Package 'rentrez'

June 26, 2015

Version 0.4.2

Date 2015-06-25

Title Entrez in R

Depends R (>= 2.6.0), XML

Imports httr, jsonlite

Suggests testthat, knitr

Description Provides an R interact to the NCBI's EUtils API allowing users to search databases like genbank and pubmed, process the resulting files and pull data into their R sessions.

VignetteBuilder knitr

License MIT + file LICENSE

NeedsCompilation no

Author David Winter [aut, cre], Scott Chamberlain [ctb]

Maintainer David Winter <david.winter@gmail.com>

Repository CRAN

Date/Publication 2015-06-26 18:23:37

R topics documented:

entrez_citmatch 2
entrez_dbs 3
entrez_db_links 3
entrez_db_searchable 4
entrez_db_summary 5
entrez_fetch 6
entrez_global_query 7
entrez_info 7
entrez_link 8
entrez_post 9
entrez_search 10

entrez_summary 11
parse_pubmed_xml 12
rentrez 13

Index **14**

`entrez_citmatch` *Fetch pubmed ids from specially formatted citation strings*

Description

Fetch pubmed ids from specially formatted citation strings

Usage

```
entrez_citmatch(bdata, db = "pubmed", retmode = "xml", config = NULL)
```

Arguments

`bdata`	character, containing citation data. Each citation must be represented in a pipe-delimited format journal_title|year|volume|first_page|author_name|your_key| The final field "your_key" is arbitrary, and can used as you see fit. Fields can be left empty, but be sure to keep 6 pipes.
`db`	character, the database to search. Defaults to pubmed, the only database currently available
`retmode`	character, file format to retrieve. Defaults to xml, as per the API documentation, though, note the API only returns plain text
`config`	vector configuration options passed to httr::GET

Value

A character vector containing PMIDs

See Also

`config` for available configs

Examples

```
ex_cites <- c("proc natl acad sci u s a|1991|88|3248|mann bj|test1|",
              "science|1987|235|182|palmenberg ac|test2|")
entrez_citmatch(ex_cites)
```

entrez_dbs *List databases avaliable from the NCBI*

Description

Retreives the names of databases available through the EUtils API

Usage

```
entrez_dbs(config = NULL)
```

Arguments

config	config vector passed to httr::GET

Value

character vector listing available dbs

See Also

Other einfo: entrez_db_links; entrez_db_searchable; entrez_db_summary; entrez_info

Examples

```
entrez_dbs()
```

entrez_db_links *Can be used in conjunction with entrez_link to find the right name for the db argument in that function.*

Description

Can be used in conjunction with entrez_link to find the right name for the db argument in that function.

Usage

```
entrez_db_links(db, config = NULL)
```

Arguments

db	character, name of database t
config	config vector passed to httr::GET

Value

An eInfoLink object (sub-classed from list) summarising linked-databases. Can be coerced to a data-frame with `as.data.frame`. Printing the object the name of each element (which is the correct name for `entrez_link`, and can be used to get (a little) more information about each linked database (see example below).

See Also

`entrez_link`

Other einfo: `entrez_db_searchable`; `entrez_db_summary`; `entrez_dbs`; `entrez_info`

Examples

```
taxid <- entrez_search(db="taxonomy", term="Osmeriformes")$ids
(tax_links <- entrez_db_links("taxonomy"))
tax_links[["pubmed"]]
entrez_link(dbfrom="taxonomy", db="pmc", id=taxid)

sra_links <- entrez_db_links("sra")
as.data.frame(sra_links)
```

---

entrez_db_searchable *Can be used in conjunction with* `entrez_search` *to find available search fields to include in the* `term` *argument of that function.*

Description

Can be used in conjunction with `entrez_search` to find available search fields to include in the `term` argument of that function.

