Package ‘europepmc’

April 20, 2018

Title  R Interface to the Europe PubMed Central RESTful Web Service
Version  0.3
License  GPL-3
Date  2018-04-19

URL  http://github.com/ropensci/europepmc/

Description  An R Client for the Europe PubMed Central RESTful Web Service (see <https://europepmc.org/RestfulWebService> for more information). It gives access to both metadata on life science literature and open access full texts. Europe PMC indexes all PubMed content and other literature sources including Agricola, a bibliographic database of citations to the agricultural literature, or Biological Patents. In addition to bibliographic metadata, the client allows users to fetch citations and reference lists. Links between life-science literature and other EBI databases, including ENA, PDB or ChEMBL are also accessible. No registration or API key is required. See the vignettes for usage examples.

LazyLoad  yes
LazyData  yes
VignetteBuilder  knitr
Depends  R (>= 3.00)
Imports  httr, jsonlite, plyr, dplyr, progress, urltools, purrr, xml2
RoxygenNote  6.0.1
Suggests  testthat, knitr, rmarkdown, ggplot2, readr
Encoding  UTF-8

Author  Najko Jahn [aut, cre, cph],
         Maëlle Salmon [ctb]
Maintainer  Najko Jahn <najko.jahn@gmail.com>

Repository  CRAN
Date/Publication  2018-04-20 07:41:01 UTC
epmc_citations

Get citations for a given publication

Finds works that cite a given publication.

Usage

epmc_citations(ext_id = NULL, data_src = "med", limit = 100, verbose = TRUE)

Arguments

- **ext_id**: character, publication identifier
- **data_src**: character, data source, by default Pubmed/MedLine index will be searched. The following three letter codes represent the sources Europe PubMed Central supports:
  - **agr**: Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
  - **cba**: Chinese Biological Abstracts
  - **ctx**: CiteXplore
  - **eth**: EthOs Theses, i.e. PhD theses (British Library)
  - **hir**: NHS Evidence
  - **med**: PubMed/Medline NLM
epmc_db

- **nbk** Europe PMC Book metadata
- **pat** Biological Patents
- **pmc** PubMed Central

**limit** integer, number of results. By default, this function returns 100 records.

**verbose** logical, print some information on what is going on.

**Value**

Metadata of citing documents as data.frame

**Examples**

```r
# Not run:
epmc_citations("PMC3166943", data_src = "pmc")
epmc_citations("9338777")

# End(Not run)
```

---

**epmc_db**

*Retrieve external database entities referenced in a given publication*

**Description**

This function returns EBI database entities referenced in a publication from Europe PMC RESTful Web Service.

**Usage**

```r
epmc_db(ext_id = NULL, data_src = "med", db = NULL, limit = 100,
        verbose = TRUE)
```

**Arguments**

- **ext_id** character, publication identifier
- **data_src** character, data source, by default Pubmed/MedLine index will be searched. The following three letter codes represent the sources Europe PubMed Central supports:
  - **agr** Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
  - **cba** Chinese Biological Abstracts
  - **ctx** CiteXplore
  - **eth** EthOs Theses, i.e. PhD theses (British Library)
  - **hir** NHS Evidence
  - **med** PubMed/Medline NLM
  - **nbk** Europe PMC Book metadata
pat  Biological Patents

pmc  PubMed Central

db  character, specify database:

'ARXPR'  Array Express, a database of functional genomics experiments [https://www.ebi.ac.uk/arrayexpress/]

'CHEBI'  a database and ontology of chemical entities of biological interest [http://www.ebi.ac.uk/chebi/]

'CHEMBL'  a database of bioactive drug-like small molecules [https://www.ebi.ac.uk/chembldb/]

'EMBL'  now ENA, provides a comprehensive record of the world’s nucleotide sequencing information [http://www.ebi.ac.uk/ena/]

'INTACT'  provides a freely available, open source database system and analysis tools for molecular interaction data [http://www.ebi.ac.uk/intact/]

'INTERPRO'  provides functional analysis of proteins by classifying them into families and predicting domains and important sites [http://www.ebi.ac.uk/interpro/]


'PDB'  European resource for the collection, organisation and dissemination of data on biological macromolecular structures [http://www.ebi.ac.uk/pdbe/]

'UNIPROT'  comprehensive and freely accessible resource of protein sequence and functional information [http://www.uniprot.org/]

'PRIDE'  PRIDE Archive - proteomics data repository [https://www.ebi.ac.uk/pride/archive/]

limit  integer, number of results. By default, this function returns 100 records.

verbose  logical, print some information on what is going on.

