Package 'insee'

August 26, 2024

Type Package

Title Tools to Easily Download Data from INSEE BDM Database

Version 1.1.7

Description Using embedded sdmx queries, get the data of more than 150 000 insee series from 'bdm' macroeconomic database.

URL https://pyr-opendatafr.github.io/R-Insee-Data/

Encoding UTF-8

License MIT + file LICENSE

VignetteBuilder knitr

BugReports https://github.com/pyr-opendatafr/R-Insee-Data/issues

Imports httr, xml2, tibble, dplyr, stringr, tidyselect (>= 1.2.0), rlang, purrr, crayon, openssl, rappdirs

Suggests lubridate, ggplot2, prettydoc, htmltools, kableExtra, knitr, rmarkdown, markdown, magrittr, testthat, covr, png

RoxygenNote 7.3.2

Depends R (>= 2.10)

NeedsCompilation no

Author Hadrien Leclerc [aut, cre], INSEE [cph]

Maintainer Hadrien Leclerc <leclerc.hadrien@gmail.com>

Repository CRAN

Date/Publication 2024-08-26 07:30:01 UTC

Contents

`add_insee_metadata` *Add metadata to the raw data*

Description

Add metadata to the raw data

Usage

```
add_insee_metadata(df)
```

Arguments

`df`	a dataframe containing data obtained from get_insee_idbank or get_insee_dataset

Details

Add metadata to the raw data obtained from get_insee_idbank or get_insee_dataset

Value

a tibble with the data given as parameter plus the corresponding metadata

Examples

```
library(magrittr)

data =
 get_insee_idbank("001694061") %>%
 add_insee_metadata()
```

`add_insee_title` *Add a title column to the idbank list dataset*

Description

Add a title column to the idbank list dataset

Usage

```
add_insee_title(df, n_split, lang = "en", split = TRUE, clean = TRUE)
```

Arguments

`df`	a dataframe containing an idbank column called "idbank" or "IDBANK"
`n_split`	number of new columns, by default the maximum is chosen
`lang`	returns an English title, by default is "en", any other value returns a French title
`split`	split the title column in several columns, by default is TRUE
`clean`	remove the columns filled with NA (missing value), by default is TRUE

Details

this function uses extensively the get_insee_title function. Then, it should be used on an already filtered dataset, not on the full idbank dataset (cf. get_insee_title). The number of separators in the official INSEE title can vary and is not normalized. Beware all title columns created may not be a cleaned dimension label.

Value

the same dataframe but with one or several title columns

Examples

```
library(magrittr)
library(dplyr)

idbank_empl =
 get_idbank_list("EMPLOI-SALARIE-TRIM-NATIONAL") %>% #employment
 slice(1:15) %>%
 add_insee_title()
```

get_column_title *Get the title of dataset's columns*

Description

Get the title of dataset's columns

Usage

```
get_column_title(dataset = NULL)
```

Arguments

dataset	an INSEE's dataset, if NULL

Value

a dataframe

Examples

```
column_titles_all_dataset = get_column_title()

column_titles = get_column_title("CNA-2014-CONSO-MEN")
```

get_dataset_list *Download a full INSEE's dataset list*

Description

Download a full INSEE's dataset list

Usage

```
get_dataset_list()
```

Details

the datasets returned are the ones available through a SDMX query

Value

a tibble with 5 columns : id, Name.fr, Name.en, url, n_series

Examples

```
insee_dataset = get_dataset_list()
```

`get_idbank_list` *Download a full INSEE's series key list*

Description

Download a full INSEE's series key list

Usage

```
get_idbank_list(..., dataset = NULL, update = FALSE)
```

Arguments

`...`	one or several dataset names
`dataset`	if a dataset name is provided, only a subset of the data is delivered, otherwise all the data is returned, and column names refer directly to data dimensions
`update`	It is FALSE by default, if it is set to TRUE, it triggers the metadata update. This update is automatically triggered once every 6 months.