Usage

```
entrez_db_searchable(db, config = NULL)
```

Arguments

db	character, name of database to get search field from
config	config vector passed to `httr::GET`

Value

An eInfoSearch object (subclassed from list) summarising linked-datbases. Can be coerced to a data-frame with `as.data.frame`. Printing the object shows only the names of each available search field.

See Also

entrez_search

Other einfo: entrez_db_links; entrez_db_summary; entrez_dbs; entrez_info

Examples

```
(pmc_fields <- entrez_db_searchable("pmc"))
pmc_fields[["AFFL"]]
entrez_search(db="pmc", term="Otago[AFFL]", retmax=0)
entrez_search(db="pmc", term="Auckland[AFFL]", retmax=0)

sra_fields <- entrez_db_searchable("sra")
as.data.frame(sra_fields)
```

---

entrez_db_summary *Retrieve summary information about an NCBI database*

Description

Retrieve summary information about an NCBI database

Usage

```
entrez_db_summary(db, config = NULL)
```

Arguments

db	character, name of database t
config	config vector passed to httr::GET

Value

Character vector with the following data

DbName Name of database

Description Brief description of the database

Count Number of records contained in the database

MenuName Name in web-interface to EUtils

DbBuild Unique ID for current build of database

LastUpdate Date of most recent update to database

See Also

Other einfo: entrez_db_links; entrez_db_searchable; entrez_dbs; entrez_info

Examples

```
entrez_db_summary("pubmed")
```

`entrez_fetch` *Download data from NCBI databases*

Description

Download data from NCBI databases

Usage

```
entrez_fetch(db, rettype, retmode = "text", config = NULL, ...)
```

Arguments

`db`	character Name of the database to use
`rettype`	character Format in which to get data (eg, fasta, xml...)
`retmode`	character Mode in which to receive data, defaults to 'text'
`config`	vector configuration options passed to httr::GET
`...`	character Additional terms to add to the request

Value

character string containing the file created

See Also

`config` for available configs

Examples

```
katipo <- "Latrodectus katipo[Organism]"
katipo_search <- entrez_search(db="nuccore", term=katipo)
kaitpo_seqs <- entrez_fetch(db="nuccore", id=katipo_search$ids, rettype="fasta")
```

entrez_global_query *See how many hits there are for a given term across all NCBI Entrez databases*

Description

See how many hits there are for a given term across all NCBI Entrez databases

Usage

```
entrez_global_query(term, config = NULL, ...)
```

Arguments

term	the search term to use
config	vector configuration options passed to httr::GET
...	additional arguments to add to the query

Value

a named vector with counts for each a database

See Also

config for available configs

Examples

```
NCBI_data_on_best_butterflies_ever <- entrez_global_query(term="Heliconius")
```

entrez_info *Get infromation about EUtils databases*

Description

Constructs a query to NCBI's einfo and returns a parsed XML object Note: The most common uses-cases for the einfo util are finding the list of search fields available for a given database or the other NCBI databases to which records in a given database might be linked. Both these use cases are implemented in higher-level functions that return just this information (entrez_db_searchable and entrez_db_links respectively). Consequently most users will not have a reason to use this function (though it is exported by rentrez for the sake of completeness.

Usage

```
entrez_info(db = NULL, config = NULL)
```

Arguments

db	characater database about which to retrieve information (optional,
config	config vector passed on to `httr::GET`

Value

XMLInternalDocument with information describing either all the databases available in Eutils (if db is not set) or one particular database (set by 'db')

See Also

`config` for available httr configurations

Other einfo: `entrez_db_links`; `entrez_db_searchable`; `entrez_db_summary`; `entrez_dbs`

Examples

```
all_the_data <- entrez_info()
xpathSApply(all_the_data, "//DbName", xmlValue)
entrez_dbs()
```

---

entrez_link *Get links to datasets related to a unique ID from an NCBI database*

Description

Constructs a query with the given arguments and downlands the XML document created by that query.