Value

Cross-references as data.frame

Examples

## Not run:
epmc_db("12368864", db = "uniprot", limit = 150)
epmc_db("25249410", db = "embl")
epmc_db("14756321", db = "uniprot")
epmc_db("11805837", db = "pride")

## End(Not run)
Description

This function returns the number of EBI database links associated with a publication.

Usage

epmc_db_count(ext_id = NULL, data_src = "med")

Arguments

- **ext_id**: character, publication identifier
- **data_src**: character, data source, by default Pubmed/MedLine index will be searched.

Details

Europe PMC supports cross-references between literature and the following databases:

- **'ARXPR’** Array Express, a database of functional genomics experiments [https://www.ebi.ac.uk/arrayexpress/](https://www.ebi.ac.uk/arrayexpress/)
- **'CHEBI’** a database and ontology of chemical entities of biological interest [http://www.ebi.ac.uk/chebi/](http://www.ebi.ac.uk/chebi/)
- **'CHEMBL’** a database of bioactive drug-like small molecules [https://www.ebi.ac.uk/chembl/db/](https://www.ebi.ac.uk/chembl/db/)
- **'EMBL’** now ENA, provides a comprehensive record of the world’s nucleotide sequencing information [http://www.ebi.ac.uk/ena/](http://www.ebi.ac.uk/ena/)
- **'INTACT’** provides a freely available, open source database system and analysis tools for molecular interaction data [http://www.ebi.ac.uk/intact/](http://www.ebi.ac.uk/intact/)
- **'INTERPRO’** provides functional analysis of proteins by classifying them into families and predicting domains and important sites [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)
- **'PDB’** European resource for the collection, organisation and dissemination of data on biological macromolecular structures [http://www.ebi.ac.uk/pdbe/](http://www.ebi.ac.uk/pdbe/)
- **'PRIDE’** PRIDE Archive - proteomics data repository [https://www.ebi.ac.uk/pride/archive/](https://www.ebi.ac.uk/pride/archive/)

Value

data.frame with counts for each database
epmc_details

Get details for individual records

Description

This function returns full metadata for a given publication ID including abstract, full text links, author details including ORCID and affiliation, MeSH terms, chemicals, grants.

Usage

epmc_details(ext_id = NULL, data_src = "med")

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ext_id</td>
<td>character, publication identifier</td>
</tr>
<tr>
<td>data_src</td>
<td>character, data source. by default Pubmed/MedLine index will be searched. Other sources Europe PubMed Central supports are:</td>
</tr>
<tr>
<td></td>
<td>agr Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.</td>
</tr>
<tr>
<td></td>
<td>cha Chinese Biological Abstracts</td>
</tr>
<tr>
<td></td>
<td>ctx CiteXplore</td>
</tr>
<tr>
<td></td>
<td>eth EthOs Theses, i.e. PhD theses (British Library)</td>
</tr>
<tr>
<td></td>
<td>hir NHS Evidence</td>
</tr>
<tr>
<td></td>
<td>med PubMed/Medline NLM</td>
</tr>
<tr>
<td></td>
<td>pat Biological Patents</td>
</tr>
<tr>
<td></td>
<td>pmc PubMed Central</td>
</tr>
</tbody>
</table>

Value

list of data frames

Examples

### Not run:
epmc_details(ext_id = "26980001")
epmc_details(ext_id = "24270414")

# PMC record
epmc_details(ext_id = "PMC4747116", data_src = "pmc")
epmc_ftxt

Fetch Europe PMC full texts

Description

This function loads full texts into R. Full texts are in XML format and are only provided for the Open Access subset of Europe PMC.

Usage

epmc_ftxt(ext_id = NULL)

Arguments

ext_id character, publication identifier. All full text publications have external IDs starting with 'PMC___'.

Value

xml_document

Examples

## Not run:
epmc_ftxt("PMC3257301")
epmc_ftxt("PMC3639880")

## End(Not run)
epmc_ftxt_book  Fetch Europe PMC books

Description
Use this function to retrieve book XML formatted full text for the Open Access subset of the Europe PMC bookshelf.

Usage
epmc_ftxt_book(ext_id = NULL)

Arguments
ext_id character, publication identifier. All book full texts are accessible either by the PMID or the 'NBK' book number.

