

# Package ‘pubchunks’

June 28, 2020

**Title** Fetch Sections of XML Scholarly Articles

**Description** Get chunks of XML scholarly articles without having to know how to work with XML. Custom mappers for each publisher and for each article section pull out the information you want. Works with outputs from package ‘fulltext’, ‘xml2’ package documents, and file paths to XML documents.

**Version** 0.2.2

**License** MIT + file LICENSE

**URL** <https://docs.ropensci.org/pubchunks>,  
<https://github.com/ropensci/pubchunks>

**BugReports** <https://github.com/ropensci/pubchunks/issues>

**Encoding** UTF-8

**Language** en-US

**Imports** xml2 (>= 1.1.1), data.table, rcrossref

**Suggests** fulltext (>= 1.0.1), testthat

**RoxygenNote** 7.1.0

**X-schema.org-applicationCategory** Literature

**X-schema.org-keywords** text-ming, literature, pdf, xml, publications,  
citations, full-text, TDM

**X-schema.org-isPartOf** <https://ropensci.org>

**NeedsCompilation** no

**Author** Scott Chamberlain [aut, cre] (<<https://orcid.org/0000-0003-1444-9135>>),  
rOpenSci [fnd] (<https://ropensci.org>)

**Maintainer** Scott Chamberlain <[sckott@protonmail.com](mailto:sckott@protonmail.com)>

**Repository** CRAN

**Date/Publication** 2020-06-28 19:40:03 UTC

## R topics documented:

pubchunks-package . . . . .	2
pub_chunks . . . . .	2
pub_guess_publisher . . . . .	6
pub_providers . . . . .	7
pub_sections . . . . .	8
pub_tabularize . . . . .	8

<b>Index</b>	<b>10</b>
--------------	-----------

---

pubchunks-package	<i>Get chunks of XML articles</i>
-------------------	-----------------------------------

---

### Description

Pass XML in various forms, and get out specific sections of articles. It makes not knowing what XPath/CSS selectors are totally fine.

### Author(s)

Scott Chamberlain [myrmecocystus+r@gmail.com](mailto:myrmecocystus+r@gmail.com)

---

pub_chunks	<i>Extract chunks of data from articles</i>
------------	---------------------------------------------

---

### Description

pub\_chunks makes it easy to extract sections of an article. You can extract just authors across all articles, or all references sections, or the complete text of each article. Then you can pass the output downstream for visualization and analysis.

### Usage

```
pub_chunks(x, sections = "all", provider = NULL)
```

### Arguments

x	one of the following: <ul style="list-style-type: none"> <li>• file path for an XML file</li> <li>• a character string of XML, a list (of file paths, or XML in a character string, or xml_document objects)</li> <li>• or an object of class fulltext::ft_data, the output from a call to fulltext::ft_get()</li> </ul>
sections	(character) What elements to get, can be one or more in a vector or list. See <a href="#">pub_sections()</a> for options. optional. Default is to get all sections. See <a href="#">Details</a> .

provider (character) a single publisher name. see `pub_providers()` for options. required. If you select the wrong provider for the XML you have you may or may not get what you need :). By default this is NULL and we use `pub_guess_publisher()` to guess the publisher; we may get it wrong. You can override our guessing by passing in a name.

## Details

Options for the `sections` parameter:

- front - Publisher, journal and article metadata elements
- body - Body of the article
- back - Back of the article, acknowledgments, author contributions, references
- title - Article title
- doi - Article DOI
- categories - Publisher's categories, if any
- authors - Authors
- aff - Affiliation (includes author names)
- keywords - Keywords
- abstract - Article abstract
- executive\_summary - Article executive summary
- refs - References
- refs\_dois - References DOIs - if available
- publisher - Publisher name
- journal\_meta - Journal metadata
- article\_meta - Article metadata
- acknowledgments - Acknowledgments
- permissions - Article permissions
- history - Dates, recieved, published, accepted, etc.

## Value

A list, named by the section selected. sections not found or not in accepted list return NULL or zero length list. A ".publisher" list element gets attached to each list output, even when no data is found. When `fulltext::ft_get` output is passed in here, the list is named by the publisher, then within each publisher is a list of articles named by their identifiers (e.g. DOIs).

