Package 'auk'

March 28, 2018

Title eBird Data Extraction and Processing with AWK

Version 0.2.1

Description Extract and process bird sightings records from eBird (<http://ebird.org>), an online tool for recording bird observations. Public access to the full eBird database is via the eBird Basic Dataset (EBD; see <http://ebird.org/ebird/data/download> for access), a downloadable text file. This package is an interface to AWK for extracting data from the EBD based on taxonomic, spatial, or temporal filters, to produce a manageable file size that can be imported into R.

License GPL-3

URL https://github.com/CornellLabofOrnithology/auk, http://CornellLabofOrnithology.github.io/auk/

BugReports https://github.com/CornellLabofOrnithology/auk/issues

Depends R (>= 3.1.2)

Imports assertthat, countrycode (>= 1.0.0), dplyr (>= 0.7.0), rlang, stringi, stringr, tidyr

Suggests covr, data.table, knitr, readr, rmarkdown, testthat

VignetteBuilder knitr

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Author Matthew Strimas-Mackey [aut, cre] (<https://orcid.org/0000-0001-8929-7776>), Eliot Miller [aut], Wesley Hochachka [aut], Cornell Lab of Ornithology [cph]

Maintainer Matthew Strimas-Mackey <mes335@cornell.edu>

Repository CRAN

Date/Publication 2018-03-28 10:51:29 UTC

R topics documented:

---

auk auk: *eBird Data Extraction and Processing with AWK*

Description

Tools for extracting and processing eBird data from the eBird Basic Dataset (EBD).

auk_breeding *Filter to only include observations with breeding codes*

Description

eBird users have the option of specifying breeding bird atlas codes for their observations, for example, if nesting building behaviour is observed. Use this filter to select only those observations with an associated breeding code. This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_breeding(x)
```

Arguments

x	auk_ebd object; reference to basic dataset file created by `auk_ebd()`.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_breeding()
```

auk_clean *Clean an eBird data file*

Description

Some rows in the eBird Basic Dataset (EBD) may have an incorrect number of columns, often resulting from tabs embedded in the comments field. This function drops these problematic records. **Note that this function typically takes at least 3 hours to run on the full dataset**

Usage

```
auk_clean(f_in, f_out, sep = "\t", remove_text = FALSE,
  overwrite = FALSE)
```

Arguments

f_in	character; input file.
f_out	character; output file.
sep	character; the input field separator, the basic dataset is tab separated by default. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.
remove_text	logical; whether all free text entry columns should be removed. These columns include comments, location names, and observer names. These columns cause import errors due to special characters and increase the file size, yet are rarely valuable for analytical applications, so may be removed. Setting this argument to TRUE can lead to a significant reduction in file size.
overwrite	logical; overwrite output file if it already exists

Details

This function can clean a basic dataset file or a sampling file.

Calling this function requires that the command line utility AWK is installed. Linux and Mac machines should have AWK by default, Windows users will likely need to install Cygwin.

Value

If AWK ran without errors, the output filename is returned, however, if an error was encountered the exit code is returned.

Examples

```
## Not run:
# get the path to the example data included in the package
# in practice, provide path to ebd, e.g. f <- "data/ebd_relFeb-2018.txt
f <- system.file("extdata/ebd-sample_messy.txt", package = "auk")
# output to a temp file for example
# in practice, provide path to output file
# e.g. f_out <- "output/ebd_clean.txt"
f_out <- tempfile()

# clean file to remove problem rows
auk_clean(f, f_out)
# number of lines in input
length(readLines(f))
# number of lines in output
length(readLines(f_out))

# note that the extra blank column has also been removed
ncol(read.delim(f, nrows = 5, quote = ""))
ncol(read.delim(f_out, nrows = 5, quote = ""))

## End(Not run)
```

auk_complete *Filter out incomplete checklists from the eBird data*

Description

Define a filter for the eBird Basic Dataset (EBD) to only keep complete checklists, i.e. those for which all birds seen or heard were recorded. These checklists are the most valuable for scientific uses since they provide presence and absence data.This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_complete(x)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by `auk_ebd()` or `auk_sampling()`.

Details

This function can also work with on an `auk_sampling` object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_complete()
```

auk_country *Filter the eBird data by country*

Description

Define a filter for the eBird Basic Dataset (EBD) based on a set of countries. This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_country(x, country, replace = FALSE)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
country	character; countries to filter by. Countries can either be expressed as English names or ISO 2-letter country codes. English names are matched via regular expressions using countrycode, so there is some flexibility in names.
replace	logical; multiple calls to auk_country() are additive, unless replace = FALSE, in which case the previous list of countries to filter by will be removed and replaced by that in the current call.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# country names and ISO2 codes can be mixed
# not case sensitive
country <- c("CA", "United States", "mexico")
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_country(country)

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_country(ebd, country)
```

---

auk_date *Filter the eBird data by date*

Description

Define a filter for the eBird Basic Dataset (EBD) based on a range of dates. This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering.

