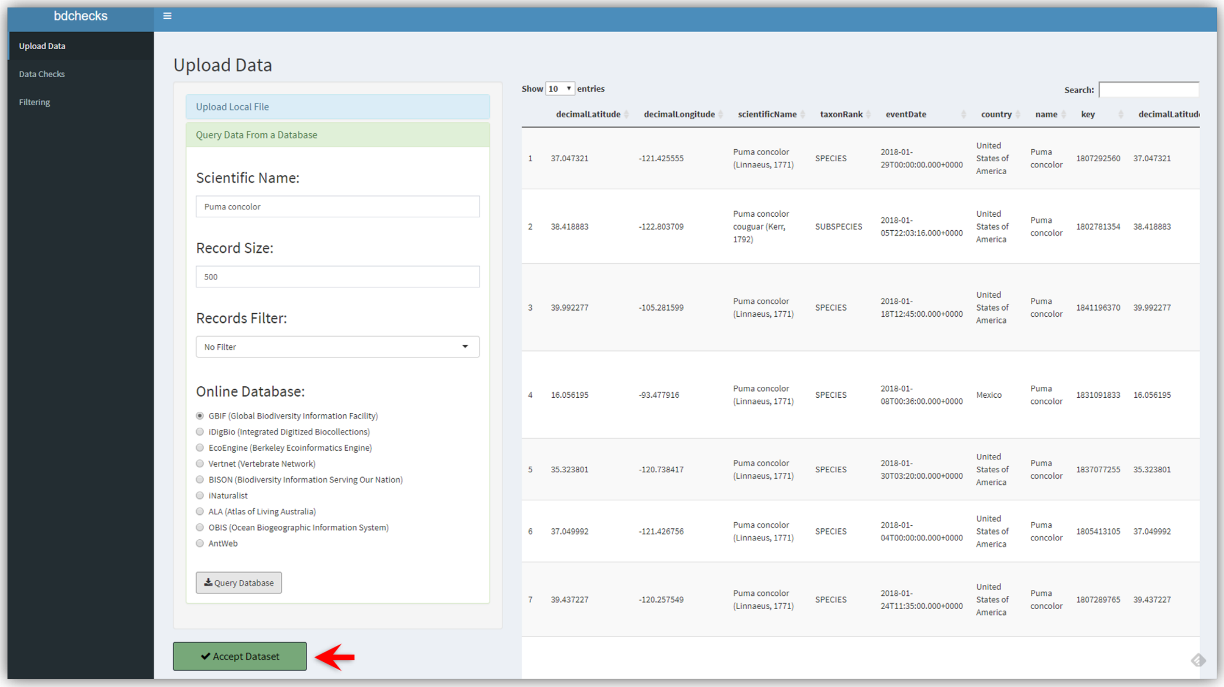
Records Filter:

No 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

Figure 2.2: Data upload from online biodiversity databases



Scientific Name:

Puma concolor

Record Size:

500

Records Filter:

No Filter

Online Database:

