Package: suppdata (via r-universe)

October 28, 2024

Type Package

Title Downloading Supplementary Data from Published Manuscripts

Version 1.1-9

Maintainer William D. Pearse <will.pearse@gmail.com>

Description Downloads data supplementary materials from manuscripts, using papers' DOIs as references. Facilitates open, reproducible research workflows: scientists re-analyzing published datasets can work with them as easily as if they were stored on their own computer, and others can track their analysis workflow painlessly. The main function suppdata() returns a (temporary) location on the user's computer where the file is stored, making it simple to use suppdata() with standard functions like read.csv().

License MIT + file LICENSE

URL https://docs.ropensci.org/suppdata/, https://github.com/ropensci/suppdata/

BugReports https://github.com/ropensci/suppdata/

VignetteBuilder knitr

LazyLoad yes

Suggests knitr (>= 1.6), testthat (>= 2.0.0), covr (>= 3.0.1)

Imports httr (>= 1.0.0), xml2 (>= 1.2.0), jsonlite (>= 1.5), rcrossref (>= 0.8.0)

Encoding UTF-8

RoxygenNote 7.2.3

Repository https://ropensci.r-universe.dev

RemoteUrl https://github.com/ropensci/suppdata

RemoteRef master

RemoteSha 2b00410c31ba8f4f645f47cbd9c6e8e330fca87a

Contents

---

suppdata *Download supplementary materials from journals*

Description

Put a call to this function where you would put a file-path - everything is cached by default, so you don't have to worry about multiple downloads in the same session.

Usage

```
suppdata(
  x,
  si,
  from = c("auto", "plos", "wiley", "science", "proceedings", "figshare",
   "esa_data_archives", "esa_archives", "biorxiv", "epmc", "peerj", "copernicus",
    "jstatsoft"),
  save.name = NA,
  dir = NA,
  cache = TRUE,
  vol = NA,
  issue = NA,
  list = FALSE,
  timeout = 10,
  zip = FALSE
)

## S3 method for class 'character'
suppdata(
  x,
  si,
  from = c("auto", "plos", "wiley", "science", "proceedings", "figshare",
   "esa_data_archives", "esa_archives", "biorxiv", "epmc", "peerj", "copernicus",
    "jstatsoft"),
  save.name = NA,
  dir = NA,
  cache = TRUE,
  vol = NA,
  issue = NA,
  list = FALSE,
  timeout = 10,
  zip = FALSE
)
```

Arguments

x	One of: vector of DOI(s) of article(s) (a character) or ESA-specific article code.
si	number of the supplementary information (SI) to be downloaded (1, 2, 3, etc.), or (for ESA, Science, and Copernicus journals) the name of the supplement (e.g., "S1_data.csv"). Can be a character or numeric.
from	Publisher of article (character). The default (auto) uses crossref (cr_works) to detect the journal's publisher. Specifying the journal can somewhat speed up your download, or be used to force a download from EPMC (see details). You *must* specify if downloading from an ESA journal (esa_data_archives, esa_archives). You can only use this argument if x is a vector of DOI(s). Must be one of: auto (i.e., auto-detect journal; default), plos, wiley, science, proceedings, figshare, esa_data_archives, esa_archives, biorxiv, epmc, peerj, copernicus, (Data)dryad), mdpi, or jstatsoft.
save.name	a name for the file to download (character). If NA (default) this will be a combination of the DOI and SI number
dir	directory to save file to (character). If NA (default) this will be a temporary directory created for your files
cache	if TRUE (default), the file won't be downloaded again if it already exists (in a temporary directory creates, or your chosen dir)
vol	Article volume (Proceedings journals only; numeric)
issue	Article issue (Proceedings journals only; numeric)
list	if TRUE, print all files within a zip-file downloaded from EPMC (default: FALSE). This is *very* useful if using EPMC (see details)
timeout	how long to wait for successful download (default 10 seconds)
zip	if TRUE, force download in binary format in order to ensure that zipped files will work on Windows (default FALSE).

Details

The examples probably give the best indication of how to use this function. In general, just specify the DOI of the article you want to download data from, and the number of the supplement you want to download (1, 5, etc.). Proceedings, and Science journals need you to give the filename of the supplement to download. The file extensions (suffixes) of files are returned as suffix attributes (see first example), which may be useful if you don't know the format of the file you're downloading.

For any DOIs not recognised (and if asked) the European PubMed Central API is used to look up articles. What this database calls a supplementary file varies by publisher; often they will simply be figures within articles, but we (obviously) have no way to check this at run-time. I strongly recommend you run any EPMC calls with list=TRUE the first time, to see the filenames that EPMC gives supplements, as these also often vary from what the authors gave them. This may actually be a 'feature', not a 'bug', if you're trying to automate some sort of meta-analysis.

Below is a list of all the publishers this supports, and examples of journals from them.

auto Default. Use a cross-ref search (cr_works) on the DOI to determine the publisher.

plos Public Library of Science journals (e.g., PLoS One;)

wiley Wiley journals (e.g., Ecology Letters)

science Science magazine (e.g., Science Advances)

proceedings Royal Society of London journals (e.g., Proceedings of the Royal Society of London B). Requires `vol` and `issue` of the article.

figshare Figshare

biorxiv Load from bioRxiv

epmc Look up an article on the Europe PubMed Central, and then download the file using their supplementary materials API. See comments above in 'notes' about EPMC.

peerj PeerJ journals (e.g., PeerJ Preprints).

copernicus Copernicus Publications journals (e.g., Biogeosciences). Only one supplemental is supported, which can be a zip archive or a PDF file. A numeric `si` parameter must be 1 to download the whole archive, which is saved using Copernicus naming scheme (<journalname>-<volume>-<firstpage>-<year>-supplement.zip) and `save.name` is ignored, or to download the PDF. If `si` matches the name of the supplemental archive (i.e. uses the Copernicus naming scheme), then the suppdata archive is not unzipped. `si` may be the name of a file in that archive, so only that file is extracted and saved to `save.name`.

Note

Make sure that the article from which you're attempting to download supplementary materials *has* supplementary materials. 404 errors and 'file not found' errors can result from such cases.

Author(s)

Will Pearse (<will.pearse@usu.edu>) and Scott Chamberlain (<myrmecocystus@gmail.com>)

Examples

```
# NOTE: The examples below are flagged as 'dontrun' to avoid
# running downloads repeatedly on CRAN servers
## Not run:
#Put the function wherever you would put a file path
crabs <- read.csv(suppdata("10.6084/m9.figshare.979288", 2))

epmc.fig <- suppdata("10.1371/journal.pone.0126524",
                     "pone.0126524.s002.jpg", "epmc")
#...note this 'SI' is not actually an SI, but rather an image from the paper.

#View the suffix (file extension) of downloaded files
# - note that not all files are uploaded/stored with useful file extensions!
attr(epmc.fig, "suffix")

copernicus.csv <- suppdata("10.5194/bg-14-1739-2017",
                           "Table S1 v2 UFK FOR_PUBLICATION.csv",
                           save.name = "data.csv")
#...note this 'SI' is not an SI but the name of a file in the supplementary information archive.
```

```
## End(Not run)
# (examples not run on CRAN to avoid downloading files repeatedly)
```

Index