Usage

```
auk_date(x, date)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
date	character or date; date range to filter by, provided either as a character vector in the format "2015-12-31" or a vector of Date objects. To filter on a range of dates, regardless of year, use "*" in place of the year.

Details

To select observations from a range of dates, regardless of year, the wildcard "*" can be used in place of the year. For example, using `date = c("*-05-01", "*-06-30")` will return observations from May and June of *any year*.

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_date(date = c("2010-01-01", "2010-12-31"))

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_date(ebd, date = c("2010-01-01", "2010-12-31"))

# the * wildcard can be used in place of year to select dates from all years
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  # may-june records from all years
  auk_date(date = c("*-05-01", "*-06-30"))
```

---

auk_distance *Filter eBird data by distance travelled*

Description

Define a filter for the eBird Basic Dataset (EBD) based on the distance travelled on the checklist. This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering. Note that stationary checklists (i.e. point counts) have no distance associated with them, however, since these checklists can be assumed to have 0 distance they will be kept if 0 is in the range defined by distance.

Usage

```
auk_distance(x, distance, distance_units)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
distance	integer; 2 element vector specifying the range of distances to filter by. The default is to accept distances in kilometers, use distance_units = "miles" for miles.
distance_units	character; whether distances are provided in kilometers (the default) or miles.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# only keep checklists that are less than 10 km long
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_distance(distance = c(0, 10))

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_distance(ebd, distance = c(0, 10))
```

---

auk_duration *Filter the eBird data by duration*

Description

Define a filter for the eBird Basic Dataset (EBD) based on the duration of the checklist. This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering. Note that checklists with no effort, such as incidental observations, will be excluded if this filter is used since they have no associated duration information.

Usage

```
auk_duration(x, duration)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
duration	integer; 2 element vector specifying the range of durations in minutes to filter by.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# only keep checklists that are less than an hour long
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_duration(duration = c(0, 60))

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_duration(ebd, duration = c(0, 60))
```

---

auk_ebd *Reference to eBird data file*

Description

Create a reference to an eBird Basic Dataset (EBD) file in preparation for filtering using AWK.

Usage

```
auk_ebd(file, file_sampling, sep = "\t")
```

Arguments

file	character; input file.
file_sampling	character; optional input sampling event data (i.e. checklists) file, required if you intend to zero-fill the data to produce a presence-absence data set. This file consists of just effort information for every eBird checklist. Any species not appearing in the EBD for a given checklist is implicitly considered to have a count of 0. This file should be downloaded at the same time as the basic dataset to ensure they are in sync.
sep	character; the input field separator, the eBird data are tab separated so this should generally not be modified. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.

Details

eBird data can be downloaded as a tab-separated text file from the eBird website after submitting a request for access. As of February 2017, this file is nearly 150 GB making it challenging to work with. If you're only interested in a single species or a small region it is possible to submit a custom download request. This approach is suggested to speed up processing time.

There are two potential pathways for preparing eBird data. Users wishing to produce presence only data, should download the eBird Basic Dataset and reference this file when calling `auk_ebd()`. Users wishing to produce zero-filled, presence absence data should additionally download the sampling event data file associated with the basic dataset This file contains only checklist information and can be used to infer absences. The sampling event data file should be provided to `auk_ebd()` via the `file_sampling` argument. For further details consult the vignettes.

Value

An `auk_ebd` object storing the file reference and the desired filters once created with other package functions.

Examples

```
# get the path to the example data included in the package
# in practice, provide path to ebd, e.g. f <- "data/ebd_relFeb-2018.txt
f <- system.file("extdata/ebd-sample.txt", package = "auk")
auk_ebd(f)
# to produce zero-filled data, provide a checklist file
f_ebd <- system.file("extdata/zerofill-ex_ebd.txt", package = "auk")
f_cl <- system.file("extdata/zerofill-ex_sampling.txt", package = "auk")
auk_ebd(f_ebd, file_sampling = f_cl)
```

---

auk_extent *Filter the eBird data by spatial extent*

Description

Define a filter for the eBird Basic Dataset (EBD) based on spatial extent. This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_extent(x, extent)
```

Arguments

`x`	`auk_ebd` or `auk_sampling` object; reference to file created by `auk_ebd()` or `auk_sampling()`.
`extent`	numeric; spatial extent expressed as the range of latitudes and longitudes in decimal degrees: `c(lng_min, lat_min, lng_max, lat_max)`. Note that longitudes in the Western Hemisphere and latitudes sound of the equator should be given as negative numbers.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# fliter to locations roughly in the Pacific Northwest
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_extent(extent = c(-125, 37, -120, 52))

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_extent(ebd, extent = c(-125, 37, -120, 52))
```

