

# Package ‘restatapi’

September 18, 2020

**Type** Package

**Title** Search and Retrieve Data from Eurostat Database

**Date** 2020-09-17

**Version** 0.8.10

**Encoding** UTF-8

**Description** Eurostat is the statistical office of the European Union and provides high quality statistics for Europe.

Large set of the data is disseminated through the Euro-

stat database (<<https://ec.europa.eu/eurostat/data/database>>).

The tools are using the REST API with the Statistical Data and Metadata eX-

change (SDMX <<https://sdmx.org>>) Web Services

(<<https://ec.europa.eu/eurostat/web/sdmx-web-services/about-this-service>>) to search and download data from

the Eurostat database using the SDMX standard.

**License** EUPL-1.1

**Imports** data.table, rjson, xml2

**Suggests** chron, knitr, rmarkdown, testthat

**LazyData** true

**NeedsCompilation** no

**URL** <https://github.com/eurostat/restatapi>

**BugReports** <https://github.com/eurostat/restatapi/issues>

**RoxygenNote** 7.1.1

**Author** Mátyás Mészáros [aut, cre]

**Maintainer** Mátyás Mészáros <[matyas.meszáros@ec.europa.eu](mailto:matyas.meszáros@ec.europa.eu)>

**Repository** CRAN

**Date/Publication** 2020-09-18 07:20:02 UTC

## R topics documented:

clean_restatapi_cache . . . . .	2
create_filter_table . . . . .	3
extract_data . . . . .	5
extract_dsd . . . . .	6
extract_toc . . . . .	6
filter_raw_data . . . . .	7
get_compressed_sdmx . . . . .	8
get_eurostat_bulk . . . . .	9
get_eurostat_cache . . . . .	11
get_eurostat_data . . . . .	12
get_eurostat_dsd . . . . .	16
get_eurostat_raw . . . . .	18
get_eurostat_toc . . . . .	20
load_cfg . . . . .	22
put_eurostat_cache . . . . .	23
search_eurostat_dsd . . . . .	25
search_eurostat_toc . . . . .	26

<b>Index</b>	<b>28</b>
--------------	-----------

---

clean\_restatapi\_cache *Clean restatapi cache*

---

### Description

Remove all objects from the `.restatapi_env` except the configuration file, API version number, download method and the country codes. In addition, it deletes all the `.rds` files from the default and selected cache directory. See [get\\_eurostat\\_data](#) for more on cache.

### Usage

```
clean_restatapi_cache(cache_dir = NULL, verbose = FALSE)
```

### Arguments

`cache_dir` A path to cache directory. If `NULL` (default) tries to clean default temporary cache directory. Directory can also be set with `options(restatapi_cache_dir=...)`.

`verbose` A boolean with default `FALSE`, so detailed messages (for debugging) will not be printed. Can be set also with `options(restatapi_verbose=TRUE)`

### Examples

```
clean_restatapi_cache(verbose=TRUE)
```

---

create\_filter\_table    *Create filter table*

---

### Description

Create filter table from the filters and date\_filter strings parameters of the [get\\_eurostat\\_data](#) to be used in the [filter\\_raw\\_data](#) function for filtering

### Usage

```
create_filter_table(
  filters,
  date_filter = FALSE,
  dsd = NULL,
  exact_match = TRUE,
  verbose = FALSE,
  ...
)
```

### Arguments

filters	string, a character or numeric vector or a named list containing words to filter by the different concepts, geographical location or time. The words can be any word, Eurostat variable code, and value which are in the Data Structure Definition (DSD) and can be retrieved by the <a href="#">search_eurostat_dsd</a> function. If a named list is used, then the name of the list elements should be the concepts from the DSD and the provided values will be used to filter the dataset for the given concept. The default is NULL, in this case the whole dataset is returned via the bulk download. To filter by time see date_filter below. In case for filtering for time and the date is defined as character string, then it should follow the format yyyy[-mm][-dd], where the month and the day part is optional.
date_filter	if TRUE the filter table is generated only for the time dimension. The default is FALSE, in this case a (dsd) should be provided which will be searched for the values given in the filters.
dsd	a table containing a DSD of an Eurostat dataset which can be retrieved by the <a href="#">get_eurostat_dsd</a> function.
exact_match	a boolean with the default value TRUE, if the strings provided in filters shall be matched exactly as it is or as a pattern in the DSD.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not be printed. Can be set also with options(restatapi_verbose=TRUE)
...	further arguments to the <a href="#">search_eurostat_dsd</a> function, e.g.: ignore.case or name. The ignore.case has the default value FALSE, then the strings provided in filters are matched as is, otherwise the case of the letters is ignored. If the name=FALSE then the pattern(s) provided in the filters argument is only searched in the code column of the DSD, and the names of the codes will not be searched.

## Details

It is a sub-function to use in the [get\\_eurostat\\_data](#) to generate url for the given filters and `date_filter` in that function. The output can be used also for filtering data on the local computer with the [get\\_eurostat\\_raw](#) and [filter\\_raw\\_data](#) function, if the direct response from REST API did not provide data because of to large data set.

## Value

a `data.table` containing in each row a distinct filtering condition to be applied to a raw Eurostat datatable. In case if `date_filter=TRUE` the output data table contains two columns with the following names:

- sd Starting date to be included in the filtered dataset, where date is formatted yyyy[-mm][-dd]
- ed End date of the period to be included in the filtered dataset, where the date is formatted yyyy[-mm][-dd]

In case if `date_filter=FALSE` the output tables have the following four columns:

- pattern Containing those parts of the filters string where the string part (pattern) was found in the dsd
- concept The name of the concepts corresponding to the result in the code/name column where the pattern was found in the
- code The list of codes where the pattern was found, or the code of a name/description where the pattern appears
- name The name/description which can be used as label for the code where the pattern was found, or the name/description

## See Also

[get\\_eurostat\\_raw](#), [search\\_eurostat\\_dsd](#), [get\\_eurostat\\_data](#), [filter\\_raw\\_data](#)