Details

Download a mapping dataset between INSEE series keys (idbank) and SDMX series names. Under the hood the get_idbank_list uses download.file function from utils, the user can change the mode argument with the following command : Sys.getenv(INSEE_download_option_idbank_list = "wb") If INSEE makes an update, the user can also change the zip file downloaded, the data file contained in the zip and data the separator : Sys.setenv(INSEE_idbank_dataset_path = "new_zip_file_link") Sys.setenv(INSEE_idbank_sep = ",") Sys.setenv(INSEE_idbank_dataset_file = "new_data_file_name")

Value

a tibble the idbank dataset

Examples

```
# download datasets list
dt = get_dataset_list()
# use a dataset name to retrieve the series key list related to the dataset
idbank_list = get_idbank_list('CNT-2014-PIB-EQB-RF')
```

get_insee *Get data from INSEE BDM database with a SDMX query link*

Description

Get data from INSEE BDM database with a SDMX query link

Usage

```
get_insee(link, step = "1/1")
```

Arguments

link	SDMX query link
step	argument used only for internal package purposes to tweak download display

Details

Get data from INSEE BDM database with a SDMX query link. This function is mainly for package internal use. It is used by the functions get_insee_dataset, get_insee_idbank and get_dataset_list. The data is cached, hence all queries are only run once per R session. The user can disable the download display in the console with the following command : Sys.setenv(INSEE_download_verbose = "FALSE"). The use of cached data can be disabled with : Sys.setenv(INSEE_no_cache_use = "TRUE"). All queries are printed in the console with this command: Sys.setenv(INSEE_print_query = "TRUE").

Value

a tibble containing the data

Examples

```
insee_link = "http://www.bdm.insee.fr/series/sdmx/data/SERIES_BDM"
insee_query = file.path(insee_link, paste0("010539365","?", "firstNObservations=1"))
data = get_insee(insee_query)
```

get_insee_dataset *Get dataset from INSEE BDM database*

Description

Get dataset from INSEE BDM database

Usage

```
get_insee_dataset(
  dataset,
  startPeriod = NULL,
  endPeriod = NULL,
  firstNObservations = NULL,
  lastNObservations = NULL,
  includeHistory = NULL,
  updatedAfter = NULL,
  filter = NULL
)
```

Arguments

dataset	dataset name to be downloaded
startPeriod	start date of data
endPeriod	end date of data
firstNObservations	get the first N observations for each key series (idbank)
lastNObservations	get the last N observations for each key series (idbank)
includeHistory	boolean to access the previous releases (not available on all series)
updatedAfter	starting point for querying the previous releases (format yyyy-mm-ddThh:mm:ss)
filter	Use the filter to choose only some values in a dimension. It is recommended to use it for big datasets. A dimension left empty means all values are selected. To select multiple values in one dimension put a "+" between those values (see example)

Details

Get dataset from INSEE BDM database

Value

a tibble with the data

Examples

```
insee_dataset = get_dataset_list()
idbank_ipc = get_idbank_list("IPC-2015")

#example 1
data = get_insee_dataset("IPC-2015", filter = "M+A.........CVS..", startPeriod = "2015-03")

#example 2
data = get_insee_dataset("IPC-2015", filter = "A..SO...VARIATIONS_A....BRUT..SO",
 includeHistory = TRUE, updatedAfter = "2017-07-11T08:45:00")
```

`get_insee_idbank` *Get data from INSEE series idbank*

Description

Get data from INSEE series idbank

Usage

```
get_insee_idbank(
  ...,
  limit = TRUE,
  startPeriod = NULL,
  endPeriod = NULL,
  firstNObservations = NULL,
  lastNObservations = NULL,
  includeHistory = NULL,
  updatedAfter = NULL
)
```

Arguments

`...`	one or several series key (idbank)
`limit`	by default, the function get_insee_idbank has a 1200-idbank limit. Set limit argument to FALSE to ignore the limit or modify the limit with the following command : Sys.setenv(INSEE_idbank_limit = 1200)
`startPeriod`	start date of data
`endPeriod`	end date of data
`firstNObservations`	get the first N observations for each key series (idbank)
`lastNObservations`	get the last N observations for each key series (idbank)
`includeHistory`	boolean to access the previous releases (not available on all series)
`updatedAfter`	starting point for querying the previous releases (format yyyy-mm-ddThh:mm:ss)

Details

Get data from INSEE series idbanks. The user can disable the download display in the console with the following command : Sys.setenv(INSEE_download_verbose = "FALSE")