---

auk_filter *Filter the eBird file using AWK*

Description

Convert the filters defined in an auk_ebd object into an AWK script and run this script to produce a filtered eBird Reference Dataset (ERD). The initial creation of the auk_ebd object should be done with auk_ebd() and filters can be defined using the various other functions in this package, e.g. auk_species() or auk_country(). **Note that this function typically takes at least a couple hours to run on the full dataset**

Usage

```
auk_filter(x, file, ...)

## S3 method for class 'auk_ebd'
auk_filter(x, file, file_sampling, keep, drop, awk_file,
  sep = "\t", filter_sampling = TRUE, execute = TRUE,
  overwrite = FALSE, ...)

## S3 method for class 'auk_sampling'
auk_filter(x, file, keep, drop, awk_file, sep = "\t",
  execute = TRUE, overwrite = FALSE, ...)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
file	character; output file.
...	arguments passed on to methods.
file_sampling	character; optional output file for sampling data.
keep	character; a character vector specifying the names of the columns to keep in the output file. Columns should be as they appear in the header of the EBD; however, names are not case sensitive and spaces may be replaced by underscores, e.g. "COMMON NAME", "common name", and "common_NAME" are all valid.
drop	character; a character vector of columns to drop in the same format as keep. Ignored if keep is supplied.
awk_file	character; output file to optionally save the awk script to.
sep	character; the input field separator, the eBird file is tab separated by default. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.
filter_sampling	logical; whether the sampling event data should also be filtered.
execute	logical; whether to execute the awk script, or output it to a file for manual execution. If this flag is FALSE, awk_file must be provided.
overwrite	logical; overwrite output file if it already exists

Details

If a sampling file is provided in the auk_ebd object, this function will filter both the eBird Basic Dataset and the sampling data using the same set of filters. This ensures that the files are in sync, i.e. that they contain data on the same set of checklists.

The AWK script can be saved for future reference by providing an output filename to awk_file. The default behavior of this function is to generate and run the AWK script, however, by setting execute = FALSE the AWK script will be generated but not run. In this case, file is ignored and awk_file must be specified.

Calling this function requires that the command line utility AWK is installed. Linux and Mac machines should have AWK by default, Windows users will likely need to install Cygwin.

Value

An auk_ebd object with the output files set. If execute = FALSE, then the path to the AWK script is returned instead.

Methods (by class)

- auk_ebd: auk_ebd object
- auk_sampling: auk_sampling object

Examples

```
# get the path to the example data included in the package
# in practice, provide path to ebd, e.g. f <- "data/ebd_relFeb-2018.txt"
f <- system.file("extdata/ebd-sample.txt", package = "auk")
# define filters
filters <- auk_ebd(f) %>%
  auk_species(species = c("Gray Jay", "Blue Jay")) %>%
  auk_country(country = c("US", "Canada")) %>%
  auk_extent(extent = c(-100, 37, -80, 52)) %>%
  auk_date(date = c("2012-01-01", "2012-12-31")) %>%
  auk_time(start_time = c("06:00", "09:00")) %>%
  auk_duration(duration = c(0, 60)) %>%
  auk_complete()

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
filters <- auk_species(ebd, species = c("Gray Jay", "Blue Jay"))
filters <- auk_country(filters, country = c("US", "Canada"))
filters <- auk_extent(filters, extent = c(-100, 37, -80, 52))
filters <- auk_date(filters, date = c("2012-01-01", "2012-12-31"))
filters <- auk_time(filters, start_time = c("06:00", "09:00"))
filters <- auk_duration(filters, duration = c(0, 60))
filters <- auk_complete(filters)

# apply filters
## Not run:
# output to a temp file for example
# in practice, provide path to output file
# e.g. f_out <- "output/ebd_filtered.txt"
f_out <- tempfile()
filtered <- auk_filter(filters, file = f_out)
str(read_ebd(filtered))

## End(Not run)
```

---

auk_getpath *OS specific path to AWK*

Description

Return the OS specific path to AWK, or highlights if it's not installed. To manually set the path to AWK, set the `AWK_PATH` environment variable in your `.Renviron` file.

Usage

```
auk_getpath()
```

Value

Path to AWK or `NA` if AWK wasn't found.

Examples

```
auk_getpath()
```

`auk_last_edited` *Filter the eBird data by last edited date*

Description

Define a filter for the eBird Basic Dataset (EBD) based on a range of last edited dates. Last edited date is typically used to extract just new or recently edited data. This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_last_edited(x, date)
```

Arguments

`x`	auk_ebd or auk_sampling object; reference to file created by `auk_ebd()` or `auk_sampling()`.
`date`	character or date; date range to filter by, provided either as a character vector in the format `"2015-12-31"` or a vector of Date objects.