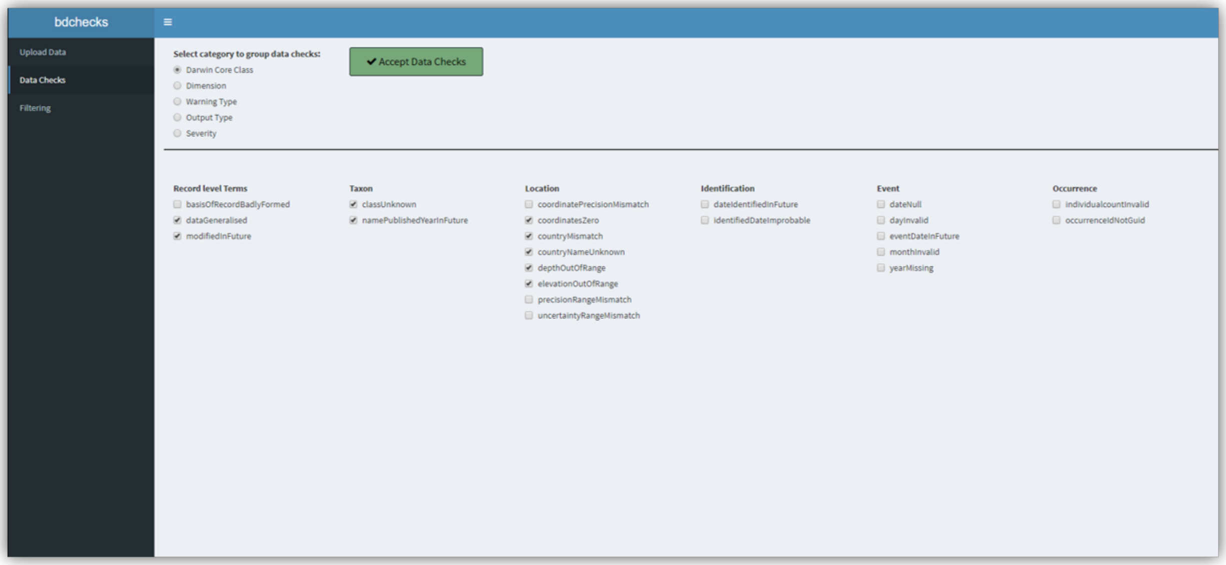
- ● GBIF (Global Biodiversity Information Facility)
- ○ iDigBio (Integrated Digitized Biocollections)
- ○ EcoEngine (Berkeley Ecoinformatics Engine)
- ○ Vertnet (Vertebrate Network)
- ○ BISON (Biodiversity Information Serving Our Nation)
- ○ iNaturalist
- ○ ALA (Atlas of Living Australia)
- ○ OBIS (Ocean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

Show 10 entries

Search:

	decimalLatitude	decimalLongitude	scientificName	taxonRank	eventDate	country	name	key	decimalLatitud
1	37.047321	-121.425555	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-29T00:00:00.000+0000	United States of America	Puma concolor	1807292560	37.047321
2	38.418883	-122.803709	Puma concolor couguar (Kerr, 1792)	SUBSPECIES	2018-01-05T22:03:16.000+0000	United States of America	Puma concolor	1802781354	38.418883
3	39.992277	-105.281599	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-18T12:45:00.000+0000	United States of America	Puma concolor	1841196370	39.992277
4	16.056195	-93.477916	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-08T00:36:00.000+0000	Mexico	Puma concolor	1831091833	16.056195
5	35.323801	-120.738417	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-30T03:20:00.000+0000	United States of America	Puma concolor	1837077255	35.323801
6	37.049992	-121.426756	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-04T00:00:00.000+0000	United States of America	Puma concolor	1805413105	37.049992
7	39.437227	-120.257549	Puma concolor (Linnaeus, 1771)	SPECIES	2018-01-24T11:35:00.000+0000	United States of America	Puma concolor	1807289765	39.437227

Figure 2.3: 'Accept dataset' to move to the next step



Select category to group data checks:

- ● Darwin Core Class
- ○ Dimension
- ○ Warning Type
- ○ Output Type
- ○ Severity

Record level Terms

- ☐ basisOfRecordBadlyFormed
- ☑ dataGeneralised
- ☑ modifiedInFuture

Taxon

- ☑ classUnknown
- ☑ namePublishedYearInFuture

Location

- ☐ coordinatePrecisionMismatch
- ☑ coordinatesZero
- ☑ countryMismatch
- ☑ countryNameUnknown
- ☑ depthOutOfRange
- ☑ elevationOutOfRange
- ☐ precisionRangeMismatch
- ☐ uncertaintyRangeMismatch

