

Package: inspercidos (via r-universe)

October 12, 2026

Title Access to Datasets from Insper Cidades

Version 0.5.0

Description Provides tools to discover, download, and cite datasets from Insper Cidades, hosted on the Insper Dataverse repository. Lists datasets and their logical resources, downloads data directly into R, links each dataset to its replication project, and generates academic citations from repository metadata.

License MIT + file LICENSE

URL <https://github.com/inspercidas/inspercidos>,
<https://inspercidas.github.io/inspercidos/>

BugReports <https://github.com/inspercidas/inspercidos/issues>

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

Depends R (>= 4.1.0)

Imports dataverse (>= 0.3.15), cli, jsonlite, readr, rlang, tibble, utils, tools

Suggests arrow (>= 8.0.0), curl, DBI, dbplyr, dplyr, duckdb (>= 1.5.5), sf, readxl, knitr, rmarkdown, testthat (>= 3.0.0), pkgdown, googlesheets4, janitor, lifecycle, stringr, withr

VignetteBuilder knitr

Config/Needs/website dplyr, ggplot2, sf

Config/testthat/edition 3

Config/pak/sysreqs libxml2-dev libssl-dev libx11-dev

Repository <https://inspercidas.r-universe.dev>

Date/Publication 2026-10-12 13:00:22 UTC

RemoteUrl <https://github.com/inspercidas/inspercidos>

RemoteRef HEAD

RemoteSha bb591d59a5b67450e8ea9695b2f9c462ebeb857

Contents

cite_dataset	2
get_dataset	3
get_dataverse	4
list_datasets	6
list_projects	7
list_resources	8
open_project	9
query_dataset	9
Index	11

cite_dataset	<i>Generate a citation for a dataset</i>
--------------	--

Description

Fetches metadata from Insper Dataverse and returns a formatted citation. The citation is printed to the console and returned invisibly as a character string.

Usage

```
cite_dataset(dataset, format = c("text", "bibtex", "ris"))
```

Arguments

- dataset A dataset identifier: alias, bare DOI, or DOI URL (see [get_dataverse\(\)](#) for the accepted DOI and URL forms).
- format Citation format. One of "text" (default), "bibtex", or "ris".

Value

A character string containing the formatted citation (invisible).

Examples

```
# Plain-text citation
cite_dataset("pemob_anual")

# BibTeX
cite_dataset("pemob_anual", format = "bibtex")

# RIS (Zotero, Mendeley, EndNote)
cite_dataset("pemob_anual", format = "ris")
```

get_dataset

*Download a dataset from Insuper Dataverse***Description**

Downloads a registered dataset into R as a tibble or sf object. Identify the dataset by its short alias; use `get_dataverse()` for a DOI or URL.

Usage

```
get_dataset(
  dataset,
  ...,
  resource = NULL,
  year = NULL,
  format = NULL,
  docs = FALSE
)
```

Arguments

dataset	A registered dataset alias, e.g. "iptu_sp". See <code>list_datasets()</code> for all aliases.
...	Must be empty. Arguments after ... must be named.
resource	The name of a logical resource within the registered dataset. When NULL, the resource marked as the default is used. See <code>list_resources()</code> for the available names.
year	An integer or character year used to filter files when a dataset contains several annual files (e.g. year = 2023).
format	An optional file format, such as "parquet", "gpkg", or "xlsx". The format must be available for the selected resource.
docs	Logical. If TRUE, returns a named list with two elements: data (the downloaded tibble/sf object) and docs. When the dataset has a documentation workbook, that file is downloaded and returned as a tibble. Otherwise docs is a named list of metadata fetched from Dataverse (title, description, authors, DOI, URL, year). Default is FALSE.

Details

A Dataverse deposit may contain several logical resources, each distributed in several formats. Use resource to select the logical dataset and format to select its representation. Deposits with one resource, or a declared default, need neither argument.

When format is omitted, spatial resources prefer GeoPackage and GeoJSON; other resources prefer RDS, Parquet, delimited text, and Excel. Formats that need an uninstalled suggested package are skipped.

Value

When docs = FALSE (default): a [tibble](#) or sf object with the "doi" attribute set. When docs = TRUE: a named list with elements data and docs.