Details

This function can also work with on an `auk_sampling` object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_last_edited(date = c("2010-01-01", "2010-12-31"))
```

auk_project *Filter the eBird data by project code*

Description

Some eBird records are collected as part of a particular project (e.g. the Virginia Breeding Bird Survey) and have an associated project code in the eBird dataset (e.g. EBIRD_ATL_VA). This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_project(x, project)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by `auk_ebd()` or `auk_sampling()`.
project	character; project code to filter by (e.g. `"EBIRD_MEX"`). Multiple codes are accepted.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_project("EBIRD_MEX")

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_project(ebd, "EBIRD_MEX")
```

auk_protocol *Filter the eBird data by protocol*

Description

Filter to just data collected following a specific search protocol: stationary, traveling, or casual. This function only defines the filter and, once all filters have been defined, `auk_filter()` should be used to call AWK and perform the filtering.

Usage

```
auk_protocol(x, protocol)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by `auk_ebd()` or `auk_sampling()`.
protocol	character; "Stationary", "Traveling", or "Incidental". Other protocols exist in the database, however, this function only extracts these three standard protocols. Multiple protocols are accepted.

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_protocol("Stationary")

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_protocol(ebd, "Stationary")
```

auk_rollup *Roll up eBird taxonomy to species*

Description

The eBird Basic Dataset (EBD) includes both true species and every other field-identifiable taxon that could be relevant for birders to report. This includes taxa not identifiable to a species (e.g. hybrids) and taxa reported below the species level (e.g. subspecies). This function produces a list of observations of true species, by removing the former and rolling the latter up to the species level. In the resulting EBD data.frame, `category` will be `"species"` for all records and the subspecies fields will be dropped. By default, `read_ebd()` calls `ebd_rollup()` when importing an eBird data file.

Usage

```
auk_rollup(x)
```

Arguments

x	data.frame; data frame of eBird data, typically as imported by `read_ebd()`

Details

When rolling observations up to species level the observed counts are summed across any taxa that resolve to the same species. However, if any of these taxa have a count of "X" (i.e. the observer did not enter a count), then the rolled up record will get an "X" as well. For example, if an observer saw 3 Myrtle and 2 Audubon's Warblers, this will roll up to 5 Yellow-rumped Warblers. However, if an "X" was entered for Myrtle, this would roll up to "X" for Yellow-rumped Warbler.

The eBird taxonomy groups taxa into eight different categories. These categories, and the way they are treated by `auk_rollup()` are as follows:

- **Species:** e.g., Mallard. Combined with lower level taxa if present on the same checklist.
- **ISSF or Identifiable Sub-specific Group:** Identifiable subspecies or group of subspecies, e.g., Mallard (Mexican). Rolled-up to species level.
- **Intergrade:** Hybrid between two ISSF (subspecies or subspecies groups), e.g., Mallard (Mexican intergrade. Rolled-up to species level.
- **Form:** Miscellaneous other taxa, including recently-described species yet to be accepted or distinctive forms that are not universally accepted (Red-tailed Hawk (Northern), Upland Goose (Bar-breasted)). If the checklist contains multiple taxa corresponding to the same species, the lower level taxa are rolled up, otherwise these records are left as is.
- **Spuh:** Genus or identification at broad level – e.g., duck sp., dabbling duck sp.. Dropped by `auk_rollup()`.
- **Slash:** Identification to Species-pair e.g., American Black Duck/Mallard). Dropped by `auk_rollup()`.
- **Hybrid:** Hybrid between two species, e.g., American Black Duck x Mallard (hybrid). Dropped by `auk_rollup()`.
- **Domestic:** Distinctly-plumaged domesticated varieties that may be free-flying (these do not count on personal lists) e.g., Mallard (Domestic type). Dropped by `auk_rollup()`.

Value

A data frame of the eBird data with taxonomic rollup applied.

References

Consult the eBird taxonomy page for further details.

