Package ‘ojsr’

November 13, 2024

Type Package

Title Crawler and Data Scraper for Open Journal System ('OJS')

Version 0.1.5

Description Crawler for 'OJS' pages and scraper for meta-data from articles.
You can crawl 'OJS' archives, issues, articles, galleys, and search results.
You can scrape articles metadata from their head tag in html,
or from Open Archives Initiative ('OAI') records.
Most of these functions rely on 'OJS' routing conventions
(<https://docs.pkp.sfu.ca/dev/documentation/en/architecture-routes>).

License GPL-3

Encoding UTF-8

Imports dplyr (>= 0.8.3), magrittr, rvest, tidyr (>= 1.0), urltools,
xml2, purrr, rlang, RCurl

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

URL https://github.com/gastonbecerra/ojsr

BugReports https://github.com/gastonbecerra/ojsr/issues

RoxygenNote 7.3.2

NeedsCompilation no

Author Gaston Becerra [aut, cre] (<https://orcid.org/0000-0001-9432-8848>)

Maintainer Gaston Becerra <gaston.becerra@gmail.com>

Repository CRAN

Date/Publication 2024-11-13 13:00:04 UTC

Contents

get_articles_from_issue

Scraping articles URLs from the ToC of OJS issues

Description

Takes a vector of OJS (issue) URLs and scrapes the links to articles from the issues table of content

Usage

```
get_articles_from_issue(input_url, verbose = FALSE)
```

Arguments

input_url	Character vector.
verbose	Logical.

Value

A long-format dataframe with the url you provided (input_url) and the articles url scrapped (output_url)

Examples

```
issue <- 'https://revistas.ucn.cl/index.php/saludysociedad/issue/view/65'
articles <- ojsr::get_articles_from_issue(input_url = issue)
```

get_articles_from_search

Scraping OJS search results for a given criteria to retrieve articles' URL

Description

takes a vector of OJS URLs and a string for search criteria to compose search result URLs, (including pagination) then it scrapes them to retrieve the articles' URLs.

Usage

```
get_articles_from_search(input_url, search_criteria, verbose = FALSE)
```

Arguments

`input_url`	Character vector.
`search_criteria`	Character string
`verbose`	Logical.

Value

A dataframe with the urls of the articles linked from the OJS issue page.

Examples

```
journals <- c(
   'https://revistapsicologia.uchile.cl/index.php/RDP/',
   'https://publicaciones.sociales.uba.ar/index.php/psicologiasocial/'
)
criteria <- "actitudes"
search_result_pages <- ojsr::get_articles_from_search(input_url = journals,
    search_criteria = criteria, verbose = TRUE)
```

`get_galleys_from_article`

Scraping galleys URLs from OJS articles

Description

Takes a vector of OJS URLs and scrapes all the galleys URLs from the article view

Usage

```
get_galleys_from_article(input_url, verbose = FALSE)
```

Arguments

`input_url`	Character vector.
`verbose`	Logical.

Value

A long-format dataframe with the url you provided (input_url), the articles url scrapped (output_url), the format of the galley (format), and the url that forces download of the galley (download_url)

Examples

```
article <- 'https://revistapsicologia.uchile.cl/index.php/RDP/article/view/55657'
galleys <- ojsr::get_galleys_from_article(input_url = article)
```

get_html_meta_from_article

Scraping metadata from the OJS articles HTML

Description

Takes a vector of OJS URLs and scrapes all metadata written in HTML from the article view

Usage

```
get_html_meta_from_article(input_url, verbose = FALSE)
```

Arguments

input_url	Character vector.
verbose	Logical.

Value

A long-format dataframe with the url you provided (input_url), the name of the metadata (meta_data_name), the content of the metadata (meta_data_content), the standard in which the content is annotated (meta_data_scheme), and the language in which the metadata was entered (meta_data_xmllang)

Examples

```
article <- 'https://dspace.palermo.edu/ojs/index.php/psicodebate/article/view/516/311'
metadata <- ojsr::get_html_meta_from_article(article)
```

get_issues_from_archive

Scraping issues' URLs from the OJS issues archive

Description

Takes a vector of OJS URLs and scrapes the issues URLs from the issue archive.

Usage

```
get_issues_from_archive(input_url, verbose = FALSE)
```

Arguments

`input_url`	Character vector.
`verbose`	Logical.

Value

A long-format dataframe with the url you provided (input_url) and the url of issues found (output_url)

Examples

```
journal <- 'https://dspace.palermo.edu/ojs/index.php/psicodebate/issue/archive'
issues <- ojsr::get_issues_from_archive(input_url = journal)
```

`get_oai_meta_from_article` *Retrieving OAI records for OJS articles*

Description

This functions access OAI records (within OJS) for any article for which you provided an URL.

Usage

```
get_oai_meta_from_article(input_url, verbose = FALSE)
```

Arguments

`input_url`	Character vector.
`verbose`	Logical.

Details

Several limitations are in place. Please refer to vignette.

Value

A long-format dataframe with the url you provided (input_url), the name of the metadata (meta_data_name), and the content of the metadata (meta_data_content).

Examples

```
article <- 'https://dspace.palermo.edu/ojs/index.php/psicodebate/article/view/516/311'
metadata_oai <- ojsr::get_oai_meta_from_article(input_url = article)
```

parse_base_url *Parses urls against OJS routing conventions and retrieves the base url*

Description

Takes a vector of urls and parses them according to OJS routing conventions, then retrieves OJS base url.

Usage

```
parse_base_url(input_url)
```

Arguments

input_url	Character vector.

Value

A vector of the same length of your input.

Examples

```
mix_links <- c(
  'https://dspace.palermo.edu/ojs/index.php/psicodebate/issue/archive',
  'https://publicaciones.sociales.uba.ar/index.php/psicologiasocial/article/view/2903'
)
base_url <- ojsr::parse_base_url(input_url = mix_links)
```

parse_oai_url *Parses urls against OJS routing conventions and retrieves the OAI url*

Description

Takes a vector of urls and parses them according to OJS routing conventions, then retrieves OAI entry url.

Usage

```
parse_oai_url(input_url)
```

Arguments

input_url	Character vector.

Value

A vector of the same length of your input.

Examples

```
mix_links <- c(
  'https://dspace.palermo.edu/ojs/index.php/psicodebate/issue/archive',
  'https://publicaciones.sociales.uba.ar/index.php/psicologiasocial/article/view/2903'
)
oai_url <- ojsr::parse_oai_url(input_url = mix_links)
```

Index