Value
xml_document

Examples
## Not run:
epmc_ftxt_book("NBK32884")
## End(Not run)

epmc_hits  Get search result count

Description
Search over Europe PMC and retrieve the number of results found

Usage
epmc_hits(query = NULL, ...)

Arguments
query query in the Europe PMC syntax
... add query parameters from 'epmc_search()', e.g. synonym=true

See Also
epmc_search
### epmc_hits_trend

**Examples**

```r
## Not run:
epmc_hits('abstract:"burkholderia pseudomallei"')
epmc_hits('AUTHORID:"0000-0002-7635-3473"')

## End(Not run)
```

---

**epmc_hits_trend**  
*Get the yearly number of hits for a query and the total yearly number of hits for a given period*

---

**Description**

Get the yearly number of hits for a query and the total yearly number of hits for a given period

**Usage**

```r
epmc_hits_trend(query, synonym = TRUE, data_src = "med", period = 1975:2016)
```

**Arguments**

- **query**  
  query in the Europe PMC syntax

- **synonym**  
  logical, synonym search. If TRUE, synonym terms from MeSH terminology and the UniProt synonym list are queried, too. Disabled by default.

- **data_src**  
  character, data source, by default Pubmed/MedLine index (med) will be searched. The following three letter codes represent the sources, which are currently supported:
  - **agr**  
    Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
  - **cba**  
    Chinese Biological Abstracts
  - **ctx**  
    CiteXplore
  - **eth**  
    EthOs Theses, i.e. PhD theses (British Library)
  - **hir**  
    NHS Evidence
  - **med**  
    PubMed/Medline NLM
  - **nbk**  
    Europe PMC Book metadata
  - **pat**  
    Biological Patents
  - **pmc**  
    PubMed Central

- **period**  
  a vector of years (numeric) over which to perform the search

**Details**

A similar function was used in [http://www.masalmon.eu/2017/05/14/evergreenreviewgraph/](http://www.masalmon.eu/2017/05/14/evergreenreviewgraph/) where it was advised not to plot no. of hits over time for a query, but to normalize it by the total no. of hits.
Value

A data frame (dplyr tbl_df) with year, total number of hits (all_hits) and number of hits for the query (query_hits).

Examples

```r
## Not run:
# aspirin as query
epmc_hits_trend('aspirin', period = 2006:2016, synonym = FALSE)
# link to cran packages in reference lists
epmc_hits_trend('REF:cran.r-project.org*', period = 2006:2016, synonym = FALSE)
# more complex with publication type review
epmc_hits_trend('(REF:cran.r-project.org*) AND (PUB_TYPE:"Review" OR PUB_TYPE:"review-article")', period = 2006:2016, synonym = FALSE)

## End(Not run)
```

epmc_lablinks

Get links to external sources

Description

With the External Link services, Europe PMC allows third parties to publish links from Europe PMC to other webpages or tools. Current External Link providers, which can be selected through Europe PMC's advanced search, include Wikipedia, Dryad Digital Repository or other open services. For more information, see [http://europepmc.org/labslink](http://europepmc.org/labslink).

Usage

epmc_lablinks(ext_id = NULL, data_src = "med", lab_id = NULL, limit = 100, verbose = TRUE)

Arguments

- **ext_id**: publication identifier
- **data_src**: data source, by default Pubmed/Medline index will be searched. The following three letter codes represents the sources Europe PubMed Central supports:
  - **agr**: Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
  - **cba**: Chinese Biological Abstracts
  - **ctx**: CiteXplore
  - **eth**: EthOs Theses, i.e. PhD theses (British Library)
  - **hir**: NHS Evidence
  - **med**: PubMed/Medline NLM
  - **nbk**: Europe PMC Book metadata
  - **pat**: Biological Patents
epmc_lablinks_count

pmc PubMed Central

lab_id character vector, identifiers of the external link service. Use Europe PMC’s advanced search form to find ids.

limit Number of records to be returned. By default, this function returns 100 records.

verbose print information about what’s going on

Value

Links found as nested data_frame

Examples

## Not run:
# Fetch links
epmc_lablinks("24007304")
# Link to Altmetric (lab_id = "1562")
epmc_lablinks("25389392", lab_id = "1562")

# Links to Wikipedia
epmc_lablinks("24007304", lab_id = "1507")

# Link to full text copy archived through the institutional repo of Bielefeld University
epmc_lablinks("12736239", lab_id = "1056")

## End(Not run)

epmc_lablinks_count Summarise links to external sources

Description

With the External Link services, Europe PMC allows third parties to publish links from Europe PMC to other webpages or tools. Current External Link providers, which can be selected through Europe PMC’s advanced search, include Wikipedia, Dryad Digital Repository or the institutional repo of Bielefeld University. For more information, see http://europepmc.org/labslink.