Examples

```
# get the path to the example data included in the package
# in practice, provide path to ebd, e.g. f <- "data/ebd_relFeb-2018.txt
f <- system.file("extdata/ebd-rollup-ex.txt", package = "auk")
# read in data without rolling up
ebd <- read_ebd(f, rollup = FALSE)
# rollup
ebd_ru <- auk_rollup(ebd)

# all taxa not identifiable to species are dropped
unique(ebd$category)
unique(ebd_ru$category)

# yellow-rump warbler subspecies rollup
library(dplyr)
# without rollup, there are three observations
ebd %>%
  filter(common_name == "Yellow-rumped Warbler") %>%
  select(checklist_id, category, common_name, subspecies_common_name,
         observation_count)
# with rollup, they have been combined
ebd_ru %>%
  filter(common_name == "Yellow-rumped Warbler") %>%
  select(checklist_id, category, common_name, observation_count)
```

---

auk_sampling *Reference to eBird sampling event file*

Description

Create a reference to an eBird sampling event file in preparation for filtering using AWK. For working with the sightings data use `auk_ebd()`, only use `auk_sampling()` if you intend to only work with checklist-level data.

Usage

```
auk_sampling(file, sep = "\t")
```

Arguments

file	character; input sampling event data file, which contains checklist data from eBird.
sep	character; the input field separator, the eBird data are tab separated so this should generally not be modified. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.

Details

eBird data can be downloaded as a tab-separated text file from the eBird website after submitting a request for access. In the eBird Basic Dataset (EBD) each row corresponds to a observation of a single bird species on a single checklist, while the sampling event data file contains a single row for every checklist. This function creates an R object to reference only the sampling data.

Value

An auk_sampling object storing the file reference and the desired filters once created with other package functions.

Examples

```
# get the path to the example data included in the package
# in practice, provide path to the sampling event data
# e.g. f <- "data/ebd_sampling_relFeb-2018.txt"
f <- system.file("extdata/zerofill-ex_sampling.txt", package = "auk")
auk_sampling(f)
```

---

auk_select *Select a subset of columns*

Description

Select a subset of columns from the eBird Basic Dataset (EBD) or the sampling events file. Subsetting the columns can significantly decrease file size.

Usage

```
auk_select(x, select, file, sep = "\t", overwrite = FALSE)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
select	character; a character vector specifying the names of the columns to select. Columns should be as they appear in the header of the EBD; however, names are not case sensitive and spaces may be replaced by underscores, e.g. "COMMON NAME", "common name", and "common_NAME" are all valid.

file	character; output file.
sep	character; the input field separator, the eBird file is tab separated by default. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.
overwrite	logical; overwrite output file if it already exists

Value

Invisibly returns the filename of the output file.

Examples

```
## Not run:
# select a minimal set of columns
out_file <- tempfile()
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
cols <- c("latitude", "longitude",
          "group identifier", "sampling event identifier",
          "scientific name", "observation count")
selected <- auk_select(ebd, select = cols, file = out_file)
str(read_ebd(selected))

## End(Not run)
```

---

auk_species *Filter the eBird data by species*

Description

Define a filter for the eBird Basic Dataset (EBD) based on species. This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering.

Usage

```
auk_species(x, species, replace)
```

Arguments

x	auk_ebd object; reference to object created by auk_ebd().
species	character; species to filter by, provided as scientific or English common names, or a mixture of both. These names must match the official eBird Taxomony (ebird_taxonomy).
replace	logical; multiple calls to auk_species() are additive, unless replace = FALSE, in which case the previous list of species to filter by will be removed and replaced by that in the current call.

Value

An auk_ebd object.

Examples

```
# common and scientific names can be mixed
species <- c("Gray Jay", "Pluvialis squatarola")
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_species(species)

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_species(ebd, species)
```

---

auk_split *Split an eBird data file by species*

Description

Given an eBird Basic Dataset (EBD) and a list of species, split the file into multiple text files, one for each species. This function is typically used after `auk_filter()` has been applied if the resulting file is too large to be read in all at once.

Usage

```
auk_split(file, species, prefix = "", ext = "txt", sep = "\t",
  overwrite = FALSE)
```

Arguments

`file`	character; input file.
`species`	species character; species to filter and split by, provided as scientific or English common names, or a mixture of both. These names must match the official eBird Taxomony (ebird_taxonomy).
`prefix`	character; a file and directory prefix. For example, if splitting by species "A" and "B" and `prefix = "data/ebd_"`, the resulting files will be "data/ebd_A.txt" and "data/ebd_B.txt".
`ext`	character; file extension, typically "txt".
`sep`	character; the input field separator, the eBird file is tab separated by default. Must only be a single character and space delimited is not allowed since spaces appear in many of the fields.
`overwrite`	logical; overwrite output files if they already exists

Value

A vector of output filenames, one for each species.

Examples

```
## Not run:
species <- c("Gray Jay", "Cyanocitta stelleri")
# get the path to the example data included in the package
# in practice, provide path to a filtered ebd file
# e.g. f <- "data/ebd_filtered.txt
f <- system.file("extdata/ebd-sample.txt", package = "auk")
# output to a temporary directory for example
# in practice, provide the path to the output location
# e.g. prefix <- "output/ebd_"
prefix <- file.path(tempdir(), "ebd_")
species_files <- auk_split(f, species = species, prefix = prefix)

