# rOpenSci tools for textmining open source science literature

Scott Chamberlain[*,a]

[a] rOpenSci, Museum of Paleontology, University of California, Berkeley, CA, USA

## Abstract

Corresponding Author:

Scott Chamberlain

rOpenSci, Museum of Paleontology, University of California, Berkeley, CA, USA

Email address: scott@ropensci.org

[*] Corresponding author
*Email address:* `scott(at)ropensci.org` (Scott Chamberlain)

*September 13, 2017*

Background. xxxx.

Methods. xxxx.

Results. xxxx.

Discussion. xxxx.

## Introduction

There's likely more than 100 million articles published (source: Crossref API), representing an enormous amount of knowledge. In addition to simply reading these articles, they contain a vast trove of information of interest to researchers.

For example, many researchers are interested in statistical outcomes of articles: questions about P-values, about effect sizes, and more. With regard to effect sizes, these are of particular interest, as they are often combined in meta-analyses to draw broad conclusions about a particular question.

Text-mining is the broad term associated with pulling information out of articles. Given the importance of text-mining, good text-mining tools are needed to make it easier for researchers to do.

Here, we do an overview of text-mining tools in the R programming language. We do not cover analysis tools per se, but rather those tools for searching for, acquiring, and "mashing up" text.

## Digital articles: technical aspects

Of digital articles some of which are available digitally, and some of which are not. Those that are digital can be split into two groups: easily machine readable and non-machine readable.

The machine readable articles are those in XML, JSON, or plain text format. The former two, XML and JSON, are ideal of the machine readable types because they are structured data, whereas plain text has no structure - it's simply a long set of characters with line breaks and spaces in between.

Of the non-machine readable kind, there's PDFs. These can be broken out into two groups: text based PDFs and scanned PDFs. The former are converted from digital versions of various kinds (MS Word, OpenOffice, markdown, etc.), while the latter are PDFs created by scanning in print articles for which there is no digital version.

## Digital articles: the access landscape

Acces to full-text is the holy grail in text-mining. Some use cases can get by with article metadata (authors, title, etc.), some with abstracts, but many use cases need full-text.

The landscape of access to full-text is a extremely hetergeous, with the majority of variation along the publisher axis. The major hurdle are paywalls. The majority of articles are published by the big three publishers - Wiley, Springer, Elsever - and the majority of their articles are behind paywalls.

A promising sign is that there's an increasing number of open access publishers. xxxx.

## The discovery problem (maybe remove section)

xxx

## Data sources

There is increasing open source scientific literature content available online. However, only a small proportion of scientific journals provide access to their full content; whereas, most publishers provide open access to their metadata only (most often through Crossref; Table 1).

Table 1. Sources of scientific literature, their content type provided via web services, whether rOpenSci has an R packages for the service, and where to find the API documentation.

| Data Provider | Content Type | rOpenSci Pkg? | API Documentation |
|---|---|---|---|
| Crossref | Metadata only | rcrossref | [1] |
| DataCite | Metadata only | rdatacite | [2] |
| Biodiversity Heritage Library | Full content/Metadata | rbhl | [3] |
| Public Library of Science (PLoS) | Full text/altmetrics | rplos | [4] |
| Scopus (Elsevier) | Full content/Metadata | fulltext | [5] |
| arXiv | Full content/Metadata | aRxiv | [6] |
| Biomed Central (via Springer) | Full content/Metadata | fulltext | [7] |
| bioRxiv | Full content/Metadata | fulltext | [8] |
| PMC/Pubmed (via Entrez) | Full content/Metadata | rentrez | [9] |
| Microsoft Academic Search | Metadata | fulltext/microdemic | [10] |

[1] http://api.crossref.org

[2] https://support.datacite.org/docs/api

[3] http://bit.ly/KYQ1Rd

[4] http://api.plos.org/solr

[5] http://bit.ly/J9S616

[6] https://arxiv.org/help/api/index

[7] https://dev.springer.com/

[8] http://www.biorxiv.org/

[9] https://www.ncbi.nlm.nih.gov/books/NBK25500

[10] https://azure.microsoft.com/en-us/services/cognitive-services

The following is a synopsis of the major data sources and associated R tools.

### *Crossref*

Crossref is a non-profit that creates (or “mints”) Digital Object Identifiers (DOIs). In addition, they maintain metadata associated with each DOI. The metadata ranges from simple (including author, title, dates, DOI, type, publisher) to including number of citations to the article, as well as references in the article, and even abstracts.

Crossref does have a text-mining opt-in program for publishers. The result of this is that some publishers deposit URLs for full text content of their articles. The majority of these links are pay-walled, while some are open access. Using any of the various tools for working with Crossref data, you can filter your search to get only articles with full text links, and further to get only articles with full text links that are open access.

The main interface for Crossref in R is rcrossref. Parallel interfaces are available in Ruby (serrano) and Python (habanero).

### *Pubmed*

Pubmed is a corpus/website of NIH funded research . . .

## How to text mine from R: Three case studies

### *Case study 1*

### *Case study 2*

### *Case study 3*

## Conclusions and future directions

xxxx

## Acknowledgments

xxxx

## Data Accessibility

All scripts and data used in this paper can be found in the permanent data archive Zenodo under the digital object identifier (DOI). This DOI corresponds to a snapshot of the GitHub repository at https://github.com/ropensci/textmine. Software can be found at https://github.com/ropensci/xxx, xxxx, all under MIT licenses.

## References