## Examples

```
dsd<-get_eurostat_dsd("avia_par_me")
create_filter_table(c("KYIV", "hu", "Quarterly"), dsd=dsd, exact_match=FALSE, ignore.case=TRUE)
create_filter_table(c("KYIV", "LHBP", "Monthly"), dsd=dsd, exact_match=FALSE, name=FALSE)
create_filter_table(c("2017-03",
                      "2001-03:2005",
                      "<2000-07-01",
                      "2012:2014",
                      "2018<",
                      "20912",
                      "<3452<",
                      ":2018-04>",
                      "2<034v",
                      "2008:2013"),
                    date_filter=TRUE,
                    verbose=TRUE)
```

---

extract_data	<i>Extract data values from SDMX XML</i>
--------------	--

---

**Description**

Extracts the data values from the SDMX XML data file

**Usage**

```
extract_data(xml_lf, keep_flags = FALSE, stringsAsFactors = FALSE, bulk = TRUE)
```

**Arguments**

xml_lf	an input XML leaf with data series from an SDMX XML file to extract the value and dimension from it
keep_flags	a boolean if to extract the observation status (flag) information from the XML file. The default value is FALSE
stringsAsFactors	if TRUE the columns are converted to factors. The default is FALSE, in this case the strings are returned as a character.
bulk	a boolean with default value TRUE if the input SDMX XML file is from the bulk download facility containing all the observations. If the input file has pre-filtered values then the FALSE should be used.

**Details**

It is a sub-function to use in the [get\\_eurostat\\_data](#) and the [get\\_eurostat\\_raw](#) functions.

**Value**

a data frame containing the values of an SDMX node, (the dimensions, value and optional the flag(s))

**Examples**

```
id<-"agr_r_milkpr"
toc<-get_eurostat_toc()
bulk_url<-toc$downloadLink.sdmx[toc$code==id]
temp<-tempfile()
download.file(bulk_url,temp)
sdmx_xml<-xml2::read_xml(unzip(temp, paste0(id,".sdmx.xml")))
xml_leafs<-xml2::xml_find_all(sdmx_xml,".//data:Series")
extract_data(xml_leafs[1])
```

---

extract_dsd	<i>Extract values from XML</i>
-------------	--------------------------------

---

**Description**

Extracts the values from the XML Data Structure Definition (DSD) file

**Usage**

```
extract_dsd(concept = NULL, dsd_xml = NULL)
```

**Arguments**

concept	a character vector with concept ids
dsd_xml	an XML file with DSD content

**Details**

It is a sub-function to use in the [get\\_eurostat\\_dsd](#) function.

**Value**

a matrix with 3 columns if the concepts has code list in the DSD file

**Examples**

```
dsd_url<-"http://ec.europa.eu/eurostat/SDMX/diss-web/rest/datastructure/ESTAT/DSD_nama_10_a10_e"
tryCatch({
  dsd_xml<-xml2::read_xml(dsd_url)},
  error=function(e){
    message("Unable to download the xml file.\n",e)},
  warning=function(w){
    message("Unable to download the xml file.\n",w)})
if (exists("dsd_xml")) {extract_dsd("GEO",dsd_xml)}
```

---

extract_toc	<i>Extract values from XML</i>
-------------	--------------------------------

---

**Description**

Extracts the values of a node from the Eurostat XML Table of contents (TOC) file

**Usage**

```
extract_toc(ns)
```

**Arguments**

ns                    An XML node set from the XML TOC file

**Details**

It is a sub-function to use in the [get\\_eurostat\\_toc](#) function.

**Value**

a character vector with all the values of the node set.

**Examples**

```
cfg<-get("cfg",envir=.restatapi_env)
rav<-get("rav",envir=.restatapi_env)
toc_endpoint<-eval(parse(text=paste0("cfg$TOC_ENDPOINT$",rav,"'$ESTAT$xml'")))

xml_leafs<-xml2::xml_find_all(xml2::read_xml(toc_endpoint),".//nt:leaf")
extract_toc(xml_leafs[1])
```

---

filter_raw_data	<i>Filter raw data locally</i>
-----------------	--------------------------------

---

**Description**

Filter downloaded full raw dataset on local computer if the [get\\_eurostat\\_data](#) has not provide data due to too large datasets for the REST API.

**Usage**

```
filter_raw_data(raw_data, filter_table, date_filter = FALSE)
```

**Arguments**

raw\_data            an input data.table dataset resulted from the call of the [get\\_eurostat\\_raw](#) function

filter\_table        a data table with values for the concepts or time to be filtered out

date\_filter         if TRUE the filter table should be applied to the time columns of the raw\_data. The default is FALSE, in this case the filters applied to the other columns of raw\_data.

**Details**

It is a sub-function to use in the [get\\_eurostat\\_data](#) to filter data if the direct response from REST API did not provide data because of to large data set. The filter\_table contains always at least two columns. In case if date\_filter=TRUE then the two columns should have the following names:

sd Starting date to be included, where date is formatted yyyy[-mm][-dd]  
 ed End date of the period to be included in the dataset formatted yyyy[-mm][-dd]

In this case these conditions are applied to the time column of the the raw\_data data.table. In case if date\_filter=FALSE then the columns should have the following names:

concept Containing concept names, which is a column name in the raw\_data data.table  
 code A possible code under the given concept, which is a value in the column of the raw\_data data.table defined by the

### Value

a filtered data.table containing only the rows of raw\_data which fulfills the conditions in the filter\_table

### See Also

[get\\_eurostat\\_raw](#), [search\\_eurostat\\_dsd](#), [get\\_eurostat\\_data](#), [create\\_filter\\_table](#)

### Examples

```
id<-"tus_00age"
rd<-get_eurostat_raw(id)
dsd<-get_eurostat_dsd(id)
ft<-create_filter_table(c("TIME_SP", "Hungary", 'T'), FALSE, dsd)
filter_raw_data(rd, ft)
```

---

get\_compressed\_sdmx *Download and extract compressed SDMX XML*

---

### Description

Downloads and extracts the data values from the SDMX XML data file

### Usage

```
get_compressed_sdmx(url = NULL, verbose = FALSE)
```

### Arguments

url an XML leaf with data series from an SDMX XML file  
 verbose A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi\_verbose=TRUE)



**Details**

It is a sub-function to use in the [get\\_eurostat\\_raw](#) and the [get\\_eurostat\\_data](#) functions.