## End(Not run)
```

---

auk_state *Filter the eBird data by state*

Description

Define a filter for the eBird Basic Dataset (EBD) based on a set of states This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering.

Usage

```
auk_state(x, state, replace = FALSE)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
state	character; states to filter by. eBird uses 4 to 6 character state codes consisting of two parts, the 2-letter ISO country code and a 1-3 character state code, separated by a dash. For example, "US-NY" corresponds to New York State in the United States. Refer to the data frame ebird_states for look up state codes.
replace	logical; multiple calls to auk_state() are additive, unless replace = FALSE, in which case the previous list of states to filter by will be removed and replaced by that in the current call.

Details

It is not possible to filter by both country and state, so calling auk_state() will reset the country filter to all countries, and vice versa.

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# state codes for a given country can be looked up in ebird_states
dplyr::filter(ebird_states, country == "Costa Rica")
# choose texas, united states and puntarenas, cost rica
states <- c("US-TX", "CR-P")
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_state(states)

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_state(ebd, states)
```

---

auk_time *Filter the eBird data by checklist start time*

Description

Define a filter for the eBird Basic Dataset (EBD) based on a range of start times for the checklist. This function only defines the filter and, once all filters have been defined, auk_filter() should be used to call AWK and perform the filtering.

Usage

```
auk_time(x, start_time)
```

Arguments

x	auk_ebd or auk_sampling object; reference to file created by auk_ebd() or auk_sampling().
start_time	character; 2 element character vector giving the range of times in 24 hour format, e.g. "06:30" or "16:22".

Details

This function can also work with on an auk_sampling object if the user only wishes to filter the sampling event data.

Value

An auk_ebd object.

Examples

```
# only keep checklists started between 6 and 8 in the morning
system.file("extdata/ebd-sample.txt", package = "auk") %>%
  auk_ebd() %>%
  auk_time(start_time = c("06:00", "08:00"))

# alternatively, without pipes
ebd <- auk_ebd(system.file("extdata/ebd-sample.txt", package = "auk"))
auk_time(ebd, start_time = c("06:00", "08:00"))
```

---

auk_unique *Remove duplicate group checklists*

Description

eBird checklists can be shared among a group of multiple observers, in which case observations will be duplicated in the database. This functions removes these duplicates from the eBird Basic Dataset (EBD) or the EBD sampling event data (with `checklists_only = TRUE`), creating a set of unique bird observations. This function is called automatically by `read_ebd()` and `read_sampling()`.

Usage

```
auk_unique(x, group_id = "group_identifier",
  checklist_id = "sampling_event_identifier",
  species_id = "scientific_name", checklists_only = FALSE)
```

Arguments

`x`	data.frame; the EBD data frame, typically as imported by `read_ebd()`.
`group_id`	character; the name of the group ID column.
`checklist_id`	character; the name of the checklist ID column, each checklist within a group will get a unique value for this field. The record with the lowest `checklist_id` will be picked as the unique record within each group.
`species_id`	character; the name of the column identifying species uniquely. This is required to ensure that removing duplicates is done independently for each species. Note that this will not treat sub-species independently and, if that behavior is desired, the user will have to generate a column uniquely identifying species and sub-species and pass that column's name to this argument.
`checklists_only`	logical; whether the dataset provided only contains checklist information as with the sampling event data file. If this argument is `TRUE`, then the `species_id` argument is ignored and removing of duplicated is done at the checklist level not the species level.

Details

This function chooses the checklist within in each that has the lowest value for the field specified by `checklist_id`. A new column is also created, `checklist_id`, whose value is the taken from the field specified in the `checklist_id` parameter for non-group checklists and from the field specified by the `group_id` parameter for grouped checklists.

Value

A data frame with unique observations, and an additional field, `checklist_id`, which is a combination of the sampling event and group IDs.

Examples

```
# read in an ebd file and don't automatically remove duplicates
f <- system.file("extdata/ebd-sample.txt", package = "auk")
ebd <- read_ebd(f, unique = FALSE)
# remove duplicates
ebd_unique <- auk_unique(ebd)
nrow(ebd)
nrow(ebd_unique)
```

---

auk_version_date *Dates of eBird Basic Dataset and taxonomy in package version*

Description

This package depends on the version of the EBD and on the eBird taxonomy. Use this function to determine the version dates for which the package is suitable. The EBD is update quarterly (March 15, June 15, September 15, December 15), while the taxonomy is updated annually in September. To ensure proper functioning, always use the latest version of the auk package and the EBD.

Usage

```
auk_version_date()
```

Value

A date vector with the eBird data date and taxonomy date that this version of the package corresponds to.

Examples

```
auk_version_date()
```

auk_zerofill *Read and zero-fill an eBird data file*

Description

Read an eBird Basic Dataset (EBD) file, and associated sampling event data file, to produce a zero-filled, presence-absence dataset. The EBD contains bird sightings and the sampling event data is a set of all checklists, they can be combined to infer absence data by assuming any species not reported on a checklist was had a count of zero.

Usage

```
auk_zerofill(x, ...)