Identification

- ☐ dateIdentifiedInFuture
- ☐ identifiedDateImprobable

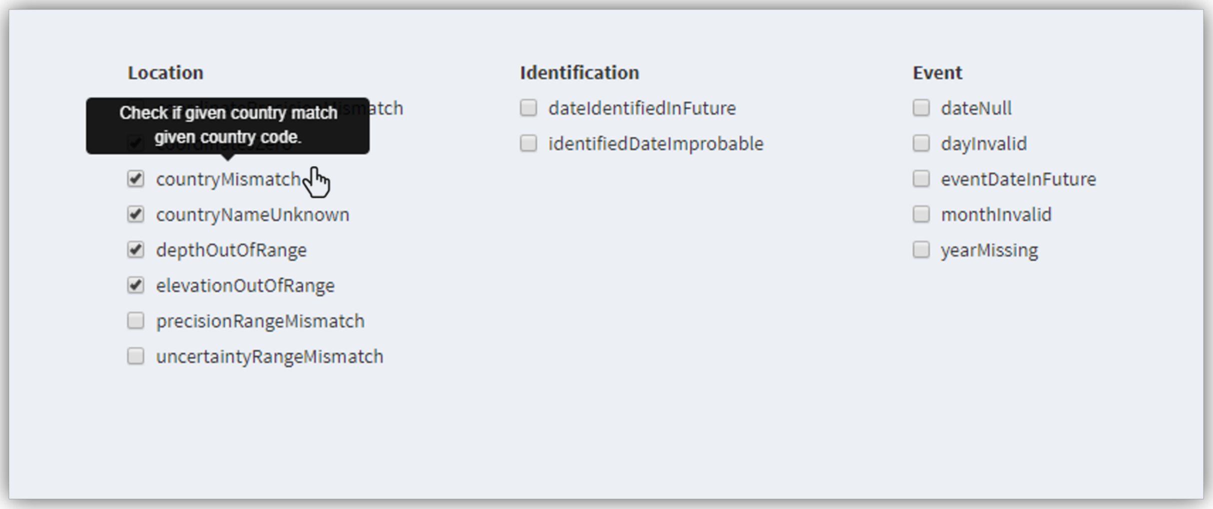
Event

- ☐ dateNull
- ☐ dayInvalid
- ☐ eventDateInFuture
- ☐ monthInvalid
- ☐ yearMissing

Occurrence

- ☐ individualcountInvalid
- ☐ occurrenceIdNotGuid

Figure 2.4: Choose a data check by checking its box



Location

Check if given country match given country code.