**Value**

an xml file with SDMX tags

**Examples**

```
base_url<-"https://ec.europa.eu/eurostat/"
url_end<-"estat-navtree-portlet-prod/BulkDownloadListing?file=data/agr_r_milkpr.sdmx.zip"
url<-paste0(base_url,url_end)
sdmx_xml<-get_compressed_sdmx(url,verbose=TRUE)
```

---

<code>get_eurostat_bulk</code>	<i>Get Eurostat data in a standardized format</i>
--------------------------------	---

---

**Description**

Download data sets from **Eurostat** database and put in a standardized format.

**Usage**

```
get_eurostat_bulk(
  id,
  cache = TRUE,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE,
  stringsAsFactors = TRUE,
  select_freq = NULL,
  keep_flags = FALSE,
  cflags = FALSE,
  check_toc = FALSE,
  verbose = FALSE,
  ...
)
```

**Arguments**

<code>id</code>	A code name for the dataset of interest. See <a href="#">search_eurostat_toc</a> for details how to get an id.
<code>cache</code>	a logical whether to do caching. Default is TRUE.
<code>update_cache</code>	a logical with a default value FALSE, whether to update cache. Can be set also with <code>options(restatapi_update=TRUE)</code>

cache_dir	a path to a cache directory. The NULL (default) uses the memory as cache. If the folder if the cache_dir directory does not exist it saves in the 'restatapi' directory under the temporary directory from tempdir(). Directory can also be set with option(restatapi_cache_dir=...).
compress_file	a logical whether to compress the RDS-file in caching. Default is TRUE.
stringsAsFactors	the default is TRUE, in this case the columns are converted to factors. If FALSE, the strings are returned as character.
select_freq	a character symbol for a time frequency when a dataset has multiple time frequencies. Possible values are: A = annual, S = semi-annual, H = half-year, Q = quarterly, M = monthly, W = weekly, D = daily. The default is NULL as most datasets have just one time frequency. In case if there are multiple frequencies and select_freq=NULL, then only the most common frequency kept. If all the frequencies needed the <a href="#">get_eurostat_raw</a> can be used.
keep_flags	a logical whether the observation status (flags) - e.g. "confidential", "provisional", etc. - should be kept in a separate column or if they can be removed. Default is FALSE. For flag values see: <a href="http://ec.europa.eu/eurostat/data/database/information">http://ec.europa.eu/eurostat/data/database/information</a> .
cflags	a logical whether the missing observations with flag 'c' - "confidential" should be kept or not. Default is FALSE, in this case these observations dropped from the dataset. If this parameter TRUE then the flags are kept and the parameter provided in keep_flags is not taken into account.
check_toc	a boolean whether to check the provided id in the Table of Contents (TOC) or not. The default value FALSE, in this case the base URL for the download link is retrieved from the configuration file. If the value is TRUE then the TOC is downloaded and the id is checked in it. If it found then the download link is retrieved form the TOC.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi_verbose=TRUE)
...	other parameter(s) to pass on the <a href="#">load_cfg</a> function

## Details

Data sets are downloaded from [the Eurostat bulk download facility](#) in TSV format as in this case smaller file has to be downloaded and processed. If there is more then one frequency then the is filtered for a unique time frequency. If no frequency is selected and there are multiple frequencies in the dataset, then the most common value is used used for frequency.

Compared to [get\\_eurostat\\_raw](#) the frequency (FREQ) and time format (TIME\_FORMAT) columns are not included and the column names for the time period, observation values and status have standardised names: "time", "values" and "flags" independently if the data was downloaded previously through SDMX or TSV format.

By default all datasets cached as they are often rather large. The datasets cached in memory (default) or can be stored in a temporary directory if cache\_dir or option(restatapi\_cache\_dir) is defined. The cache can be emptied with [clean\\_restatapi\\_cache](#).

The id, is a value from the code column of the table of contents ([get\\_eurostat\\_toc](#)), and can be searched for with the [search\\_eurostat\\_toc](#) function. The id value can be retrieved from the

**Eurostat database** as well. The Eurostat database gives codes in the Data Navigation Tree after every dataset in parenthesis.

### Value

a data.table with the following columns: #'

dimension names	One column for each dimension in the data
time	A column for the time dimension
values	A column for numerical values
flags	A column for flags if the keep_flags=TRUE otherwise this column is not included in the data table

The data.table does not include all missing values. The missing values are dropped if the value and flag are missing on a particular time.

### See Also

[get\\_eurostat\\_data](#), [get\\_eurostat\\_raw](#)

### Examples

```
dt<-get_eurostat_bulk("agr_r_milkpr", keep_flags=TRUE)
options(restatapi_update=TRUE)
dt<-get_eurostat_bulk("avia_par_ee", check_toc=TRUE)
dt<-get_eurostat_bulk("avia_par_ee", select_freq="A", verbose=TRUE)
options(restatapi_update=FALSE)
dt<-get_eurostat_bulk("agr_r_milkpr", cache_dir=tempdir(), compress_file=FALSE, verbose=TRUE)
```

---

get\_eurostat\_cache      *Load an object from cache*

---

### Description

Search and load the object (dataset/toc/DSD) from cache

### Usage

```
get_eurostat_cache(oname, cache_dir = NULL, verbose = FALSE)
```

### Arguments

oname	a character string with the name of the object (toc, dataset id, DSD id)
cache_dir	a path to a cache directory to search in. The default is NULL, in this case the object is searched in the memory (in the '.restatapi_env'). Otherwise if the cache_dir directory does not exist it searches the 'restatapi' directory in the temporary directory from tempdir(). Directory can also be set with options(restatapi_cache_dir=.

verbose      A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi\_verbose=TRUE)

### Details

If the given name or the beginning of the name (for datasets) found in the cache then it returns the value of the object otherwise it returns NULL.

### Value

The requested object if exists in the '.restatapi\_env' or in the cache\_dir, otherwise it returns the NULL value.

### Examples

```
dt<-data.frame(txt=c("a","b","c"),nr=c(1,2,3))
put_eurostat_cache(dt,"teszt")
get_eurostat_cache("teszt",verbose=TRUE)
```

---

get\_eurostat\_data      *Download/extract Eurostat Data*

---

### Description

Download full or partial data set from **Eurostat** database.