Usage

epmc_lablinks_count(ext_id = NULL, data_src = "med")

Arguments

data_src publication identifier
data_src (data source, by default Pubmed/MedLine index will be searched. The following three letter codes represents the sources Europe PubMed Central supports:

agr Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
Value

data.frame with counts for each database

Examples

## Not run:
epmc_lablinks_count("24023770")
epmc_lablinks_count("PMC3986813", data_src = "pmc")

## End(Not run)

epmc_profile

Obtain a summary of hit counts

Description

This function returns the number of results found for your query, and breaks it down to the various publication types, data sources, and subsets Europe PMC provides.

Usage

ePMC_profile(query = NULL, synonym = TRUE)

Arguments

query character, search query. For more information on how to build a search query, see http://europepmc.org/Help

synonym logical, synonym search. If TRUE, synonym terms from MeSH terminology and the UniProt synonym list are queried, too. Enabled by default.

Examples

## Not run:
ePMC_profile('malaria')
# use field search, e.g. query materials and reference section for
# mentions of "ropensci"
ePMC_profile('(METHODS:"ropensci")')

## End(Not run)
epmc_refs

Get references for a given publication

Description
This function retrieves all the works listed in the bibliography of a given article.

Usage
epmc_refs(ext_id = NULL, data_src = "med", limit = 100, verbose = TRUE)

Arguments
- ext_id: character, publication identifier
- data_src: character, data source, by default Pubmed/MedLine index will be searched. The following three letter codes represent the sources Europe PubMed Central supports:
  - agr: Agricola is a bibliographic database of citations to the agricultural literature created by the US National Agricultural Library and its co-operators.
  - cba: Chinese Biological Abstracts
  - ctx: CiteXplore
  - eth: EthOs Theses, i.e. PhD theses (British Library)
  - hir: NHS Evidence
  - med: PubMed/Medline NLM
  - nbk: Europe PMC Book metadata
  - pat: Biological Patents
  - pmc: PubMed Central
- limit: integer, number of results. By default, this function returns 100 records.
- verbose: logical, print some information on what is going on.

Value
returns reference section as tibble

Examples
```r
## Not run:
epmc_refs("PMC3166943", data_src = "pmc")
epmc_refs("25378340")
epmc_refs("21753913")

## End(Not run)
```
Description

This is the main function to search Europe PMC RESTful Web Service (http://europepmc.org/RestfulWebService). It fully supports the comprehensive Europe PMC query language. Simply copy & paste your query terms to R. To get familiar with the Europe PMC query syntax, check the Advanced Search Query Builder https://europepmc.org/advancesearch.

Usage

epmc_search(query = NULL, output = "parsed", synonym = TRUE, verbose = TRUE, limit = 100, sort = NULL)

Arguments

query character, search query. For more information on how to build a search query, see http://europepmc.org/Help
output character, what kind of output should be returned. One of 'parsed', 'id_list' or 'raw' As default, parsed key metadata will be returned as data.frame. 'id_list' returns a list of IDs and sources. Use 'raw' to get full metadata as list. Please be aware that these lists can become very large.
synonym logical, synonym search. If TRUE, synonym terms from MeSH terminology and the UniProt synonym list are queried, too. Disabled by default.
verbose logical, print progress bar. Activated by default.
limit integer, limit the number of records you wish to retrieve. By default, 100 are returned.
sort character, relevance ranking is used by default. Use sort = 'cited' for sorting by the number of citations, or sort = 'date' by the most recent publications.

Value

tibble

See Also

http://europepmc.org/Help

Examples

## Not run:
#Search articles for 'Gabi-Kat'
my.data <- epmc_search(query='Gabi-Kat')

#Get article metadata by DOI
my.data <- epmc_search(query = 'DOI:10.1007/bf00197367')

# Get article metadata by PubMed ID (PMID)
my.data <- epmc_search(query = 'EXT_ID:22246381')

# Get only PLOS Genetics article with EMBL database references
my.data <- epmc_search(query = 'ISSN:1553-7404 HAS_EMBL:y')
# Limit search to 250 PLOS Genetics articles
my.data <- epmc_search(query = 'ISSN:1553-7404', limit = 250)

# Exclude MeSH synonyms in search
my.data <- epmc_search(query = 'aspirin', synonym = FALSE)

# Get 100 most cited articles from PLOS ONE published in 2014
epmc_search(query = '(ISSN:1932-6203) AND FIRST_PDATE:2014', sort = 'cited')

# Print number of records found
attr(my.data, "hit_count")

# Change output

## End(Not run)

---

**Description**

In general, use `epmc_search` instead. It calls this function, calling all pages within the defined limit.