Usage

```
entrez_link(db, dbfrom, config = NULL, ...)
```

Arguments

db	character Name of the database to search for links (or use "all" to search all databases available for db. `entrez_db_links` allows you to discover databases that might have linked information (see examples).
dbfrom	character Name of database from which the Id(s) originate
config	vector configuration options passed to httr::GET
...	character Additional terms to add to the request

Value

An elink object containing vectors of unique IDs the vectors names take the form [db_from]_[db_to]

file XMLInternalDocument xml file resulting from search, parsed with `xmlTreeParse`

See Also

config for available configs

Examples

```
(pubmed_search <- entrez_search(db = "pubmed", term ="10.1016/j.ympev.2010.07.013[doi]"))
(linked_dbs <- entrez_db_links("pubmed"))
nucleotide_data <- entrez_link(dbfrom = "pubmed", id = pubmed_search$ids, db ="nuccore")
#All the links
entrez_link(dbfrom="pubmed", db="all", id=pubmed_search$ids)
```

---

entrez_post *Post IDs to Eutils for later use*

Description

Post IDs to Eutils for later use

Usage

```
entrez_post(db, id, config = NULL, ...)
```

Arguments

db	character Name of the database from which the IDs were taken
id	integer ID(s) for which data is being collected
config	vector configuration options passed to httr::GET
...	character Additional terms to add to the request

Value

QueryKey integer identifier for specific query in webhistory

WebEnv character identifier for session key to use with history

See Also

config for available configs

Examples

```
## Not run:
so_many_snails <- entrez_search(db="nuccore",
                      "Gastropoda[Organism] AND COI[Gene]", retmax=200)
upload <- entrez_post(db="nuccore", id=so_many_snails$ids)
cookie <- upload$WebEnv
first <- entrez_fetch(db="nuccore", file_format="fasta", WebEnv=cookie,
                      query_key=upload$QueryKey, retend=10)
second <- entrez_fetch(db="nuccore", file_format="fasta", WebEnv=cookie,
                       query_key=upload$QueryKey, retstart=10)

## End(Not run)
```

`entrez_search` *Search the NCBI databases using EUtils*

Description

Constructs a query with the given arguments, including a search term, and a database name, then retrieves the XML document created by that query.

Usage

```
entrez_search(db, term, config = NULL, retmode = "xml", ...)
```

Arguments

`db`	character Name of the database to search for
`term`	character The search term
`config`	vector configuration options passed to httr::GET
`retmode`	character One of json (default) or xml. This will make no difference in most cases.
`...`	character Additional terms to add to the request

Value

ids integer Unique IDS returned by the search

count integer Total number of hits for the search

retmax integer Maximum number of hits returned by the search

QueryKey integer identifier for specific query in webhistory

WebEnv character identifier for session key to use with history

file either and XMLInternalDocument xml file resulting from search, parsed with `xmlTreeParse` or, if `retmode` was set to json a list resulting from the returned JSON file being parsed with `fromJSON`.

See Also

config for available configs

entrez_db_searchable to get a set of search fields that can be userd in term for any base

Examples

```
## Not run:
   query <- "Gastropoda[Organism] AND COI[Gene]"
   web_env_search <- entrez_search(db="nuccore", query, usehistory="y")
   cookie <- web_env_search$WebEnv
   qk <- web_env_search$QueryKey
   snail_coi <- entrez_fetch(db = "nuccore", WebEnv = cookie, query_key = qk,
                             file_format = "fasta", retmax = 10)

## End(Not run)


fly_id <- entrez_search(db="taxonomy", term="Drosophila")
#Oh, right. There is a genus and a subgenus name Drosophila...
#how can we limit this seach
(tax_fields <- entrez_db_searchable("taxonomy"))
#"RANK" loots promising
tax_fields$RANK
entrez_search(db="taxonomy", term="Drosophila & Genus[RANK]")
```

---

entrez_summary

Get summaries of objects in NCBI datasets from a unique ID Constructs a query from the given arguments, including a database name and list of of unique IDs for that database.