### Usage

```
get_eurostat_data(
  id,
  filters = NULL,
  exact_match = TRUE,
  date_filter = NULL,
  label = FALSE,
  select_freq = NULL,
  cache = TRUE,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE,
  stringsAsFactors = TRUE,
  keep_flags = FALSE,
  cflags = FALSE,
  check_toc = FALSE,
  local_filter = TRUE,
  force_local_filter = FALSE,
```

```

    verbose = FALSE,
    ...
)

```

### Arguments

id	A code name for the dataset of interest. See <a href="#">search_eurostat_toc</a> for details how to get an id.
filters	a string, a character vector or named list containing words to filter by the different concepts or geographical location. If filter applied only part of the dataset is downloaded through the API. The words can be any word, Eurostat variable code, and value which are in the DSD <a href="#">search_eurostat_dsd</a> . If a named list is used, then the name of the list elements should be the concepts from the DSD and the provided values will be used to filter the dataset for the given concept. The default is NULL, in this case the whole dataset is returned via the bulk download. To filter by time see <code>date_filter</code> below. If after filtering still the dataset has more observations than the limit per query via the API, then the raw download is used to retrieve the whole dataset and apply the filter on the local computer. This option can be disabled with the <code>local_filter=FALSE</code> parameter.
exact_match	a boolean with the default value TRUE, if the strings provided in <code>filters</code> shall be matched exactly as it is or as a pattern.
date_filter	a vector which can be numeric or character containing dates to filter the dataset. If date is defined as character string it should follow the format <code>yyyy[-mm][-dd]</code> , where the month and the day part is optional. If date filter applied only part of the dataset is downloaded through the API. The default is NULL, in this case the whole dataset is returned via the bulk download. If after filtering still the dataset has more observations than the limit per query via the API, then the raw download is used to retrieve the data and apply the filter on the local computer. This option can be disabled with the <code>local_filter=FALSE</code> parameter.
label	a boolean with the default FALSE. If it is TRUE then the code values are replaced by the name from the Data Structure Definition (DSD) <a href="#">get_eurostat_dsd</a> . For example instead of "D1110A", "Raw cows' milk from farmtype" is used or "HU32" is replaced by "Észak-Alföld".
select_freq	a character symbol for a time frequency when a dataset has multiple time frequencies. Possible values are: A = annual, S = semi-annual, H = half-year, Q = quarterly, M = monthly, W = weekly, D = daily. The default is NULL as most datasets have just one time frequency and in case there are multiple frequencies, then only the most common frequency kept. If all the frequencies needed the <a href="#">get_eurostat_raw</a> can be used.
cache	a logical whether to do caching. Default is TRUE. Affects only queries without filtering. If <code>filters</code> or <code>date_filter</code> is used then there is no caching.
update_cache	a logical with a default value FALSE, whether to update the data in the cache. Can be set also with <code>options(restatapi_update=TRUE)</code>
cache_dir	a path to a cache directory. The NULL (default) uses the memory as cache. If the folder <code>cache_dir</code> directory does not exist it saves in the 'restatapi' directory

	under the temporary directory from <code>tempdir()</code> . Directory can also be set with <code>option(restatapi_cache_dir=...)</code> .
<code>compress_file</code>	a logical whether to compress the RDS-file in caching. Default is TRUE.
<code>stringsAsFactors</code>	if TRUE (the default) the non-numeric columns are converted to factors. If the value FALSE they are returned as characters.
<code>keep_flags</code>	a logical whether the observation status (flags) - e.g. "confidential", "provisional", etc. - should be kept in a separate column or if they can be removed. Default is FALSE. For flag values see: <a href="http://ec.europa.eu/eurostat/data/database/information">http://ec.europa.eu/eurostat/data/database/information</a> .
<code>cflags</code>	a logical whether the missing observations with flag 'c' - "confidential" should be kept or not. Default is FALSE, in this case these observations dropped from the dataset. If this parameter TRUE then the flags are kept and the parameter provided in <code>keep_flags</code> is not taken into account.
<code>check_toc</code>	a boolean whether to check the provided id in the Table of Contents (TOC) or not. The default value FALSE, in this case the base URL for the download link is retrieved from the configuration file. If the value is TRUE then the TOC is downloaded and the id is checked in it. If it found then the download link is retrieved from the TOC.
<code>local_filter</code>	a boolean whether do the filtering on the local computer or not in case after filtering still the dataset has more observations than the limit per query via the API would allow to download. The default is TRUE, in this case if the response footer contains information that the result cannot be downloaded because it is too large, then the whole raw dataset is downloaded and filtered on the local computer.
<code>force_local_filter</code>	a boolean with the default value FALSE. In case, if there are existing filter conditions, then it will do the filtering on the local computer and not requesting through the REST API. It can be useful, if the values are not numeric as these are provided as NaN (Not a Number) through the REST API, but it is fully listed in the raw dataset.
<code>verbose</code>	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with <code>options(restatapi_verbose=TRUE)</code>
<code>...</code>	further arguments to the <code>search_eurostat_dsd</code> function, e.g.: <code>ignore.case</code> or <code>name</code> . The <code>ignore.case</code> has the default value FALSE, then the strings provided in <code>filters</code> are matched as is, otherwise the case of the letters is ignored. If the <code>name=FALSE</code> then the pattern(s) provided in the <code>filters</code> argument is only searched in the code column of the DSD, and the names of the codes will not be searched.

## Details

Data sets are downloaded from the Eurostat Web Services **SDMX API** if there is a filter otherwise the **the Eurostat bulk download facility** is used. If only the table id is given, the whole table is downloaded from the bulk download facility. If also `filters` or `date_filter` is defined then the SDMX REST API is used. In case after filtering the dataset has more rows than the limitation of

the SDMX REST API (1 million values at one time) then the bulk download is used to retrieve the whole dataset .

By default all datasets cached as they are often rather large. The datasets cached in memory (default) or can be stored in a temporary directory if `cache_dir` or `option(restatpi_cache_dir)` is defined. The cache can be emptied with `clean_restatapi_cache`.

