

# Package: capesR (via r-universe)

September 21, 2026

**Type** Package

**Title** Access to CAPES Data

**Version** 0.2.1

**Date** 2026-09-19

**Description** Provides simplified access to the data from the Catalog of Theses and Dissertations of the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES, <https://catalogodeteses.capes.gov.br>) for the years 1987 through 2024. The dataset includes variables such as Higher Education Institution (institution), Area of Concentration (area), Graduate Program Name (program\_name), Type of Work (type), Language of Work (language), Author Identification (author), Abstract (abstract), Advisor Identification (advisor), Development Region (region), State (state).

**License** GPL (>= 2)

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 4.0.0)

**Imports** arrow, dplyr, magrittr, rlang, stringr, utils

**Suggests** knitr, rmarkdown, testthat (>= 3.0.0)

**VignetteBuilder** knitr

**URL** <https://github.com/StrategicProjects/capesR>,  
<https://strategicprojects.github.io/capesR/>

**BugReports** <https://github.com/StrategicProjects/capesR/issues>

**Config/Needs/website** tidyverse/tidytemplate

**Config/roxygen2/version** 8.0.0

**Config/testthat/edition** 3

**Config/pak/sysreqs** cmake libicu-dev libssl-dev

**Repository** <https://milkway.r-universe.dev>

**Date/Publication** 2026-09-19 13:25:49 UTC  
**RemoteUrl** <https://github.com/milkway/capesr>  
**RemoteRef** HEAD  
**RemoteSha** 53f5c885d31766329c8d915a3f9dd1c784f8e1e3

Contents

|                               |          |
|-------------------------------|----------|
| capes_synthetic_df . . . . .  | 2        |
| capes_years . . . . .         | 3        |
| download_capes_data . . . . . | 4        |
| read_capes_data . . . . .     | 5        |
| search_capes_text . . . . .   | 5        |
| <b>Index</b>                  | <b>7</b> |

---

|                    |                             |
|--------------------|-----------------------------|
| capes_synthetic_df | <i>Synthetic CAPES Data</i> |
|--------------------|-----------------------------|

---

Description

Aggregated data from the CAPES Catalog of Theses and Dissertations, containing summarized information by year, institution, area, program, type, region, and state (UF).

Usage

capes\_synthetic\_df

Format

- A data frame with the following columns:
- base\_year** Reference year of the data.
  - institution** Higher Education Institution.
  - area** Area of Concentration.
  - program\_name** Name of the Graduate Program.
  - type** Type of work (e.g., Master’s, Doctorate).
  - region** Region of Brazil.
  - state** Federative Unit (state).
  - n** Total number of works.

Source

Synthetic data created from the CAPES Catalog of Theses and Dissertations.

## Examples

```
data(capes_synthetic_df)
head(capes_synthetic_df)
```

---

|             |                                                                      |
|-------------|----------------------------------------------------------------------|
| capes_years | <i>Annual files of the CAPES Catalog of Theses and Dissertations</i> |
|-------------|----------------------------------------------------------------------|

---

## Description

A data frame listing, for each year, the Parquet file distributed by the package and the URL from which it is downloaded. Used internally by [download\_capes\_data()]; the host can be overridden with the 'capesR.base\_url' option (see [download\_capes\_data()]).

## Usage

```
capes_years
```

## Format

A data frame with one row per year (1987-2024) and the columns:

**year** Year of the data.

**file** File name (e.g. 'capes\_1987.parquet').

**url** Download URL of the file.

**bytes** File size in bytes, used to check the download.

**md5** MD5 checksum of the file.

## Source

Derived from the CAPES open data portal, <<https://dadosabertos.capes.gov.br/group/catalogo-de-teses-e-dissertacoes-brasil>>.

## Examples

```
data(capes_years)
head(capes_years)
```

---

download\_capes\_data     *Download CAPES Data*

---

## Description

Downloads the yearly Parquet files of the CAPES Catalog of Theses and Dissertations for the selected years. The file list and download URLs come from [capes\_years].