Description

The NCBI offer two distinct formats for summary documents. Version 1.0 is a relatively limited summary of a database record based on a shared Document Type Definition. Version 1.0 summaries are only available as XML and are not available for some newer databases Version 2.0 summaries generally contain more information about a given record, but each database has its own distinct format. 2.0 summaries are available for records in all databases and as JSON and XML files. As of version 0.4, rentrez fetches version 2.0 summaries by default and uses JSON as the exchange format (as JSON object can be more easily converted into native R types). Existing scripts which relied on the structure and naming of the "Version 1.0" summary files can be updated by setting the new version argument to "1.0".

Usage

```
entrez_summary(db, version = c("2.0", "1.0"), config = NULL, ...)
```

Arguments

db	character Name of the database to search for
version	either 1.0 or 2.0 see above for description
config	vector configuration options passed to httr::GET
...	character Additional terms to add to the request. Requires either id (unique id(s) for records in a given database) or WebEnv (a character containing a cookie created by a previous entrez query).

Value

A list of esummary records (if multiple IDs are passed) or a single record.

file XMLInternalDocument xml file resulting from search, parsed with xmlTreeParse

See Also

config for available configs

Examples

```
pop_ids = c("307082412", "307075396", "307075338", "307075274")
 pop_summ <- entrez_summary(db="popset", id=pop_ids)
 sapply(pop_summ, "[[", "title")

 # clinvar example
 res <- entrez_search(db = "clinvar", term = "BRCA1", retmax=10)
 cv <- entrez_summary(db="clinvar", id=res$ids)
 cv[[1]] # get the names of the list for each result
 sapply(cv, "[[", "title") # titles
 lapply(cv, "[[", "trait_set")[1:2] # trait_set
 sapply(cv, "[[", "gene_sort") # gene_sort
```

---

parse_pubmed_xml *Summarise an XML record from pubmed.*

Description

Summarise an XML record from pubmed.

Usage

```
parse_pubmed_xml(raw_xml)
```

Arguments

raw_xml	character the record to be parsed (as a character, expected to come from entrez_fetch)

Value

Either a single pubmed_record object, or a list of several

Examples

```
hox_paper <- entrez_search(db="pubmed", term="10.1038/nature08789[doi]")
hox_rel <- entrez_link(db="pubmed", dbfrom="pubmed", id=hox_paper$ids)
recs <- entrez_fetch(db="pubmed",
                     id=hox_rel$pubmed_pubmed[1:3],
                     rettype="xml")
parse_pubmed_xml(recs)
```

`rentrez` *rentrez*

Description

rentrez provides functions to search for, discover and download data from the NCBI's databeses using their EUtils function.

Details

Users are expected to know a little bit about the EUtils API, which is well documented: http://www.ncbi.nlm.nih.gov/books/NBK25500/

The NCBI will ban IPs that don't use EUtils within their user guidelines. In particular /enumerated /item Don't sen more than three request per second (rentrez enforces this limit) /item If you plan on sending a sequence of more than ~100 requests, do so outside of peak times for the US /item For large requests use the web history method (see examples for `entrez_search` or use `entrez_post` to upload IDs)

Index

config, *2*, *6–9*, *11, 12*

entrez_citmatch, 2
entrez_db_links, *3*, 3, *5*, *8*
entrez_db_searchable, *3, 4*, 4, *5*, *8*, *11*
entrez_db_summary, *3–5*, 5, *8*
entrez_dbs, 3, *4, 5*, *8*
entrez_fetch, 6, *12*
entrez_global_query, 7
entrez_info, *3–5*, 7
entrez_link, *3, 4*, 8
entrez_post, 9, *13*
entrez_search, *4, 5*, 10, *13*
entrez_summary, 11

fromJSON, *10*

parse_pubmed_xml, 12

rentrez, 13
rentrez-package *(*rentrez*)*, 13

xmlTreeParse, *8*, *10*, *12*