## S3 method for class 'data.frame'
auk_zerofill(x, sampling_events, species, unique = TRUE,
  collapse = FALSE, ...)

## S3 method for class 'character'
auk_zerofill(x, sampling_events, species, sep = "\t",
  unique = TRUE, collapse = FALSE, ...)

## S3 method for class 'auk_ebd'
auk_zerofill(x, species, sep = "\t", unique = TRUE,
  collapse = FALSE, ...)

collapse_zerofill(x)
```

Arguments

x	filename, data.frame of eBird observations, or auk_ebd object with associated output files as created by auk_filter(). If a filename is provided, it must point to the EBD and the sampling_events argument must point to the sampling event data file. If a data.frame is provided it should have been imported with read_ebd(), to ensure the variables names have been set correctly, and it must have been passed through auk_unique() to ensure duplicate group checklists have been removed.
...	additional arguments passed to methods.
sampling_events	character or data.frame; filename for the sampling event data or a data.frame of the same data. If a data.frame is provided it should have been imported with read_sampling(), to ensure the variables names have been set correctly, and it must have been passed through auk_unique() to ensure duplicate group checklists have been removed.
species	character; species to include in zero-filled dataset, provided as scientific or English common names, or a mixture of both. These names must match the official

	eBird Taxomony (ebird_taxonomy). To include all species, leave this argument blank.
unique	logical; should auk_unique() be run on the input data if it hasn't already.
collapse	logical; whether to call collapse_zerofill() to return a data frame rather than an auk_zerofill object.
sep	character; single character used to separate fields within a row.

Details

auk_zerofill() generates an auk_zerofill object consisting of a list with elements observations and sampling_events. observations is a data frame giving counts and binary presence/absence data for each species. sampling_events is a data frame with checklist level information. The two data frames can be connected via the checklist_id field. This format is efficient for storage since the checklist columns are not duplicated for each species, however, working with the data often requires joining the two data frames together.

To return a data frame, set collapse = TRUE. Alternatively, zerofill_collapse() generates a data frame from an auk_zerofill object, by joining the two data frames together to produce a single data frame in which each row provides both checklist and species information for a sighting.

Value

By default, an auk_zerofill object, or a data frame if collapse = TRUE.

Methods (by class)

- data.frame: EBD data frame.
- character: Filename of EBD.
- auk_ebd: auk_ebd object output from auk_filter(). Must have had a sampling event data file set in the original call to auk_ebd().

Examples

```
# read and zero-fill the sampling data
f_ebd <- system.file("extdata/zerofill-ex_ebd.txt", package = "auk")
f_smpl <- system.file("extdata/zerofill-ex_sampling.txt", package = "auk")
auk_zerofill(x = f_ebd, sampling_events = f_smpl)

# use the species argument to only include a subset of species
auk_zerofill(x = f_ebd, sampling_events = f_smpl,
             species = "Collared Kingfisher")

# to return a data frame use collapse = TRUE
ebd_df <- auk_zerofill(x = f_ebd, sampling_events = f_smpl, collapse = TRUE)
```

ebird_species *Lookup species in eBird taxonomy*

Description

Given a list of common or scientific names, check that they appear in the official eBird taxonomy and convert them all to scientific names, or common names if `scientific = FALSE`. Un-matched species are returned as `NA`.

Usage

```
ebird_species(x, scientific = TRUE)
```

Arguments

`x`	character; species to look up, provided as scientific or English common names, or a mixture of both. Case insensitive.
`scientific`	logical; whether to return scientific (`TRUE`) or English common names (`FALSE`).

Value

Character vector of scientific names or common names if names if `scientific = FALSE`.

Examples

```
# mix common and scientific names, case-insensitive
species <- c("Blackburnian Warbler", "Poecile atricapillus",
             "american dipper", "Caribou")
# species not in the ebird taxonomy return NA
ebird_species(species)
```

ebird_states *eBird States*

Description

A data frame of state codes used by eBird. These codes are 4 to 6 characters, consisting of two parts, the 2-letter ISO country code and a 1-3 character state code, separated by a dash. For example, `"US-NY"` corresponds to New York State in the United States. These state codes are required to filter by state using `auk_state()`.

Usage

