wsl.gbif

October 18, 2022

Type Package

Title A toolbox to efficiently download and filter large GBIF observational datasets for sound spatial analyses

Version 0.1.0

Depends R (>= 4.0.0), raster, terra, rgbif, CoordinateCleaner

Description Package aiming at easing the workflow of retrieving GBIF observations at large spatial scale for all species accepted names and synonyms, and filtering them according to the specific scale of an analysis.

License GPL (>=3)

BugReports https://github.com/8Ginette8/wsl.gbif/issues

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Authors Yohann Chauvier [cre,aut] (https://orcid.org/0000-0001-9399-3192)

Collate 'make_tiles.R'
'wsl_gbif.R'
'wsl_taXnames.R'

R topics documented:

`make_tiles` *Create a specific number of tiles based on an extent*

Description

Not to be called directly by the user

Usage

```
make_tiles(geo, Ntiles, meta = TRUE)
```

Author(s)

Yohann Chauvier

`wsl_gbif` *Massively download and filter GBIF observations for sound spatial analyses*

Description

Implement an user-friendly workflow to download and clean gbif taxa observations. The function uses the rgbif R package but (1) implements the same search result found if www.gbif.org is employed i.e., based on the input taxa name, all species records related to its accepted name and synonyms are extracted. The function also (2) bypasses rgbif hard limit on the number of records (100'000 max). For this purpose, a dynamic moving window is created and used across the geographic extent defined by the user. This window automatically fragments the specified study area in succesive tiles of different sizes, until all tiles include < 100'000 observations. The function also (3) automatically applies a post-filtering of observations based on the chosen resolution of the study/analysis and by partly employing the CoordinateCleaner R package. Filtering options may be chosen and involve several choices: study's extent, removal of duplicates, removal of absences, basis of records selection, removal of invalid/uncertain xy coordinates (WGS84), time period selection and removal of raster centroids. By default, the argument hasGeospatialIssue in occ_search() (implemented rgbif function) is set to FALSE. To get the custom DOI of the downloaded GBIF data, the derived_dataset() function from the rgbif package must be used with the column 'datasetKey' of one or several outputs.

Usage

```
wsl_gbif(
  sp_name = NULL,
  conf_match = 90,
  geo = NULL,
  grain = 1000,
  duplicates = FALSE,
  absences = FALSE,
  no_xy = FALSE,
  basis = c("OBSERVATION", "HUMAN_OBSERVATION", "MACHINE_OBSERVATION",
   "MATERIAL_SAMPLE", "PRESERVED_SPECIMEN", "FOSSIL_SPECIMEN", "LIVING_SPECIMEN",
     "LITERATURE", "UNKNOWN"),
  add_infos = NULL,
  time_period = c(1000, 3000),
  identic_xy = FALSE,
  wConverted_xy = FALSE,
  centroids = FALSE,
  ntries = 10,
  error.skip = TRUE,
  ...
)
```

Arguments

sp_name	Character. Scientific name to run an online search (i.e. with GBIF-API) for species observations. Works also for genus and higher taxa levels.
conf_match	Numeric from 0 to 100. Determine the confidence threshold of match of 'sp_name' with the GBIF backbone taxonomy. Default is 90.
geo	Object of class 'Extent', 'SpatExtent', 'SpatialPolygon', 'SpatialPolygonDataframe', or 'SpaVector' (WGS84) to define the study's area extent. Default is NULL i.e. the whole globe.
grain	Numeric. Specify in meters the study resolution. Used to filter gbif records (x2) according to their uncertainties and number of coordinate decimals. Records with no information on coordinate uncertainties (column 'coordinateUncertaintyInMeters') are be kept by default. See details.
duplicates	Logical. Should duplicated records be kept?
absences	Logical. Should absence records be kept?
no_xy	Logical. Default is FALSE i.e. only records with coordinates are downloaded. If TRUE, only records with no coordinates are downloaded.
basis	Character. Which basis of records should be selected? Default is all i.e. c("OBSERVATION", "HUMAN_OBSERVATION", "MACHINE_OBSERVATION", "MATERIAL_SAMPLE", "PRESERVED_SPECIMEN", "FOSSIL_SPECIMEN", "LIVING_SPECIMEN", "LITERATURE", "UNKNOWN"). Description may be found here: https://docs.gbif.org/course-data-use/en/basis-of-record.html
add_infos	Character. Infos that may be added to the default output information. List of IDs may be found at: https://www.gbif.org/developer/occurrence. Default IDs contain 'taxonKey', 'scientificName', 'acceptedTaxonKey', 'acceptedScientificName', 'individualCount', 'decimalLatitude', 'decimalLongitude', 'basisOfRecord', 'coordinateUncertaintyInMeters', 'country', 'year', 'datasetKey', 'institutionCode', 'publishingOrgKey', 'taxonomicStatus' and 'taxonRank'.
time_period	Numerical vector. Observations will be downloaded according to the chosen year range. Default is c(1000,3000). Observations with year = NA are kept by default.
identic_xy	Logical. Should records with identical xy be kept?
wConverted_xy	Logical. Should incorrectly lon/lat converted xy be kept? Uses 'cd_ddmm' from 'CoordinateCleaner' R package.
centroids	Logical. Should species records from raster centroids be kept? Uses 'cd_round' from 'CoordinateCleaner' R package.
ntries	Numeric. In case of failure from GBIF server or within the rgbif package, how many download attempts should the function request? Default is '10' with a 2 seconds interval between tries. If attempts failed, an empty data.frame is return by default.
error.skip	Logical. Should the search process continues if ntries failed ?
...	Additonnal parameters for the function 'cd_round' of the 'CoordinateCleaner' R package.