## Examples

```
# a file path to an XML file
x <- system.file("examples/elsevier_1.xml", package = "pubchunks")
pub_chunks(x, "title")
pub_chunks(x, "authors")
pub_chunks(x, "acknowledgments")
```

```
pub_chunks(x, "refs")
pub_chunks(x, c("title", "refs"))

# works the same with the xml already in a string
xml <- paste0(readLines(x), collapse = "")
pub_chunks(xml, "title")

# also works if you've already read in the XML (with xml2 pkg)
xml <- paste0(readLines(x), collapse = "")
xml <- xml2::read_xml(xml)
pub_chunks(xml, "title")

# Hindawi
x <- system.file("examples/hindawi_1.xml", package = "pubchunks")
pub_chunks(x, "abstract")
pub_chunks(x, "authors")
pub_chunks(x, "aff")
pub_chunks(x, "title")
pub_chunks(x, c("abstract", "title", "authors", "refs"))

# Pensoft
x <- system.file("examples/pensoft_1.xml", package = "pubchunks")
pub_chunks(x, "abstract")
pub_chunks(x, "aff")
pub_chunks(x, "title")
pub_chunks(x, c("abstract", "title", "authors", "refs"))

# Peerj
x <- system.file("examples/peerj_1.xml", package = "pubchunks")
pub_chunks(x, "abstract")
pub_chunks(x, "authors")
pub_chunks(x, "aff")
pub_chunks(x, "title")
pub_chunks(x, c("abstract", "title", "authors", "refs"))

# Frontiers
x <- system.file("examples/frontiers_1.xml", package = "pubchunks")
pub_chunks(x, "authors")
pub_chunks(x, "aff")
pub_chunks(x, c("doi", "abstract", "title", "authors", "refs", "abstract"))

# eLife
x <- system.file("examples/elife_1.xml", package = "pubchunks")
pub_chunks(x, "authors")
pub_chunks(x, "aff")
pub_chunks(x, c("doi", "title", "authors", "refs"))

# f1000research
x <- system.file("examples/f1000research_1.xml", package = "pubchunks")
pub_chunks(x, "title")
pub_chunks(x, "aff")
pub_chunks(x, c("doi", "title", "authors", "keywords", "refs"))
```

```
## Not run:
# Copernicus
x <- system.file("examples/copernicus_1.xml", package = "pubchunks")
pub_chunks(x, c("doi", "abstract", "title", "authors", "refs"))
pub_chunks(x, "aff")

# MDPI
x <- system.file("examples/mdpi_1.xml", package = "pubchunks")
x <- system.file("examples/mdpi_2.xml", package = "pubchunks")
pub_chunks(x, "title")
pub_chunks(x, "aff")
vv <- pub_chunks(x, c("doi", "title", "authors", "keywords", "refs",
  "abstract", "categories"))
vv$doi
vv$title
vv$authors
vv$keywords
vv$refs
vv$abstract
vv$categories

# Many inputs at once
x <- system.file("examples/frontiers_1.xml", package = "pubchunks")
y <- system.file("examples/elife_1.xml", package = "pubchunks")
z <- system.file("examples/f1000research_1.xml", package = "pubchunks")
pub_chunks(list(x, y, z), c("doi", "title", "authors", "refs"))

# non-XML files/content are xxx?
# pub_chunks('foo bar')

# Pubmed brief XML files (abstract only)
x <- system.file("examples/pubmed_brief_1.xml", package = "pubchunks")
pub_chunks(x, "title")

# Pubmed full XML files
x <- system.file("examples/pubmed_full_1.xml", package = "pubchunks")
pub_chunks(x, "title")

# using output of fulltext::ft_get()
if (requireNamespace("fulltext", quietly = TRUE)) {
  library("fulltext")

  # single
  x <- fulltext::ft_get('10.7554/eLife.03032')
  pub_chunks(fulltext::ft_collect(x), sections="authors")

  # many
  dois <- c('10.1371/journal.pone.0086169', '10.1371/journal.pone.0155491',
    '10.7554/eLife.03032')
  x <- fulltext::ft_get(dois)
  pub_chunks(fulltext::ft_collect(x), sections="authors")

  # as.ft_data() function
```

```
x <- ft_collect(as.ft_data())
names(x)
x$cached
pub_chunks(x, "title")
pub_chunks(x, "title") %>% pub_tabularize()
}

## End(Not run)
```