```
ebird_states
```

Format

A data frame with four variables and 3,145 rows:

- `country`: short form of English country name.
- `country_code`: 2-letter ISO country code.
- `state`: state name.
- `state_code`: 4 to 6 character state code.

Details

Note that some countries are not broken into states in eBird and therefore do not appear in this data frame.

`ebird_taxonomy` *eBird Taxonomy*

Description

A simplified version of the taxonomy used by eBird. Includes proper species as well as various other categories such as `spuh` (e.g. *duck sp.*) and *slash* (e.g. *American Black Duck/Mallard*). This taxonomy is based on the Clements Checklist, which is updated annually, typically in the late summer. Non-ASCII characters (e.g. those with accents) have been converted to ASCII equivalents in this data frame.

Usage

```
ebird_taxonomy
```

Format

A data frame with eight variables and 15,251 rows:

- `taxon_order`: numeric value used to sort rows in taxonomic order.
- `category`: whether the entry is for a species or another field-identifiable taxon, such as `spuh`, `slash`, `hybrid`, etc.
- `species_code`: a unique alphanumeric code identifying each species.
- `common_name`: the common name of the species as used in eBird.
- `scientific_name`: the scientific name of the species.
- `order`: the scientific name of the order that the species belongs to.
- `family`: the family of the species, in the form "Parulidae (New World Warblers)".
- `report_as`: for taxa that can be resolved to true species (i.e. species, subspecies, and recognizable forms), this field links to the corresponding species code. For taxa that can't be resolved, this field is `NA`.

For further details, see http://help.ebird.org/customer/en/portal/articles/1006825-the-ebird-taxonomy

read_ebd *Read an EBD file*

Description

Read an eBird Basic Dataset file using `data.table::fread()`, `readr::read_delim()`, or `read.delim()` depending on which packages are installed. `read_ebd()` reads the EBD itself, while read_sampling()‘ reads a sampling event data file.

Usage

```
read_ebd(x, reader, sep = "\t", unique = TRUE, rollup = TRUE)

## S3 method for class 'character'
read_ebd(x, reader, sep = "\t", unique = TRUE,
  rollup = TRUE)

## S3 method for class 'auk_ebd'
read_ebd(x, reader, sep = "\t", unique = TRUE,
  rollup = TRUE)

read_sampling(x, reader, sep = "\t", unique = TRUE)

## S3 method for class 'character'
read_sampling(x, reader, sep = "\t", unique = TRUE)

## S3 method for class 'auk_ebd'
read_sampling(x, reader, sep = "\t", unique = TRUE)

## S3 method for class 'auk_sampling'
read_sampling(x, reader, sep = "\t", unique = TRUE)
```

Arguments

x	filename or auk_ebd object with associated output files as created by `auk_filter()`.
reader	character; the function to use for reading the input file, options are `"fread"`, `"readr"`, or `"base"`, for `data.table::fread()`, `readr::read_delim()`, or `read.delim()`, respectively. This argument should typically be left empty to have the function choose the best reader based on the installed packages.
sep	character; single character used to separate fields within a row.
unique	logical; should duplicate grouped checklists be removed. If `unique = TRUE`, `auk_unique()` is called on the EBD before returning.
rollup	logical; should taxonomic rollup to species level be applied. If `rollup = TRUE`, `auk_rollup()` is called on the EBD before returning. Note that this process can be time consuming for large files, try turning rollup off if reading is taking too long.

Details

This functions performs the following processing steps:

- Data types for columns are manually set based on column names used in the February 2017 EBD. If variables are added or names are changed in later releases, any new variables will have data types inferred by the import function used.
- Variables names are converted to `snake_case`.
- Duplicate observations resulting from group checklists are removed using `auk_unique()`, unless `unique = FALSE`.

Value

A data frame of EBD observations. An additional column, `checklist_id`, is added to output files if `unique = TRUE`, that uniquely identifies the checklist from which the observation came. This field is equal to `sampling_event_identifier` for non-group checklists, and `group_identifier` for group checklists.

Methods (by class)

- `character`: Filename of EBD.
- `auk_ebd`: `auk_ebd` object output from `auk_filter()`
- `character`: Filename of sampling event data file
- `auk_ebd`: `auk_ebd` object output from `auk_filter()`. Must have had a sampling event data file set in the original call to `auk_ebd()`.
- `auk_sampling`: `auk_sampling` object output from `auk_filter()`.

Examples

```
f <- system.file("extdata/ebd-sample.txt", package = "auk")
read_ebd(f)
# read a sampling event data file
x <- system.file("extdata/zerofill-ex_sampling.txt", package = "auk") %>%
  read_sampling()
```

Index