The `id`, is a value from the code column of the table of contents (`get_eurostat_toc`), and can be searched for with the `search_eurostat_toc` function. The `id` value can be retrieved from the [Eurostat database](#) as well. The Eurostat database gives codes in the Data Navigation Tree after every dataset in parenthesis.

Filtering can be done by the codes as described in the API documentation providing in the correct order and connecting with "." and "+". If we do not know the codes we can filter based on words or by the mix of the two putting in a vector like `c("AT$", "Belgium", "persons", "Total")`. Be careful that the filter is case sensitive, if you do not know the code or label exactly you can use the option `ignore.case=TRUE` and `exact_match=FALSE`, but in this case the results may include unwanted elements as well. In the `filters` parameter regular expressions can be used as well. We do not have to worry about the correct order of the filter, it will be put in the correct place based on the DSD.

The `date_filter` shall be a string in the format `yyyy[-mm][-dd]`. The month and the day part is optional, but if we use the years and we have monthly frequency then all the data for the given year is retrieved. The string can be extended by adding the "<" or ">" to the beginning or to the end of the string. In this case the date filter is treated as range, and the date is used as a starting or end date. The data will include the observation of the start/end date. A single date range can be defined as well by concatenating two dates with the ":", e.g. `"2016-08:2017-03-15"`. As seen in the example the dates can have different length: one defined only at year/month level, the other by day level. If a date range is defined with ":", it is not possible to use the "<" or ">" characters in the date filter. If there are multiple dates which is not a continuous range, it can be put in vector in any order like `c("2016-08", 2013:2015, "2017-07-01")`. In this case, as well, it is not possible to use the "<" or ">" characters.

## Value

a `data.table`. a `data.table` with the following columns: #

<code>freq</code>	A column for the frequency of the data in case there are multiple frequencies, for single frequency this column is not included
<code>dimension names</code>	One column for each dimension in the data
<code>time</code>	A column for the time dimension
<code>values</code>	A column for numerical values
<code>flags</code>	A column for flags if the <code>keep_flags=TRUE</code> or <code>cf_flags=TRUE</code> otherwise this column is not included in the <code>data.table</code>

The `data.table` does not include all missing values. The missing values are dropped if the value and flag are missing on a particular time.

## See Also

[search\\_eurostat\\_toc](#), [search\\_eurostat\\_dsd](#)

**Examples**

```

load_cfg()
eu<-get("cc",envir=.restatapi_env)

dt<-get_eurostat_data("NAMA_10_GDP")
dt<-get_eurostat_data("htec_cis3",update_cache=TRUE,check_toc=TRUE)
dt<-get_eurostat_data("agr_r_milkpr",cache_dir="/tmp",cflags=TRUE)
options(restatapi_update=FALSE)
options(restatapi_cache_dir=file.path(tempdir(),"restatapi"))
dt<-get_eurostat_data("avia_gonc",select_freq="A",cache=FALSE)
dt<-get_eurostat_data("agr_r_milkpr",date_filter=2008,keep_flags=TRUE)
dt<-get_eurostat_data("avia_par_me",
  filters="BE$",
  exact_match=FALSE,
  date_filter=c(2016,"2017-03","2017-07-01"),
  select_freq="Q",
  label=TRUE,
  name=FALSE)
dt<-get_eurostat_data("agr_r_milkpr",
  filters=c("BE$","Hungary"),
  date_filter="2007-06<",
  keep_flags=TRUE)
dt<-get_eurostat_data("nama_10_a10_e",
  filters=c("Annual","EU28","Belgium","AT","Total","EMP_DC","person"),
  date_filter=c("2008",2002,2013:2018))
dt<-get_eurostat_data("vit_t3",
  filters=c("EU28",eu$EA15,"HU$"),
  date_filter=c("2015",2007))
dt<-get_eurostat_data("avia_par_me",
  filters="Q...ME_LYPG_HU_LHBP+ME_LYTV_UA_UKKK",
  date_filter=c("2016-08","2017-07-01"),
  select_freq="M")
dt<-get_eurostat_data("htec_cis3",
  filters="lu",
  ignore.case=TRUE)
dt<-get_eurostat_data("bop_its6_det",
  filters=list(bop_item="SC",
    currency="MIO_EUR",
    partner="EXT_EU28",
    geo=c("EU28","HU"),
    stk_flow="BAL"),
  date_filter="2010:2012",
  select_freq="A",
  label=TRUE,
  name=FALSE)

```



**Description**

Download Data Structure Definition (DSD) of a Eurostat dataset if it is not cached previously.

**Usage**

```
get_eurostat_dsd(
  id,
  cache = TRUE,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE,
  verbose = FALSE,
  ...
)
```

**Arguments**

id	a character string with the id of the dataset. It is the value from the codename column of the <a href="#">get_eurostat_toc</a> function.
cache	a boolean whether to load/save the TOC from/in the cache or not. The default value is TRUE, so that the TOC is checked first in the cache and if does not exist then downloaded from Eurostat and cached.
update_cache	a boolean to update cache or not. The default value is FALSE, so the cache is not updated. Can be set also with <code>options(restatapi_update=TRUE)</code>
cache_dir	a path to a cache directory. The default is NULL, in this case the TOC is cached in the memory (in the <code>restatapi_env</code> ). Otherwise if the <code>cache_dir</code> directory does not exist it creates the <code>restatapi</code> directory in the temporary directory from <code>tempdir()</code> to save the RDS-file. Directory can also be set with <code>option(restatapi_cache_dir=...)</code> .
compress_file	a logical whether to compress the RDS-file in caching. Default is TRUE.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with <code>options(restatapi_verbose=TRUE)</code>
...	parameter to pass on the <code>load_cfg</code> function

**Details**

The DSD is downloaded from Eurostat's website, through the REST API in XML (SDMX) format.

**Value**

If the DSD does not exist it returns NULL otherwise the result is a table with the 3 columns:

concept	The name of the concepts in the order of the data structure
code	The possible list of codes under the concept
name	The name/description of the code

## References

For more information see the detailed documentation of the [API](#).

## See Also

[get\\_eurostat\\_data](#), [search\\_eurostat\\_toc](#).