☑ countryMismatch

☑ countryNameUnknown

☑ depthOutOfRange

☑ elevationOutOfRange

☐ precisionRangeMismatch

☐ uncertaintyRangeMismatch

Identification

☐ dateIdentifiedInFuture

☐ identifiedDateImprobable

Event

☐ dateNull

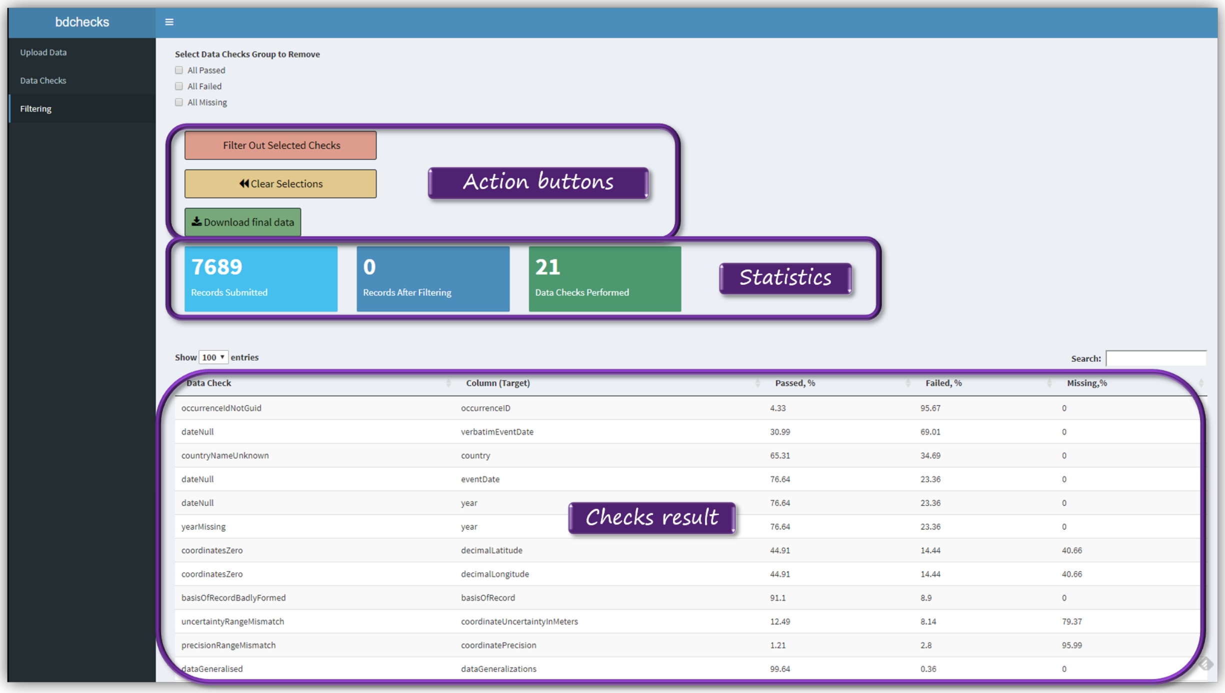
☐ dayInvalid

☐ eventDateInFuture

☐ monthInvalid

☐ yearMissing

Figure 2.5: Hovering over a data check name shows a short description



Select Data Checks Group to Remove

☐ All Passed

☐ All Failed

☐ All Missing

7689 Records Submitted

0 Records After Filtering

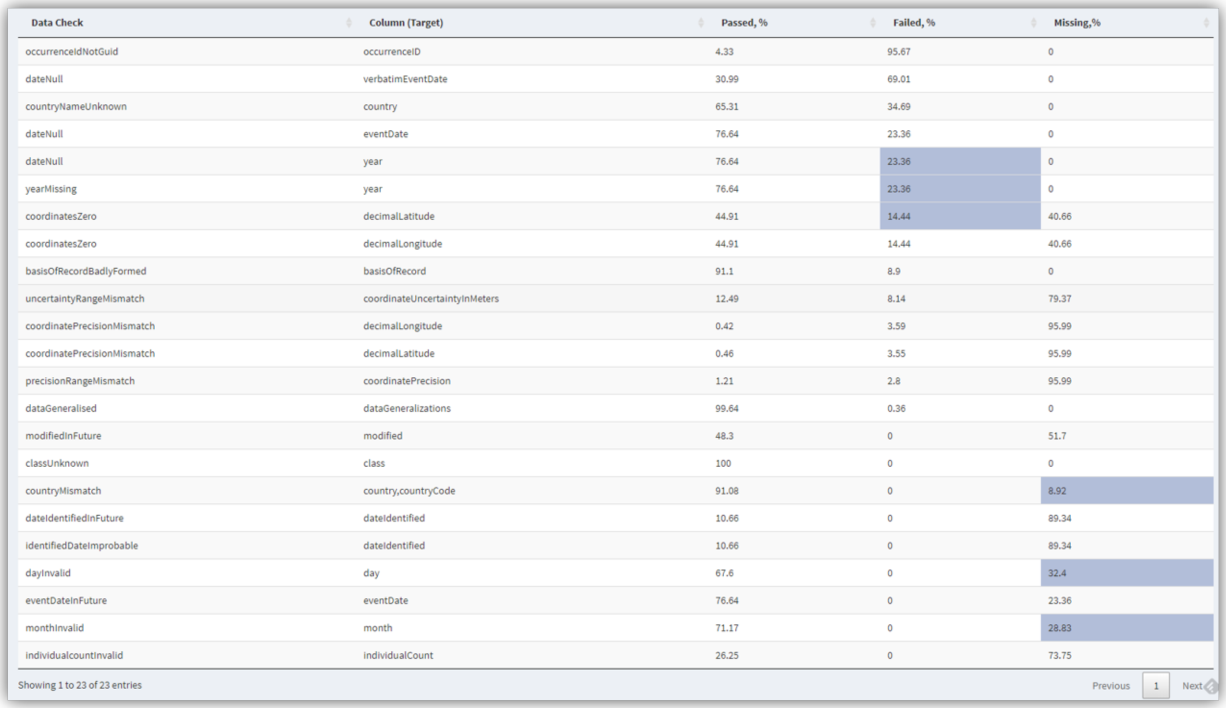
21 Data Checks Performed

Show 100 ▾ entries

Search:

Data Check	Column (Target)	Passed, %	Failed, %	Missing,%
occurrenceIdNotGuid	occurrenceID	4.33	95.67	0
dateNull	verbatimEventDate	30.99	69.01	0
countryNameUnknown	country	65.31	34.69	0
dateNull	eventDate	76.64	23.36	0
dateNull	year	76.64	23.36	0
yearMissing	year	76.64	23.36	0
coordinatesZero	decimalLatitude	44.91	14.44	40.66
coordinatesZero	decimalLongitude	44.91	14.44	40.66
basisOfRecordBadlyFormed	basisOfRecord	91.1	8.9	0
uncertaintyRangeMismatch	coordinateUncertaintyInMeters	12.49	8.14	79.37
precisionRangeMismatch	coordinatePrecision	1.21	2.8	95.99
dataGeneralised	dataGeneralizations	99.64	0.36	0