Examples

```
# By alias
embarques <- get_dataset("embarques_mensais")

# Pick one year from a multi-year dataset
pemob_2023 <- get_dataset("pemob_anual", year = 2023)

# Pick a resource and an explicit format
pontos <- get_dataset(
  "qualidade_ar_mare",
  resource = "pontos",
  format = "gpkg"
)

# Return the data with its documentation
result <- get_dataset("embarques_mensais", docs = TRUE)
result$docs

# Spatial datasets return an sf object
faixa_azul <- get_dataset("faixa_azul_sp")
```

get_dataverse

Download any Insper Dataverse dataset from a DOI or URL

Description

Downloads any deposit on Insper Dataverse, registered or not. Copy the DOI or the landing-page URL from Insper Dataverse and paste it straight into R.

Usage

```
get_dataverse(
  x,
  filename = NULL,
  file_pattern = NULL,
  year = NULL,
  files = FALSE
)
```

Arguments

x	A DOI ("10.60873/FK2/AOLEOI"), a DOI URL, or a Dataverse landing-page URL containing persistentId=doi:....
filename, file_pattern, year	Optional selectors passed to the same file-picking logic used by get_dataset() .
files	Logical. If TRUE, list the deposit's files and return their names instead of downloading anything. Useful for inspecting an unfamiliar deposit first.

Details

Registered datasets are better served by [get_dataset\(\)](#), which knows which file to pick. This function has no such knowledge, so it lists the deposit and applies the same format preference to whatever it finds. Deposits with an unusual file layout may need `filename` or `file_pattern` to resolve, and formats the package cannot read will raise an error naming the file.

DOIs under the 10.60873 and 10.48808 prefixes are served.

Value

A [tibble](#) or `sf` object with the "doi" attribute set. When `files = TRUE`, a character vector of file-names.

See Also

[get_dataset\(\)](#) for registered datasets.

Examples

```
# Inspect a deposit before downloading
get_dataverse("10.60873/FK2/YWXLQS", files = TRUE)

# Paste a DOI straight from the Dataverse page
linhas <- get_dataverse("10.60873/FK2/YWXLQS", filename = "dim_line.rds")

# Or the landing-page URL
linhas <- get_dataverse(
  paste0(
    "https://dataverse.datascience.insper.edu.br/dataset.xhtml?",
    "persistentId=doi:10.60873/FK2/YWXLQS"
  ),
  filename = "dim_line.rds"
)
```

list_datasets	<i>List available datasets</i>
---------------	--------------------------------

Description

Returns a tibble of datasets in the inspercidos registry.

Usage

```
list_datasets(
  search = NULL,
  theme = NULL,
  region = NULL,
  project = NULL,
  spatial = NULL,
  include = c("active", "catalogued", "all")
)
```

Arguments

search	A character string to filter results. Matched case-insensitively against alias, title, description, keywords, theme, and region. Pass NULL (default) to return all datasets.
theme, region, project	Optional filters, matched case-insensitively as regular expressions against the corresponding column.
spatial	Logical. If TRUE, keep only datasets that ship a GeoPackage or GeoJSON. If FALSE, drop them. NULL (default) keeps both.
include	Which registry entries to return. "active" (default) lists datasets you can download. "catalogued" adds entries held in Insper's secure data room, which appear with access = "secure_room" and cannot be downloaded. "all" also includes retired aliases.

Value

A [tibble](#) with one row per dataset and columns alias, title, description, theme, region, project, access, is_spatial, formats, keywords, and doi.

Registry updates

The registry is read from the package's GitHub repository once per session, so datasets added after a release appear without reinstalling. When GitHub cannot be reached, or the online registry needs a newer version of the package, the copy shipped with the package is used instead. Set `options(inspercidos.registry = "bundled")` to always use the shipped copy, for example to keep a script's catalog fixed.

See Also

[list_projects\(\)](#) for the studies these datasets come from.

Examples

```
# List all downloadable datasets
list_datasets()

# Search by keyword or theme
list_datasets("metro")
list_datasets("IPTU")

# Only spatial datasets
list_datasets(spatial = TRUE)

# Datasets from one study
list_datasets(project = "faixa_azul")

# Include datasets that are catalogued but not downloadable
list_datasets(include = "catalogued")
```

list_projects

List the studies behind the datasets

Description

Most datasets come from a research project with its own processing pipeline. One project usually produces several datasets, and its pipeline is published as one repository.