## Examples

```
dsd<-get_eurostat_dsd("nama_10_gdp",cache=FALSE,verbose=TRUE)
head(dsd)
```

---

get_eurostat_raw	<i>Get Eurostat data as it is</i>
------------------	-----------------------------------

---

## Description

Download data sets from [Eurostat](#) database .

## Usage

```
get_eurostat_raw(
  id,
  mode = "txt",
  cache = TRUE,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE,
  stringsAsFactors = FALSE,
  keep_flags = FALSE,
  check_toc = FALSE,
  verbose = FALSE,
  ...
)
```

## Arguments

id	A code name for the dataset of interest. See <a href="#">search_eurostat_toc</a> for details how to get an id.
mode	defines the format of the downloaded dataset. It can be txt (the default value) for Tab Separated Values (TSV), or xml for the SDMX version.
cache	a logical whether to do caching. Default is TRUE.
update_cache	a logical with a default value FALSE, whether to update cache. Can be set also with <code>options(restatapi_update=TRUE)</code>

cache_dir	a path to a cache directory. The NULL (default) uses the memory as cache. If the folder if the cache_dir directory does not exist it saves in the 'restatapi' directory under the temporary directory from tempdir(). Directory can also be set with option(restatapi_cache_dir=...).
compress_file	a logical whether to compress the RDS-file in caching. Default is TRUE.
stringsAsFactors	if TRUE the variables which are not numeric are converted to factors. The default value FALSE, in this case they are returned as characters.
keep_flags	a logical whether the observation status (flags) - e.g. "confidential", "provisional", etc. - should be kept in a separate column or if they can be removed. Default is FALSE. For flag values see: <a href="http://ec.europa.eu/eurostat/data/database/information">http://ec.europa.eu/eurostat/data/database/information</a> .
check_toc	a boolean whether to check the provided id in the Table of Contents (TOC) or not. The default value FALSE, in this case the base URL for the download link is retrieved from the configuration file. If the value is TRUE then the TOC is downloaded and the id is checked in it. If it found then the download link is retrieved from the TOC.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi_verbose=TRUE)
...	further argument for the <code>load_cfg</code> function

## Details

Data sets are downloaded from [the Eurostat bulk download facility](#) in TSV or SDMX format.

The id, should be a value from the code column of the table of contents (`get_eurostat_toc`), and can be searched for with the `search_eurostat_toc` function. The id value can be retrieved from the [Eurostat database](#) as well. The Eurostat database gives codes in the Data Navigation Tree after every dataset in parenthesis. By default all datasets downloaded in TSV format and cached as they are often rather large. The datasets cached in memory (default) or can be stored in a temporary directory if `cache_dir` or `option(restatapi_cache_dir)` is defined. The cache can be emptied with `clean_restatapi_cache`. If the id is checked in TOC then the data will saved in the cache with the date from the "lastUpdate" column from the TOC, otherwise it is saved with the current date.

## Value

a data.table with the following columns:

FREQ	The frequency of the data (Annual, Semi-annual, Half-year, Quarterly, Monthly, Weekly, Daily)
dimension names	One column for each dimension in the data
TIME_FORMAT	A column for the time format, if the source file SDMX and the data was not loaded from a previously c
time/TIME_PERIOD	A column for the time dimension, where the name of the column depends on the source file (TSV/SDMX
values/OBS_VALUE	A column for numerical values, where the name of the column depends on the source file (TSV/SDMX
flags/OBS_STATUS	A column for flags if the <code>keep_flags=TRUE</code> otherwise this column is not included in the data table, and

The data does not include all missing values. The missing values are dropped if the value and flags are missing on a particular time.

**See Also**

[get\\_eurostat\\_data](#), [get\\_eurostat\\_bulk](#)

**Examples**

```
dt<-get_eurostat_raw("agr_r_milkpr",keep_flags=TRUE)
dt<-get_eurostat_raw("avia_par_ee",mode="xml",check_toc=TRUE,update_cache=TRUE)
options(restatapi_update=FALSE)
dt<-get_eurostat_raw("avia_par_me",mode="txt",cache_dir=tempdir(),compress_file=FALSE,verbose=TRUE)
```

---

<code>get_eurostat_toc</code>	<i>Download the Table of Contents of Eurostat datasets</i>
-------------------------------	--

---

**Description**

Download Table of Contents (TOC) of Eurostat datasets if it is not cached previously.

**Usage**

```
get_eurostat_toc(
  mode = "xml",
  cache = TRUE,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE,
  lang = "en",
  verbose = FALSE,
  ...
)
```

**Arguments**

<code>mode</code>	a character string either <code>xml</code> or <code>txt</code> defining the download mode. Depending on the mode the 'xml' version or the 'text' version of the TOC is downloaded. The default value is <code>xml</code> as it provides more information (e.g. number of values, short description and download links in different formats (SDMX, TSV))
<code>cache</code>	a boolean whether to load/save the TOC from/in the cache or not. The default value is <code>TRUE</code> , so that the TOC is checked first in the cache and if does not exist then downloaded from Eurostat and cached.
<code>update_cache</code>	a boolean to update cache or not. The default value is <code>FALSE</code> , so the cache is not updated. Can be set also with <code>options(restatapi_update=TRUE)</code>

cache_dir	a path to a cache directory. The default is NULL, in this case the TOC is cached in the memory (in the '.restatapi_env'). Otherwise if the cache_dir directory does not exist it creates the 'restatapi' directory in the temporary directory from tempdir() to save the RDS- file. Directory can also be set with option(restatapi_cache_dir=...).
compress_file	a logical whether to compress the RDS-file in caching. Default is TRUE.
lang	a character string either en, de or fr to define the language version for the table of contents. The default is en - English.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi_verbose=TRUE)
...	parameter to pass on the load_cfg function

### Details

The TOC is downloaded from Eurostat websites through the REST API for the xml (default) version or from the bulk download facilities for txt version. From the downloaded TOC the values in the 'code' column can be used as id in the [get\\_eurostat\\_dsd](#), [get\\_eurostat\\_raw](#), [get\\_eurostat\\_bulk](#), and [get\\_eurostat\\_data](#) functions.