**Usage**

```
epmc_search_(query = NULL, limit = 100, output = "parsed",
             page_token = NULL, ...)
```

**Arguments**

- **query** character, search query. For more information on how to build a search query, see [http://europepmc.org/Help](http://europepmc.org/Help)
- **limit** integer, limit the number of records you wish to retrieve. By default, 25 are returned.
- **output** character, what kind of output should be returned. One of 'parsed', 'id_list' or 'raw' As default, parsed key metadata will be returned as data.frame. 'id_list returns a list of IDs and sources. Use 'raw' to get full metadata as list. Please be aware that these lists can become very large.
- **page_token** cursor marking the page
- **...** further params from `epmc_search`
See Also

epmc_search

epmc_tm  Get text-mined terms

Description

Retrieve a count and a list of text-mined terms gathered by Europe PMC.

Usage

epmc_tm(ext_id = NULL, data_src = "med", semantic_type = NULL,
limit = 100, verbose = TRUE)

Arguments

ext_id character, publication identifier
data_src character, data source, by default Pubmed/MedLine index will be searched. The only sources relevant to this methods are med and pmc.
  med   Publications from PubMed and MEDLINE
  pmc   Publications from PubMed Central
semantic_type controlled vocabulary. Specify the semantic type you wish to retrieve. The following types are supported:
  ACCESSION accession IDs for DNA or protein sequences
  CHEMICAL chemicals
  DISEASE diseases
  EFO    Experimental Factor Ontology (http://www.ebi.ac.uk/efo/)
  GENE_PROTEIN gene proteins
  GO_TERM Gene Ontology Terms (http://geneontology.org/)
  ORGANISM organism
limit Number of records to be returned. By default, this function returns 100 records.
verbose print information about what's going on

Value

Terms found as tibble

Examples

## Not run:
epmc_tm("25249410")
epmc_tm("25249410", semantic_type = "GO_TERM")
epmc_tm("PMC4348542", data_src = "pmc", semantic_type = "GO_TERM")

## End(Not run)
Retrieve the number of database links from Europe PMC publication database

Description

This function returns the number of database links for each EBI database from Europe PMC RESTful Web Service.

Usage

epmc_tm_count(ext_id = NULL, data_src = "med")

Arguments

- **ext_id** character, publication identifier
- **data_src** character, data source, by default Pubmed/MedLine index will be searched.

Details

Europe PMC supports the following semantic types:

- **ACCESSION** accession IDs for DNA or protein sequences
- **CHEMICAL** chemicals
- **DISEASE** diseases
- **EFO** Experimental Factor Ontology ([http://www.ebi.ac.uk/efo/](http://www.ebi.ac.uk/efo/))
- **GENE_PROTEIN** gene proteins
- **ORGANISM** organism

Value

data.frame with counts for each semantic type

Examples

```r
## Not run:
epmc_tm_count("PMC4340542", data_src = "pmc")
```

## End(Not run)
Description

europepmc - an R client for the Europe PMC RESTful article API

What is europepmc?

europepmc facilitates access to Europe PMC RESTful Web Service. Europe PMC covers life science literature and gives access to open access full texts. Coverage is not only restricted to Europe, but articles and abstracts are indexed from all over the world. Europe PMC ingests all PubMed content and extends its index with other sources, including Agricola, a bibliographic database of citations to the agricultural literature, or Biological Patents.

Besides searching abstracts and full text, europepmc can be used to retrieve reference sections and citations, text-mined terms or cross-links to other databases hosted by the European Bioinformatics Institute (EBI).

Index

epmc_citations, 2
epmc_db, 3
epmc_db_count, 5
epmc_details, 6
epmc_ftxt, 7
epmc_ftxt_book, 8
epmc_hits, 8
epmc_hits_trend, 9
epmc_lablinks, 10
epmc_lablinks_count, 11
epmc_profile, 12
epmc.refs, 13
epmc_search, 8, 14, 15, 16
epmc_search_, 15
epmc_tm, 16
epmc_tm_count, 17
europepmc, 18
europepmc-package (europepmc), 18