Details

Argument 'grain' used for two distinct gbif records filtering. (1) Records filtering according to gbif 'coordinateUncertaintyInMeters'; every records uncertainty > grain/2 are removed. Note: Records

with no information on coordinate uncertainties are kept by default. (2) Records filtering according to the number of longitude/latitude decimals; if 110km < grain <= 11km, lon/lat with >= 1 decimal are kept, if 11km < grain <= 1100m, lon/lat with >= 2 decimals kept; if 1100m < grain <= 110m, lon/lat with >= 3 decimals are kept; if 110m < grain <= 11m, lon/lat with >= 4 decimals are kept; if 11m < grain <= 1.1m, lon/lat with >= 5 decimals are kept etc...

Value

Object of class 'data.frame' with requested GBIF information. Although the function works accurately, error outputs might still occur depending on the 'sp_name' used. Therefore, default information detailed in 'add_infos' is stored so that sanity checks may still be applied afterwards. Although crucial preliminary checks of species records are done by the function, addtional post exploration with the 'CoordinateCleaner' R package is still highly recommended.

Author(s)

Yohann Chauvier

Examples

```
# Necessary libraries
#library(raster)
#library(terra)
#library(rgbif)
#library(CoordinateCleaner)

# Load the Alps Extend
data(AlpineConvention_lonlat)

# Downloading worldwide the observations of Panthera tigris
test1 = wsl_gbif("Panthera tigris")

# Downloading in the Alps the observations of Cypripedium calceolus (with a 100m grain and
# by adding the 'issues' column)
test3 = wsl_gbif("Cypripedium calceolus", geo = shp.lonlat, grain = 100, add_infos = c("issue"))
plot(shp.lonlat)
points(test1[,c("decimalLongitude","decimalLatitude")],pch=20,col="#238b4550",cex=1)

# Downloading worlwide the observations of Ailuropoda melanoleuca (with a 100km grain, after
# 1990 and by keeping duplicates and by adding the name of the person who collected the species records)
test3 = wsl_gbif("Ailuropoda melanoleuca", grain = 100000 , duplicates = TRUE,
    time_period = c(1990,3000), add_infos = c("recordedBy","issue"))

# Downloading worlwide the observations of Phascolarctos cinereus (with a 1km grain, after 1980,
# and keeping raster centroids)
test4 = wsl_gbif("Phascolarctos cinereus", grain = 1000,
    time_period = c(1990,3000), centroids = TRUE)

# Just an example on how to retrieve the DOI for the first downloaded dataset using
# derived_dataset() from the rgbif R package. Note that multiple datasets may be combined
# and derived_dataset() used once to only obtain one unique DOI.
d.target = table(test1$datasetKey)
d.summary = data.frame(datasetKey = names(d.target),count = as.numeric(d.target))
rgbif::derived_dataset(d.summary,"GBIF_test",
    "Filetred and cleaned based on CoordinateCleaner",source_url="https://example.com/",
```

```
user="your_gbif_user",pwd="your_gbif_password")
```

`wsl_taXnames` *Retrieve from GBIF all scientific names of a specific Taxa*

Description

Allows to extract from the gbif backbone taxonomy all names from an input species name (accepted, synonyms, children, related...).

Usage

```
wsl_taXnames(sp_name = NULL, conf_match = 90, all = FALSE)
```

Arguments

`sp_name`	Character. Species name from which the user wants to retrieve all existing GBIF names
`conf_match`	Numeric. From 0 to 100. Determine the confidence threshold of match of 'sp_name' with the GBIF backbone taxonomy. Default is 90.
`all`	Logical. Default is FALSE. Should all species names be retrieved or only the accepted name and its synonyms?

Value

A data.frame with two columns: (1) Names and (2) Backbone Taxonomy Status

Author(s)

Yohann Chauvier

Examples

```
wsl_taXnames("Cypripedium calceolus",all=FALSE)
wsl_taXnames("Cypripedium calceolus",all=TRUE)
```

Index