### Value

A data table with the following columns:

title	The name of dataset/table in the language provided by the lang parameter
code	The codename of dataset/table which can be used as id in other functions
type	The type of information: 'dataset' or 'table'
lastUpdate	The date when the data was last time updated for tables and datasets
lastModified	The date when the structure of the dataset/table was last time modified
dataStart	The start date of the data in the dataset/table
dataEnd	The end date of the data in the dataset/table
values	The number of values in the dataset/table, and it is filled only if the download mode is "xml"
unit	The unit name for tables in the language provided by the lang parameter, for dataset it is empty and th
shortDescription	The short description of the values for tables in the language provided by the lang parameter, for data
metadata.html	The link to the metadata in html format, and this column exists only if the download mode is "xml"
metadata.sdmx	The link to the metadata in SDMX format, and this column exists only if the download mode is "xml"
downloadLink.tsv	The link to the whole dataset/table in tab separated values format in the bulk download facility and thi
downloadLink.sdmx	The link to the whole dataset/table in SDMX format in the bulk download facility and this column exi

### References

For more technical information see the detailed documentation of the [API](#).

### See Also

[search\\_eurostat\\_toc](#), [get\\_eurostat\\_dsd](#), [get\\_eurostat\\_raw](#), [get\\_eurostat\\_bulk](#), [get\\_eurostat\\_data](#).

### Examples

```

toc_xml<-get_eurostat_toc(cache=FALSE,verbose=TRUE)
head(toc_xml)
toc_txt<-get_eurostat_toc(mode="txt", lang="de")
head(toc_txt)

```

---

load\_cfg

*Load configuration data from JSON*


---

### Description

Load the configuration information to the '.restatapi\_env' from the JSON configuration file.

### Usage

```

load_cfg(
  api_version = "current",
  load_toc = FALSE,
  parallel = TRUE,
  max_cores = FALSE,
  verbose = FALSE
)

```

### Arguments

api_version	It can be either "old", "new", "test" or "current". The default value is "current".
load_toc	The default value FALSE, which means that the XML version of the Table of contents (TOC) will not be downloaded and cached automatically in the '.restatapi_env' when the package is loaded.
parallel	A boolean with the default value TRUE. If there are multiple cores/logical processors then part of the data extraction is made in parallel reducing significantly the time needed for large datasets. If the value is FALSE the option restatapi_cores set to 1.
max_cores	A boolean with the default value FALSE. If the parameter 'parallel' is TRUE then this parameter is taken into account otherwise it is ignored. If the value is TRUE, then the maximum minus one cores/logical processors are used for parallel computing. If the parameter FALSE, then the default value of getOption("mc.cores") is used, if it is defined. If mc.cores is NULL then depending on the memory size and number of available cores/threads the restatapi_cores are set to 2 or 4 cores/logical processors. Otherwise the parallel processing turned off by setting the option restatapi_cores to 1. The number of cores used for parallel computing can be changed any time with options(restatapi_cores=...)
verbose	A boolean if the verbose message about the configuration to be showed or not. The default is FALSE. Can be set also with options(restatapi_verbose=TRUE)

## Details

Loads configuration data from a JSON file. The function first tries to load the configuration file from GitHub. If it is not possible it loads from the file delivered with the package. By this way different version of the API can be tested. Since in many cases there is http/https redirection in the download which can cause problems with the 'wininet' download method, the 'libcurl' method is used when it is available. This configuration code sets up the parallel processing to handle large XML files efficiently. By default if there is more then 4 cores/logical processors and at least 32 GB of RAM then 4 cores are used for parallel computing. If there is more then 2 cores then 2 cores are used. This default configuration can be overwritten with `options(restatapi_cores=...)` or with the `max_cores=TRUE` parameter. In the second case part of the computation distributed over the maximum number minus one cores. By using the `max_cores=TRUE` option there is a higher probability that the program will run out off memory for larger datasets. In addition, the list of country codes are loaded to the variable `cc` (country codes), based on the [Eurostat standard code list](#)

## Value

it returns 4 objects in the '.restatapi\_env'

- `cfg` a list with all the configuration data
- `rav` a character string with a number defining the `API_VERSION` from the configuration file to be used later. It is determined based on the `api_version` parameter.
- `cc` a list containing the 2 character country codes of the member states for different EU composition like EU15, EU28 or EA (Euro Area).
- `dmethod` the download method to be used to access Eurostat database. If the 'libcurl' method exists then it will be the method for file download otherwise it will be 'auto'.

## Examples

```
load_cfg(parallel=FALSE)
load_cfg(api_version="test", verbose=TRUE, max_cores=FALSE)
load_cfg()
eu<-get("cc", envir=.restatapi_env)
eu$EU28
eu$EA15
```

---

put\_eurostat\_cache      *Put an object to cache*

---

## Description

Save the object (dataset/toc/DSD) to cache

**Usage**

```
put_eurostat_cache(
  obj,
  oname,
  update_cache = FALSE,
  cache_dir = NULL,
  compress_file = TRUE
)
```

**Arguments**

obj	an object (toc, dataset, DSD)
oname	a character string with the name of the object to reference later in the cache
update_cache	a logical with a default value FALSE, whether to update the cache. In this case the existing value in the cache is overwritten. Can be set also with <code>options(restatapi_update = TRUE)</code>
cache_dir	a path to a cache directory. The default is NULL, in this case the object is saved in the memory (in the <code>restatapi_env</code> ). Otherwise if the <code>cache_dir</code> directory does not exist it saves in the <code>restatapi</code> directory under the temporary directory from <code>tempdir()</code> . Directory can also be set with <code>options(restatapi_cache_dir=...)</code> .
compress_file	a logical whether to compress the RDS-file in caching. Default is TRUE.

**Details**

Saves a given object in cache. This can be the memory `restatapi_env` or on the hards disk. If the given `cache_dir` does not exist then the file is saved in the R temp directory (`tempdir()`). If the file or object with the `oname` exists in the cache, then the object is not cached.

**Value**

The function returns the place where the object was cached: either it creates an the object in the memory (`restatapi_env`) or creates an RDS-file.