Usage

```
list_projects(project = NULL)
```

Arguments

project Optional project slug or regular expression to filter on.

Value

A [tibble](#) with columns project, title, repo_url, visibility, n_datasets, and datasets. repo_url is NA for projects whose code has not been published yet, and visibility is "private" for repositories that only Insper Cidades members can open.

See Also

[open_project\(\)](#) to open a repository, [list_datasets\(\)](#) for the datasets themselves.

Examples

```
list_projects()  
list_projects("motiva")
```

list_resources	<i>List a dataset's logical resources</i>
----------------	---

Description

Returns the logical resources registered within a dataset. A resource is one dataset that may be distributed in several file formats or split by a declared dimension such as year. The information comes from the registry (see [list_datasets\(\)](#)), so Dataverse is not queried.

Usage

```
list_resources(dataset)
```

Arguments

dataset	A registered dataset alias. See list_datasets() .
---------	---

Value

A [tibble](#) with one row per resource and columns resource, title, is_spatial, years, formats, and default.

See Also

[get_dataset\(\)](#) to download a resource.

Examples

```
list_resources("qualidade_ar_mare")  
list_resources("pemob_anual")
```

open_project	<i>Open the study behind a dataset</i>
--------------	--

Description

Prints the study a dataset came from, the other datasets it produced, and the repository holding its processing code, then opens that repository in a browser.

Usage

```
open_project(x, open = TRUE)
```

Arguments

x	A dataset alias (see list_datasets()) or a project slug (see list_projects()).
open	Logical. If TRUE (default), open the repository in a browser when one is registered and the session is interactive.

Details

One study normally produces several datasets, and its pipeline is published as one repository. Some repositories are private and open only for Insper Cidades members; [list_projects\(\)](#) reports which.

Value

The repository URL, invisibly, or NULL when no repository is registered for the project.

Examples

```
# By dataset alias
open_project("sinistros_sp", open = FALSE)

# By project slug
open_project("motiva", open = FALSE)
```

query_dataset	<i>Query a large dataset without downloading it</i>
---------------	---

Description

Opens a registered parquet dataset as a lazy table backed by DuckDB. DuckDB reads the file directly from Insper Dataverse and fetches only the columns and row groups a query needs, so large datasets such as "itbi_sp" can be filtered and summarised without loading them into memory. Use dplyr verbs to build the query and [dplyr::collect\(\)](#) to bring the result into R.

Usage

```
query_dataset(dataset, ..., resource = NULL, year = NULL)
```

Arguments

dataset	A registered dataset alias, e.g. "iptu_sp". See list_datasets() for all aliases.
...	Must be empty. Arguments after ... must be named.
resource	The name of a logical resource within the registered dataset. When NULL, the resource marked as the default is used. See list_resources() for the available names.
year	An integer or character year used to keep a single annual file. When NULL, every annual file is included.

Details

Only resources distributed as parquet can be queried; use [get_dataset\(\)](#) for the others. When a resource is split into annual files and year is NULL, all years are combined into one table.

DuckDB loads httpfs from its shared extension cache or installs it on the first call. The cache is stored under ~/.duckdb.

Value

A lazy [dbplyr](#) table (tbl_duckdb_connection) with the "doi" attribute set.

See Also

[get_dataset\(\)](#) to download a dataset into memory.

Examples

```
itbi <- query_dataset("itbi_sp")

# Transactions per year, computed by DuckDB
itbi |>
  dplyr::count(ano = dplyr::sql("year(data_de_transacao)")) |>
  dplyr::arrange(ano) |>
  dplyr::collect()
```

Index

`cite_dataset`, [2](#)

`dbplyr`, [10](#)

`dplyr::collect()`, [9](#)

`get_dataset`, [3](#)

`get_dataset()`, [5](#), [8](#), [10](#)

`get_dataverse`, [4](#)

`get_dataverse()`, [2](#), [3](#)

`list_datasets`, [6](#)

`list_datasets()`, [3](#), [7–10](#)

`list_projects`, [7](#)

`list_projects()`, [7](#), [9](#)

`list_resources`, [8](#)

`list_resources()`, [3](#), [10](#)

`open_project`, [9](#)

`open_project()`, [7](#)

`query_dataset`, [9](#)

`tibble`, [4–8](#)