

wsl.gbif

October 19, 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

Maintainer Yohann Chauvier <yohann.chauvier@wsl.ch>

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_taNames.R'

'wsl_doi.R'

'wsl_obs_filter.R'

R topics documented:

make_tiles	2
wsl_doi	3
wsl_gbif	4
wsl_obs_filter	7
wsl_taNames	8

Index	9
--------------	----------

make_tiles

Create a specific number of tiles based on a raster extent

Description

Based on a specific extent, one or several tiles are generated. Tiles can be smaller numerical extents or geometry arguments POLYGON(). The original extent is therefore either converted into a POLYGON() argument, or divided into Ntiles of regular fragments which are converted into POLYGON() arguments and smaller numerical extents.

Usage

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

Arguments

geo	Object of class 'Extent', 'SpatExtent', 'SpatialPolygon', 'SpatialPolygonDataframe', or 'SpaVector' (WGS84 or planar) to define the study's area extent. Default is NULL i.e. the whole globe.
Ntiles	Numeric. In how many tiles/fragments should geo be divided approximately?
sext	Logical. Should a list of numerical extents 'c(xmin,xmax,ymin,ymax)' also be returned for each generated POLYGON()?

Value

A list of geometry arguments POLYGON() of length Ntiles (and of numerical extents if sext=TRUE)

Author(s)

Yohann Chauvier

References

Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., ... & Zimmermann, N. E. (2021). Influence of climate, soil, and land cover on plant species distribution in the European Alps. Ecological monographs, 91(2), e01433. 10.1002/ecm.1433

Examples

```
# Load the European Alps Extent
data(geo_dat)

# Apply the function to divide the extent in ~20 fragments
make_tiles(geo=shp.lonlat,Ntiles=20,sext=TRUE)
```

wsl_doi*Get a custom DOI for a GBIF filtered dataset*

Description

A small user friendly wrapper of the `derived_dataset()` function of the `rgbif` R package, compatible with one or several `wsl_gbif()` outputs.

Usage

```
wsl_doi(  
  wsl_gbif = list(),  
  title = NULL,  
  description = NULL,  
  source_url = "https://example.com/",  
  usr = "",  
  pwd = "",  
  ...  
)
```

Arguments

<code>wsl_gbif</code>	List. List of one or several <code>wsl_gbif</code> outputs.
<code>title</code>	The title for your derived dataset.
<code>source_url</code>	A link to where the dataset is stored.
<code>pwd</code>	Your GBIF password.
<code>...</code>	Additonnal parameters for <code>derived_dataset()</code> in <code>rgbif</code> . R package.
<code>descriptipion</code>	A description of the dataset.
<code>user</code>	Your GBIF username.

Details

see `derived_dataset()` function from the `rgbif` R package

Value

One citable DOI and its information.

References

Chamberlain, S., Oldoni, D., & Waller, J. (2022). `rgbif`: interface to the global biodiversity information facility API. 10.5281/zenodo.6023735

See Also

The `rgbif` package for additional and more general approaches to get GBIF DOI

Examples

```
# Downloading worldwide the observations of Panthera tigris and Ailuropoda melanoleuca
test1 = wsl_gbif("Panthera tigris")
test32 = wsl_gbif("Ailuropoda melanoleuca")

# Just an example on how to retrieve the DOI for only one dataset

# Just an example on how to retrieve the DOI for only one dataset

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_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 successive 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 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.

<code>ntries</code>	Numeric. In case of failure from GBIF server or within the <code>rgbif</code> package, how many download attempts should the function request? Default is '10' with a 2 seconds interval between tries. If attempts failed, an empty <code>data.frame</code> is returned by default.
<code>error.skip</code>	Logical. Should the search process continue if <code>ntries</code> failed?
<code>...</code>	Additional parameters for the function <code>cd_round()</code> of <code>CoordinateCleaner</code> .

Details

Argument 'grain' used for two distinct `gbif` records filtering. (1) Records filtering according to `gbif` 'coordinateUncertaintyInMeters'; every record with uncertainty $> \text{grain}/2$ is 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 $110\text{km} < \text{grain} \leq 11\text{km}$, lon/lat with ≥ 1 decimal are kept; if $11\text{km} < \text{grain} \leq 1100\text{m}$, lon/lat with ≥ 2 decimals are kept; if $1100\text{m} < \text{grain} \leq 110\text{m}$, lon/lat with ≥ 3 decimals are kept; if $110\text{m} < \text{grain} \leq 11\text{m}$, lon/lat with ≥ 4 decimals are kept; if $11\text{m} < \text{grain} \leq 1.1\text{m}$, 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, additional post exploration with the `CoordinateCleaner` R package is still highly recommended.

References

- Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., ... & Zimmermann, N. E. (2021). Influence of climate, soil, and land cover on plant species distribution in the European Alps. *Ecological monographs*, 91(2), e01433. [10.1002/ecm.1433](https://doi.org/10.1002/ecm.1433)
- Chamberlain, S., Oldoni, D., & Waller, J. (2022). `rgbif`: interface to the global biodiversity information facility API. [10.5281/zenodo.6023735](https://doi.org/10.5281/zenodo.6023735)
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... & Antonelli, A. (2019). `CoordinateCleaner`: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10(5), 744-751. [10.1111/2041-210X.13152](https://doi.org/10.1111/2041-210X.13152)
- Hijmans, Robert J. "terra: Spatial Data Analysis. R Package Version 1.6-7." (2022). Terra - CRAN

See Also

The (1) `rgbif` and (2) `CoordinateCleaner` packages for additional and more general approaches on (1) downloading GBIF observations and (2) post-filtering those.

Examples

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

# Load the Alps Extent
data(geo_dat)
```

```
# 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 worldwide 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 worldwide 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)
```

wsl_obs_filter

Filter a set of GBIF observations according to a defined grain

Description

Filter a set of GBIF observations through a chosen raster grid that defines the resolution of the spatial analysis. Depending on the desired resolution of the analysis, one might want to keep only one GBIF observation per grid cell to avoid e.g., modelling and sampling bias issues.

Usage

```
wsl_obs_filter(wsl.gbif, grid)
```

Arguments

wsl.gbif	one wsl.gbif output including one or several species
grid	Object of class 'SpatRaster', 'RasterLayer', 'RasterBrick' or 'RasterStack' of desired resolution and extent.

Value

a data frame with two columns named 'x' and 'y' comprising the new set of observations filtered at grid resolution.

Examples

```
# Load the European Alps extent and a raster of a random resolution
data(geo_dat)
data(exrst)

# Downloading in the European Alps the observations of Arctostaphylos alpinus
```

```
obs.arcto = wsl_gbif("Arctostaphylos alpinus", geo=shp.lonlat)
plot(shp.lonlat)
points(obs.arcto[,c("decimalLongitude", "decimalLatitude")], pch=20, col="#238b4550", cex=1)
```

wsl_taNAMES

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_taNAMES(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

References

Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., ... & Zimmermann, N. E. (2021). Influence of climate, soil, and land cover on plant species distribution in the European Alps. *Ecological monographs*, 91(2), e01433. 10.1002/ecm.1433

Chamberlain, S., Oldoni, D., & Waller, J. (2022). *rgbif*: interface to the global biodiversity information facility API. 10.5281/zenodo.6023735

See Also

The *rgbif* package for additional and more general approaches on how to retrieve scientific names from the GBIF backbone taxonomy.

Examples

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


Index

`make_tiles`, [2](#)

`wsl_doi`, [3](#)

`wsl_gbif`, [4](#)

`wsl_obs_filter`, [7](#)

`wsl_taxnames`, [8](#)