---

pub\_guess\_publisher    *Guess the publisher from an XML document*

---

## Description

Guess the publisher from an XML document

## Usage

```
pub_guess_publisher(x)
```

## Arguments

x                    an XML file, a character string of XML, or a xml\_document object (as from  
xml2::read\_xml)

## Value

a list, with two named character strings, one for full\_name and the other a short\_name

## Examples

```
(x <- system.file("examples/pensoft_1.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/copernicus_2.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/peerj_1.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/hindawi_1.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/frontiers_1.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/elife_1.xml", package = "pubchunks"))
pub_guess_publisher(x)
```

```
(x <- system.file("examples/elsevier_1.xml", package = "pubchunks"))
```

```
pub_guess_publisher(x)

x <- system.file("examples/f1000research_1.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/plos_1.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/mdpi_1.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/pubmed_brief_1.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/pubmed_full_1.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/pubmed_full_2.xml", package = "pubchunks")
pub_guess_publisher(x)

x <- system.file("examples/pubmed_full_3.xml", package = "pubchunks")
pub_guess_publisher(x)
```

---

pub\_providers

*Providers*

---

### **Description**

The possible providers to select from

### **Usage**

```
pub_providers()
```

### **Value**

character vector

### **Examples**

```
pub_providers()
```

---

pub_sections	<i>Sections</i>
--------------	-----------------

---

**Description**

The possible sections of an XML article that are supported for retrieval

**Usage**

```
pub_sections()
```

**Value**

character vector

**Examples**

```
pub_sections()
```

---

pub_tabularize	<i>Tabularize chunks output</i>
----------------	---------------------------------

---

**Description**

Tabularize chunks output

**Usage**

```
pub_tabularize(x, bind = FALSE)
```

**Arguments**

x	the output of <a href="#">pub_chunks()</a>
bind	(logical) whether to bind list of data.frames or not. ignored unless list input to x. default: FALSE

**Value**

a data.frame or list

**Examples**

```
# one at a time
## example 1, a file path
x <- system.file("examples/elife_1.xml", package = "pubchunks")
(res <- pub_chunks(x, c("doi", "title", "keywords")))
pub_tabularize(res)

## example 2, a file path
y <- system.file("examples/frontiers_1.xml", package = "pubchunks")
(res2 <- pub_chunks(y, c("doi", "title", "keywords")))
pub_tabularize(res2)

# > 1, a list of file paths
x <- system.file("examples/elife_1.xml", package = "pubchunks")
y <- system.file("examples/frontiers_1.xml", package = "pubchunks")
(res <- pub_chunks(list(x, y), c("doi", "title", "keywords")))
pub_tabularize(res)
pub_tabularize(res, bind = TRUE)

## Not run:
# using output of fulltext::ft_get()
if (requireNamespace("fulltext", quietly = TRUE)) {
  dois <- c('10.1371/journal.pone.0086169', '10.1371/journal.pone.0155491',
    '10.7554/eLife.03032')
  x <- fulltext::ft_get(dois)
  (tmp <- pub_chunks(fulltext::ft_collect(x), sections=c("doi","title")))
  pub_tabularize(tmp)
  pub_tabularize(tmp, bind = TRUE)
}
## End(Not run)
```

# Index

## \*Topic **package**

pubchunks-package, 2

pub\_chunks, 2

pub\_chunks(), 8

pub\_guess\_publisher, 6

pub\_guess\_publisher(), 3

pub\_providers, 7

pub\_providers(), 3

pub\_sections, 8

pub\_sections(), 2

pub\_tabularize, 8

pubchunks (pubchunks-package), 2

pubchunks-package, 2