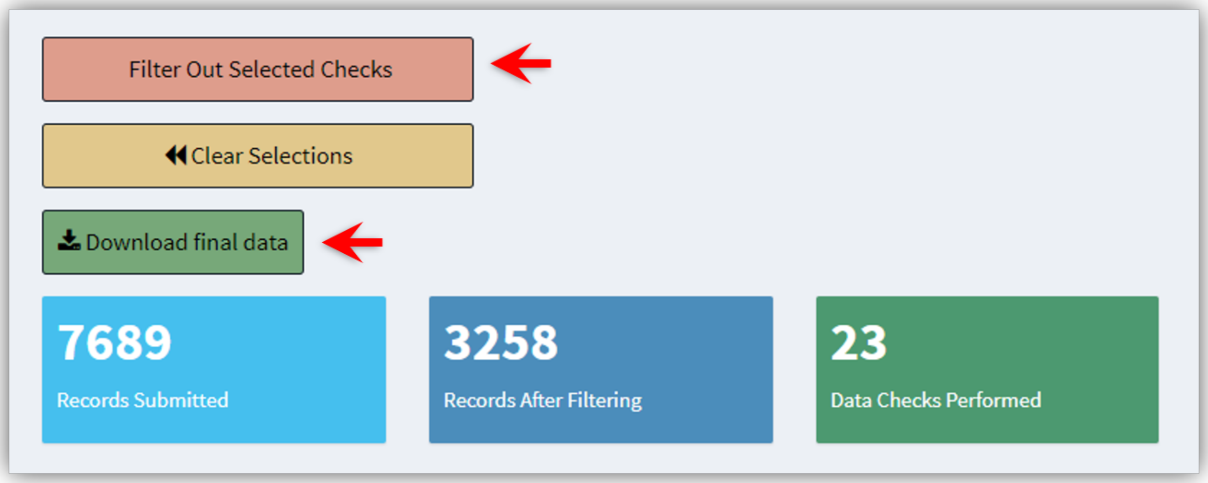
Figure 2.6: Results page overview



Data Check	Column (Target)	Passed, %	Failed, %	Missing,%
occurrenceIdNotGuid	occurrenceID	4.33	95.67	0
dateNull	verbatimEventDate	30.99	69.01	0
countryNameUnknown	country	65.31	34.69	0
dateNull	eventDate	76.64	23.36	0
dateNull	year	76.64	23.36	0
yearMissing	year	76.64	23.36	0
coordinatesZero	decimalLatitude	44.91	14.44	40.66
coordinatesZero	decimalLongitude	44.91	14.44	40.66
basisOfRecordBadlyFormed	basisOfRecord	91.1	8.9	0
uncertaintyRangeMismatch	coordinateUncertaintyInMeters	12.49	8.14	79.37
coordinatePrecisionMismatch	decimalLongitude	0.42	3.59	95.99
coordinatePrecisionMismatch	decimalLatitude	0.46	3.55	95.99
precisionRangeMismatch	coordinatePrecision	1.21	2.8	95.99
dataGeneralised	dataGeneralizations	99.64	0.36	0
modifiedInFuture	modified	48.3	0	51.7
classUnknown	class	100	0	0
countryMismatch	country,countryCode	91.08	0	8.92
dateIdentifiedInFuture	dateIdentified	10.66	0	89.34
identifiedDateImprobable	dateIdentified	10.66	0	89.34
dayInvalid	day	67.6	0	32.4
eventDateInFuture	eventDate	76.64	0	23.36
monthInvalid	month	71.17	0	28.83
individualcountInvalid	individualCount	26.25	0	73.75

Showing 1 to 23 of 23 entries

Figure 2.7: Choose specific results to filter out



7689 Records Submitted

3258 Records After Filtering

23 Data Checks Performed

Figure 2.8: Filter the data and download your filtered data

Chapter 3

Command line operations

3.1 Load package

Load the bdchecks package

```
    library(bdchecks)
```

3.2 Perform data checks

bdchecks contains a dataset on bats named dataBats.

