

Package ‘OAIHarvester’

January 24, 2012

Version 0.1-3

Date 2011-02-22

Title Harvest Metadata Using OAI-PMH v2.0

Description Harvest metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 2.0.

Author Kurt Hornik

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

License GPL-2

Imports XML, RCurl

Repository CRAN

Date/Publication 2012-01-24 16:11:38

R topics documented:

harvest	2
transform	3
verb	4
Index	6

harvest

OAI-PMH Harvester

Description

Harvest a repository using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) requests.

Usage

```
oaih_harvest(baseurl,  
             prefix = "oai_dc", from = NULL, until = NULL, set = NULL,  
             transform = TRUE)
```

Arguments

<code>baseurl</code>	a character string giving the base URL of the repository.
<code>prefix</code>	a character vector with the formats in which metadata should be obtained, or NULL, indicating all available formats. The default ("oai_dc") corresponds to the mandatory OAI unqualified Dublin Core metadata schema.
<code>from, until</code>	character strings or Date or POSIXt date/time objects giving timestamps to be used as lower or upper bounds, respectively, for datestamp-based selective harvesting (i.e., only harvest only records with datestamps in the given range). If character, dates and times must be encoded using ISO 8601 in either '%F' or '%FT%TZ' format (see strptime). The trailing 'Z' must be used when including time. OAI-PMH implies UTC for data/time specifications.
<code>set</code>	a character vector giving the sets to be used for selective harvesting (i.e., only harvest records in the given sets), or NULL.
<code>transform</code>	a logical indicating whether the OAI-PMH XML results to "useful" R data structures via oaih_transform . Default: true.

Details

This is a high-level function for conveniently harvesting metadata from a repository, allowing specifying several metadata formats or sets. It also maps datestamps specified as R date or date/time objects to valid OAI-PMH datestamps according to the granularity of the repository.

Value

If the OAI-PMH request was successful, the result of the request as XML or (default) transformed to "useful" R data structures.

transform

Transform OAI-PMH XML Results

Description

Transform OAI-PMH XML results to “useful” R data structures (lists of character vectors or XML nodes) for further processing or analysis.

Usage

```
oaih_transform(x)
```

Arguments

x an XML node, or a list of character vectors or XML nodes.

Details

In a “list context”, i.e., if x conceptually contains information on several cases, transformation gives a “list matrix” (a list of character vector or XML node observations with a dim attribute) providing a rectangular case by variables data layout; otherwise, a list of variables. See the vignette for details.

Value

A list of character vectors or XML nodes, arranged as a matrix in the “list context”.

Examples

```
baseurl <- "http://epub.wu.ac.at/cgi/oai2"
## Get a single record to save bandwidth.
x <- oaih_get_record(baseurl,
                    "oai:epub.wu-wien.ac.at:852",
                    transform = FALSE)
## The result of the request is a single OAI-PMH XML <record> node:
x
## Transform this (turning identifier, datestamp and setSpec into
## character data):
x <- oaih_transform(x)
x
## This has its metadata in the default Dublin Core form, encoded in
## XML. Transform these to character data:
oaih_transform(x$metadata)
```

 verb

OAI-PMH Verb Functions

Description

Perform Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) requests for harvesting repositories.

Usage

```
oaih_get_record(baseurl, identifier, prefix = "oai_dc", transform = TRUE)
oaih_identify(baseurl, transform = TRUE)
oaih_list_identifiers(baseurl, prefix = "oai_dc", from = NULL,
                     until = NULL, set = NULL, transform = TRUE)
oaih_list_metadata_formats(baseurl, identifier = NULL, transform = TRUE)
oaih_list_records(baseurl, prefix = "oai_dc", from = NULL,
                 until = NULL, set = NULL, transform = TRUE)
oaih_list_sets(baseurl, transform = TRUE)
```

Arguments

baseurl	a character string giving the base URL of the repository.
identifier	a character string giving the unique identifier for an item in a repository.
prefix	a character string to specify the metadata format in OAI-PMH requests issued to the repository. The default ("oai_dc") corresponds to the mandatory OAI unqualified Dublin Core metadata schema.
from, until	character strings giving timestamps to be used as lower or upper bounds, respectively, for timestamp-based selective harvesting (i.e., only harvest only records with timestamps in the given range). Dates and times must be encoded using ISO 8601 in either '%F' or '%FT%TZ' format (see strptime). The trailing 'Z' must be used when including time. OAI-PMH implies UTC for data/time specifications.
set	a character string giving a set to be used for selective harvesting (i.e., only harvest records in the given set).
transform	a logical indicating whether the OAI-PMH XML results to "useful" R data structures via oaih_transform . Default: true.

Value

If the OAI-PMH request was successful, the result of the request as XML or (default) transformed to "useful" R data structures.

Examples

```
## Harvest ePubWU metadata.
baseurl <- "http://epub.wu.ac.at/cgi/oai2"
## Identify.
```

```
oaih_identify(baseurl)
## List metadata formats.
oaih_list_metadata_formats(baseurl)
## List sets.
sets <- oaih_list_sets(baseurl)
sets
## List records in the 'theses' set.
spec <- unlist(sets[sets[, "setName"] == "Type = Thesis", "setSpec"])
x <- oaih_list_records(baseurl, set = spec)
## Drop deleted records and extract the metadata.
m <- x[, "metadata"]
m <- oaih_transform(m[sapply(m, length) > 0L])
## Find the most frequent keywords.
sep <- "[[:space:]]*/[[:space:]]*"
keywords <- unlist(strsplit(unlist(m[, "subject"]), sep))
head(sort(table(keywords), decreasing = TRUE))
```

Index

Date, 2

harvest, 2

oaih_get_record (verb), 4

oaih_harvest (harvest), 2

oaih_identify (verb), 4

oaih_list_identifiers (verb), 4

oaih_list_metadata_formats (verb), 4

oaih_list_records (verb), 4

oaih_list_sets (verb), 4

oaih_transform, 2, 4

oaih_transform (transform), 3

POSIXt date/time objects, 2

strptime, 2, 4

transform, 3

verb, 4