## Usage

```
download_capes_data(
  years,
  destination = tempdir(),
  timeout = 120,
  base_url = getOption("capesR.base_url")
)

baixar_dados_capes(
  years,
  destination = tempdir(),
  timeout = 120,
  base_url = getOption("capesR.base_url")
)
```

## Arguments

|             |                                                                                                                                                                              |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| years       | A vector with the desired years (1987-2024).                                                                                                                                 |
| destination | The directory where the files will be saved (default: temporary directory).                                                                                                  |
| timeout     | The timeout in seconds for the download process (default: 120 seconds).                                                                                                      |
| base_url    | Optional base URL of a mirror hosting the files listed in [capes_years]. Defaults to 'getOption("capesR.base_url")'; when 'NULL', the 'url' column of [capes_years] is used. |

## Value

A named list (by year) of file paths for the downloaded or already existing files.

## Examples

```
# Download data for the years 1987 and 1990
capes_files <- download_capes_data(c(1987, 1990))
```

---

|                 |                                                                                |
|-----------------|--------------------------------------------------------------------------------|
| read_capes_data | <i>Read and filter data from the CAPES Catalog of Theses and Dissertations</i> |
|-----------------|--------------------------------------------------------------------------------|

---

### Description

This function combines data from multiple Parquet files and applies optional filters, including text-based searches.

### Usage

```
read_capes_data(files, filters = list())
```

```
ler_dados_capes(files, filters = list())
```

### Arguments

|         |                                                                                                                                                                                                                                                                                                                                                                       |
|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| files   | A vector or list of paths to Parquet files.                                                                                                                                                                                                                                                                                                                           |
| filters | A named list of filters. Values for the text columns 'titulo' and 'resumo' (or 'title'/'abstract') are matched as case-insensitive substrings, and a vector of terms keeps the rows containing any of them (e.g., 'list(titulo = c("varicela", "catapora"))'); all other columns are matched exactly (e.g., 'list(ano_base = 1987, uf = "SP", titulo = "educação")'). |

### Value

A 'data.frame' containing the combined and filtered data.

### Examples

```
# Download data for the years 1987 and 1990
capes_files <- download_capes_data(c(1987, 1990))
# Combine all selected data
combined_data <- read_capes_data(capes_files)
```

---

|                   |                                                                                              |
|-------------------|----------------------------------------------------------------------------------------------|
| search_capes_text | <i>Search for terms in text fields of the CAPES Catalog of Theses and Dissertations data</i> |
|-------------------|----------------------------------------------------------------------------------------------|

---

### Description

Searches one or more terms in one or more text columns of a previously loaded 'data.frame'. Matching is literal (not a regular expression) and case-insensitive.

**Usage**

```
search_capes_text(data, term, field, match = c("any", "all"))
```

```
buscar_texto_capes(data, term, field, match = c("any", "all"))
```

**Arguments**

|                    |                                                                                                                                                                                            |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data</code>  | A 'data.frame' containing the CAPES Catalog of Theses and Dissertations data.                                                                                                              |
| <code>term</code>  | A character vector with one or more terms to search for.                                                                                                                                   |
| <code>field</code> | A character vector with the name(s) of the column(s) to search in (e.g., "titulo", 'c("titulo", "resumo")').                                                                               |
| <code>match</code> | How several terms combine: "any" (default) keeps rows where at least one term occurs in at least one field; "all" keeps rows where every term occurs (each in at least one of the fields). |

**Value**

A 'data.frame' with the matching rows. When nothing matches, an empty 'data.frame' with the same columns as 'data' is returned with a message.

**Examples**

```
# Download data for the years 1987 and 1990
capes_files <- download_capes_data(c(1987, 1990))
# Combine all selected data
combined_data <- read_capes_data(capes_files)
# Titles mentioning "Educação"
results <- search_capes_text(combined_data, term = "Educação", field = "titulo")
# Titles or abstracts mentioning either synonym
results <- search_capes_text(
  combined_data,
  term = c("varicela", "catapora"),
  field = c("titulo", "resumo")
)
# Abstracts mentioning both terms
results <- search_capes_text(
  combined_data,
  term = c("saúde", "escola"),
  field = "resumo",
  match = "all"
)
```

# Index

## \* **datasets**

    capes\_synthetic\_df, [2](#)

    capes\_years, [3](#)

baixar\_dados\_capes

    (download\_capes\_data), [4](#)

buscar\_texto\_capes (search\_capes\_text),  
    [5](#)

capes\_synthetic\_df, [2](#)

capes\_years, [3](#)

download\_capes\_data, [4](#)

ler\_dados\_capes (read\_capes\_data), [5](#)

read\_capes\_data, [5](#)

search\_capes\_text, [5](#)