Value

a tibble with the data

Examples

```
#example 1 : import price index of industrial products and turnover index : manufacture of wood
data = get_insee_idbank("001558315", "010540726")

#example 2 : unemployment data

library(magrittr)
library(dplyr)
library(ggplot2)

df_idbank_list_selected =
  get_idbank_list("CHOMAGE-TRIM-NATIONAL") %>%  #unemployment dataset
  filter(SEXE == 0) %>% #men and women
  add_insee_title()

idbank_list_selected = df_idbank_list_selected %>% pull(idbank)

unem = get_insee_idbank(idbank_list_selected)

#example 3 : French GDP growth rate

df_idbank_list_selected =
  get_idbank_list("CNT-2014-PIB-EQB-RF") %>%  # Gross domestic product balance
  filter(FREQ == "T") %>% #quarter
  filter(OPERATION == "PIB") %>% #GDP
  filter(NATURE == "TAUX") %>% #rate
  filter(CORRECTION == "CVS-CJO") #SA-WDA, seasonally adjusted, working day adjusted

idbank = df_idbank_list_selected %>% pull(idbank)

data = get_insee_idbank(idbank) %>%
  add_insee_metadata()

#plot
ggplot(data, aes(x = DATE, y = OBS_VALUE)) +
geom_col() +
ggtitle("French GDP growth rate, quarter-on-quarter, sa-wda") +
labs(subtitle = sprintf("Last updated : %s", data$TIME_PERIOD[1]))
```

`get_insee_title` *Get title from INSEE series idbank*

Description

Get title from INSEE series idbank

Usage

```
get_insee_title(..., lang = "en")
```

Arguments

...	list of series key (idbank)
lang	language of the title, by default it is Engligh, if lang is different from "en" then French will be the title's language

Details

Query INSEE website to get series title from series key (idbank). Any query to INSEE database can handle around 400 idbanks at maximum, if necessary the idbank list will then be splitted in several lists of 400 idbanks each. Consequently, it is not advised to use it on the whole idbank dataset, the user should filter the idbank dataset first.

Value

a character vector with the titles

Examples

```
#example 1 : industrial production index on manufacturing and industrial activities
title = get_insee_title("010537900")

#example 2 : automotive industry and overall industrial production
library(magrittr)
library(dplyr)
library(stringr)

idbank_list_selected =
  get_idbank_list("IPI-2015") %>% #industrial production index dataset
  filter(FREQ == "M") %>% #monthly
  filter(NATURE == "INDICE") %>% #index
  filter(CORRECTION  == "CVS-CJO") %>% #Working day and seasonally adjusted SA-WDA
 filter(str_detect(NAF2,"^29$|A10-BE")) %>% #automotive industry and overall industrial production
  mutate(title = get_insee_title(idbank))
```

---

search_insee *Search a pattern among insee datasets and idbanks*

Description

Search a pattern among insee datasets and idbanks

Usage

```
search_insee(pattern = ".*")
```

Arguments

pattern	string used to filter the dataset and idbank list

Details

The data related to idbanks is stored internally in the package and might the most up to date. The function ignores accents and cases.

Value

the dataset and idbank table filtered with the pattern

Examples

```
# example 1 : search one pattern, the accents do not matter
writeLines("the word 'enqu\U00EAte' (meaning survey in French) will match with 'enquete'")
dataset_enquete = search_insee("enquete")

# example 2 : search multiple patterns
dataset_survey_gdp = search_insee("Survey|gdp")

# example 3 : data about paris
data_paris = search_insee('paris')

# example 4 : all data
data_all = search_insee()
```

---

split_title *Split the title column in several columns*

Description

Split the title column in several columns

Usage

```
split_title(df, title_col_name, pattern, n_split = "max", lang = NULL)
```

Arguments

df	a dataframe containing a title column
title_col_name	the column name to be splitted, if missing it will be either TITLE_EN
pattern	the value by default is stored in the package and it is advised to use it, but in some cases it is useful to use one's pattern
n_split	number of new columns, by default the maximum is chosen
lang	by default it returns both the French and the English title provided by INSEE

Details

The number of separators in the official INSEE title can vary and is not normalized. Beware all title columns created may not be a cleaned dimension label.

Value

the same dataframe with the title column splitted

Examples

```
library(magrittr)

# quarterly payroll enrollment in the construction sector
data_raw = get_insee_idbank("001577236")

data = data_raw %>%
  split_title()
```

Index