**Examples**

```
dt<-data.frame(txt=c("a","b","c"),nr=c(1,2,3))
put_eurostat_cache(dt,"teszt")
get("teszt",envir=.restatapi_env)
put_eurostat_cache(dt,"teszt",cache_dir=tempdir())
readRDS(file.path(tempdir(),"teszt.rds"))
```



---

search\_eurostat\_dsd    *Search the downloaded Data Structure Definition of a dataset*

---

### Description

Search the Data Structure Definition (DSD) of a Eurostat dataset for a given pattern. It returns the rows where the pattern appears in the code and name column of the output of the [get\\_eurostat\\_dsd](#) function.

### Usage

```
search_eurostat_dsd(pattern, dsd = NULL, name = TRUE, exact_match = FALSE, ...)
```

### Arguments

pattern	a character string or a vector of character string.
dsd	a table containing Data Structure Definition (DSD) of a Eurostat dataset which can be retrieved by the <a href="#">get_eurostat_dsd</a> function.
name	a boolean with the default value TRUE, if the search shall look for the pattern in the name of the code. If the value FALSE, then only the 'code' column of the DSD will be searched.
exact_match	a boolean with the default value FALSE, if the strings provided in pattern shall be matched exactly as it is or as a pattern.
...	additional arguments to the <code>grep</code> function like <code>ignore.case=TRUE</code> if the pattern should be searched case sensitive or not. The default value for <code>ignore.case</code> is FALSE.

### Details

The function returns the line(s) where the searched pattern appears in the code or in the name column.

### Value

If the pattern found then the function returns a data.frame with the 4 columns:

pattern	The pattern which was searched
concept	The name of the concepts in the data structure
code	The list of codes where the pattern was found, or the code of a name where the pattern appears
name	The name/description of the code where the pattern found, or the name of the code where the pattern appears

Otherwise returns the value NULL.

### See Also

[get\\_eurostat\\_dsd](#), [search\\_eurostat\\_toc](#).

## Examples

```

dsd_example<-get_eurostat_dsd("nama_10_gdp", verbose=TRUE)
search_eurostat_dsd("EU", dsd_example)
search_eurostat_dsd("EU", dsd_example, ignore.case=TRUE)
search_eurostat_dsd("EU27_2019", dsd_example, name=FALSE)
search_eurostat_dsd("EU27_2019", dsd_example, exact_match=TRUE)

```

---

search\_eurostat\_toc     *Search in Eurostat datasets titles, units and short description*

---

## Description

Lists names of dataset from Eurostat with the particular pattern in the title, units or short description.

## Usage

```
search_eurostat_toc(pattern, lang = "en", verbose = FALSE, ...)
```

## Arguments

pattern	Character string to search for in the table of contents of Eurostat tables/datasets
lang	a character string either en, de or fr to define the language version for the table of contents. The default is en - English.
verbose	A boolean with default FALSE, so detailed messages (for debugging) will not printed. Can be set also with options(restatapi_verbose=TRUE)
...	other additional parameters to pass to the grepl function like ignore.case=TRUE if the pattern should be searched case sensitive or not. The default value for ignore.case is FALSE.

## Details

Downloads the list of all tables and datasets available in the Eurostat database and returns all the details from the table of contents of the tables/datasets that contains particular pattern in the dataset title, unit or short description. E.g. all tables/datasets mentioning 'energy'.

## Value

A table with the following columns:

title	The name of dataset/table in the language provided by the lang parameter
code	The codename of dataset/table which can be used by the get_eurostat function
type	The type of information: 'dataset' or 'table'
lastUpdate	The date when the data was last time updated for tables and datasets
lastModified	The date when the structure of the dataset/table was last time modified
dataStart	The start date of the data in the dataset/table

dataEnd	The end date of the data in the dataset/table
values	The number of values in the dataset/table
unit	The unit name for tables in the language provided by the lang parameter, if the type 'dataset' this column is not present
shortDescription	The short description of the values for tables in the language provided by the lang parameter if the type is 'table'
metadata.html	The link to the metadata in html format
metadata.sdmx	The link to the metadata in SDMX format
downloadLink.tsv	The link to the whole dataset/table in tab separated values format in the bulk download facility
downloadLink.sdmx	The link to the whole dataset/table in SDMX format in the bulk download facility

The value in the code column can be used as an id in the [get\\_eurostat\\_data](#), [get\\_eurostat\\_bulk](#), [get\\_eurostat\\_raw](#) and [get\\_eurostat\\_dsd](#) functions. If there is no hit for the search query it returns NULL.

### See Also

[search\\_eurostat\\_dsd](#), [get\\_eurostat\\_data](#), [get\\_eurostat\\_toc](#)

### Examples

```
head(search_eurostat_toc("energy", verbose=TRUE))
nrow(search_eurostat_toc("energy"))
head(search_eurostat_toc("energie", lang="de", ignore.case=TRUE))
nrow(search_eurostat_toc("energie", lang="de", ignore.case=TRUE))
```

# Index

`clean_restatapi_cache`, [2](#), [10](#), [15](#), [19](#)  
`create_filter_table`, [3](#), [8](#)

`extract_data`, [5](#)  
`extract_dsd`, [6](#)  
`extract_toc`, [6](#)

`filter_raw_data`, [3](#), [4](#), [7](#)

`get_compressed_sdmx`, [8](#)  
`get_eurostat_bulk`, [9](#), [20](#), [21](#), [27](#)  
`get_eurostat_cache`, [11](#)  
`get_eurostat_data`, [2–5](#), [7–9](#), [11](#), [12](#), [18](#), [20](#),  
[21](#), [27](#)  
`get_eurostat_dsd`, [3](#), [6](#), [13](#), [16](#), [21](#), [25](#), [27](#)  
`get_eurostat_raw`, [4](#), [5](#), [7–11](#), [13](#), [18](#), [21](#), [27](#)  
`get_eurostat_toc`, [7](#), [10](#), [15](#), [17](#), [19](#), [20](#), [27](#)

`load_cfg`, [10](#), [19](#), [22](#)

`put_eurostat_cache`, [23](#)

`search_eurostat_dsd`, [3](#), [4](#), [8](#), [13–15](#), [25](#), [27](#)  
`search_eurostat_toc`, [9](#), [10](#), [13](#), [15](#), [18](#), [19](#),  
[21](#), [25](#), [26](#)