To perform all data checks use performDataCheck:

```
resultDC <- bdchecks::performDataCheck(bdchecks::dataBats)
```

replace bdchecks::dataBats with your own dataset name.

3.3 Review performed checks

See which data checks were performed:

```
resultDC
```

Review data checks result (% of records that passed, failed or have missing data)

```
# Nice summary
summary_DC(resultDC)
```

3.4 Filtering your data

[**TBA**]

Chapter 4

Data checks YAML file

The YMAL file holds the code and metadata of all data checks. The checks are derived from a core suite of tests and assertions being developed by TDWG's Biodiversity Data Quality **Task Group 2 (Data Quality Tests and Assertions)**. More information and links can be found in the Learn more section.

4.1 Data check example

```
DC_b23110e7-1be7-444a-a677-cdee0cf4330c:
  name: countryMismatch
  meta:
    Description:
      Main: Check if given country match given country code.
      InputQuestion: Does country and country code match?
      Example:
        Fail: Country name (dwc:country) and ISO country code (dwc:countryCode) do
          not match
        Pass: Country name (dwc:country) and ISO country code (dwc:countryCode) match
        InputFail: country=Australia, countryCode=4
        InputPass: country=Australia, countryCode=AU
        OutputFail: Failed
        OutputPass: Passed
      Resolution:
        Record: SingleRecord
        Term: MultiTerm
      DarwinCoreClass: Location
      Keywords: location,iso,country
      guid: b23110e7-1be7-444a-a677-cdee0cf4330c
    Flags:
      Severity: Warning
      Warning: Inconsistent
      Output: Validation
      Dimension: Consistency
    Pseudocode: |
      get.Country($countryCode) == $country
    Source:
      Reference:
```

```
      CreatedBy: Povilas Gibas
      MaintainedBy: Povilas Gibas
      CreationDate: 2018-06-27
      ModificationDate: 2018-06-27
      ModificationHist:
  Input:
    Target: country,countryCode
    Dependency:
      DependencyType: Internal
      DataChecks:
      Rpackages: rgbif
      Data: isocodes$name,isocodes$code
  Functionality: |
      FUNC <- function() {
          result <- sapply(seq_along(TARGET1), function(i) {
              if (is.na(TARGET1[i]) | is.na(TARGET2[i])) {
                  NA
              } else {
                  which(DEPEND1 == TARGET1[i]) == which(DEPEND2 == TARGET2[i])
              }
          })
          result <- unlist(result)
          return(result)
       }
```

4.2 Manage your own data checks

After adding/ removing/ editing the YAML file, you can load data checks into R using `getDC()` function.

```r
DC <- getDC("path to your YAML file")
```

You can also export data checks from your YAML file to .rda and roxygen2 comments.

```r
exportDC("path to your YAML file")
```

Chapter 5

bdchecks architecture

5.1 The overall architecture

[TBA]

5.2 Component 1****

[TBA]

5.3 Component 2****

[TBA]

Chapter 6

Getting your feedback

Loading...

6.1 Report a bug

Submit an issue at https://github.com/bd-R/bdchecks/issues

6.2 Contribute

Contribute: https://github.com/bd-R/bdchecks

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

Chapter 7

bdchecks citation

```
citation("bdchecks")
```

```
##
## To cite package 'bdchecks' in publications use:
##
##   Povilas Gibas, Tomer Gueta, Vijay Barve, Thiloshon Nagarajah and
##   Yohay Carmel (2018). bdchecks: Biodiversity Data Checks. R
##   package version 0.1.2. https://github.com/bd-R/bdchecks
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {bdchecks: Biodiversity Data Checks},
##     author = {Povilas Gibas and Tomer Gueta and Vijay Barve and Thiloshon Nagarajah and Yohay Carmel
##     year = {2018},
##     note = {R package version 0.1.2},
##     url = {https://github.com/bd-R/bdchecks},
##   }
```

Chapter 8

Learn more

- **TDWG's data quality tests and assertions Task Group**
- **Core suite of tests and assertions**
- **Core tests and assertions as GitHub issues**
- **A conceptual framework for quality assessment and management of biodiversity data (Veiga et al., 2017)**

References

Bibliography

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.