

jstor: Import and Analyse Data from Scientific Texts

Thomas Klebel¹

¹ Department of Sociology, University of Graz

DOI: [10.21105/joss.00882](https://doi.org/10.21105/joss.00882)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Submitted: 07 August 2018

Published: 07 August 2018

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

The interest in text as data has seen a sharp increase in the past few years, mostly due to the advent of methods for automated text analysis. At the same time, researchers within the field of scientometrics have analysed citations and other aspects of the scholarly literature with great sophistication. The archival content of [JSTOR](#) offers a rich and diverse set of primary sources like research articles or book chapters for both approaches. [Data for Research \(DfR\)](#) by JSTOR gives all researchers, regardless of whether they have access to JSTOR or not, the opportunity to analyse metadata, n-grams and, upon special request, full-text materials about all available articles and books from JSTOR. The package `jstor` (Klebel 2018) helps in analysing these datasets by enabling researchers to easily import the metadata to R (R Core Team 2018), a task, for which no other integrated solution exists to date.

The metadata from DfR can either be analysed on their own or be used in conjunction with n-grams or full-text data. Commonly, metadata from DfR include information on the articles' authors, their title, journal, date of publishing, and quite frequently all footnotes and references. All this information can be of interest for specific research questions. For the analysis of n-grams or full-texts, the metadata imported with `jstor` allow the researchers to filter articles based on specific journals, the dates of publication, the authors, keywords in titles and other aspects.

`jstor` provides functions for three main tasks within the research process:

- Importing different parts of metadata, either from XML-files or directly from the .zip-archive provided by DfR.
- Importing n-gram and full-text files.
- Performing common tasks of cleaning metadata like unifying the journal id or cleaning page numbers.

Full documentation of `jstor`, including a comprehensive case study about analysing n-grams from DfR, is available at <https://ropensci.github.io/jstor/>. The package can be obtained from CRAN (<https://CRAN.R-project.org/package=jstor>) or from GitHub (<https://github.com/ropensci/jstor>). Archived versions of all releases are available at Zenodo (<https://doi.org/10.5281/zenodo.1169861>).

Acknowledgements

I am indebted to Jason Becker and Elin Waring for their helpful requests during [package review](#). Benjamin Klebel, Antonia Schirgi and Matthias Duller provided helpful comments and requests for clarifications when writing the case study.

Work on `jstor` benefited from financial support for the project “Academic Super-Elites in Sociology and Economics” by the Austrian Science Fund (FWF), project number “P 29211 Einzelprojekte”. `jstor` is currently being used by the project team to analyse academic elites in sociology and economics.

References

Klebel, Thomas. 2018. *Jstor: Read Data from Jstor/Dfr*. <https://ropensci.github.io/jstor/>.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.