Using `tidy_bib`

Paul C. Bauer* Denis Cohen†

2020-06-25

Problem

The package/function deals with two problems:

1. We normally use (one or several) large .bib files as input for our paper .rmd files. These are based on one or several authors' literature databases. However, we ideally end up with one single .bib file that includes only those references that were cited in our paper, e.g., if we prepare reproduction files.
2. Papers normally have at least two sections namely a main section and an appendix that also have independent bibliographies. Ideally, in compiling our RMarkdown paper two independent bibliographies that correspond to the section and are located at their end would be automatically generated.

The idea is to provide functions that automatically solve these problems for us.

An example

In this example, we use the `tidy_bib()` function to

1. combine two separate `.bib`-files, `references1.bib` and `references2.bib`;
2. clean and repair the resulting bib-files (e.g., by removing unwanted fields such as the ISSN, ISBN, DOI, and URL);
3. freely cite works from the combined `.bib`-files in an RMarkdown document which consists of a main text and an appendix;
4. create separate `.bib`-files for the main text and the appendix, each containing only those entries which were cited in the respective sections;
5. embed the corresponding separate bibliographies for the main text and the appendix.

*University of Mannheim, paul.bauer@mzes.uni-mannheim.de
†University of Mannheim, denis.cohen@mzes.uni-mannheim.de

YAML Header

In order for this to work, we must specify some arguments in the YAML header of our `.Rmd` file:

```
output:
  bookdown::pdf_document2:
    number_sections: no
    toc: false
    pandoc_args: --lua-filter=multiple-bibliographies.lua
    keep_tex: no

bibliography_main:
  - partial_bib_1.bib
bibliography_app:
  - partial_bib_2.bib
link-citations: yes
linkcolor: blue
```

These are the important additions:

1. The argument `pandoc_args: --lua-filter=multiple-bibliographies.lua` calls the lua-filter `multiple-bibliographies` for the creation and inclusion of multiple bibliographies using `pandoc-citeproc`.
2. The definitions of `bibliography_main` and `bibliography_app` specify the suffixes of our partial bibliographies as well as the file names of the corresponding `.bib`files, which will be produced by `tidy_bib()` in the next step.

`tidy_bib()` code chunk

After the YAML header, make sure to include a (hidden, yet evaluating) R code chunk in your RMarkdown document. Here, we specify the following:

```
tidy_bib(
  rmarkdown_file = "manuscript.Rmd",
  bib_input = c("references1.bib", "references2.bib"),
  bib_output = "partial_bib.bib",
  by_sections = c("<!--- appendix split -->"),
  repair = TRUE,
  removeISSN = TRUE,
  removeISBN = TRUE,
  removeDOI = TRUE,
  removeURL = TRUE
)
```

- `rmarkdown_file = "manuscript.Rmd"` specifies that the very same RMarkdown script in which we are writing our paper will be scanned for citations.

- `bib_input = c("references1.bib", "references2.bib")` means that we supply two larger bib files which will be combined and cleaned before the entries matching the citations in `manuscript.Rmd` will be extracted.
- `bib_output = "partial_bib.bib"` specifies the name of the newly created (partial) `.bib` files. If we request multiple separate `.bib` files for different sections of the document, these will by default be saved with numerical suffixes before the file extensions, e.g. `partial_bib_1.bib`, `partial_bib_2.bib`, etc. *Note that we must already supply matching file names for these bibliographies in our YAML header.*
- `by_sections = c("<!--- appendix split -->")` defines the split point included in our `.Rmd` script. Here, we only supply one split point, which means that `tidy_bib()` will extract citations separately before and after the split point (and thus produce two separate `.bib` files). Note that we can easily add more split points, e.g., by including distinct comments in the `.Rmd` script and adding them to the input vector for the `by_sections` argument.
- `repair`, `removeISSN`, `removeISBN`, `removeDOI`, and `remove URL` are additional options that define how we want to tidy up our new `.bib` file(s).

Document body

The following is the text body of the script `manuscript.Rmd`:

```
## Main Text

This is the main text. It cites one paper by @Athey2019-fy inline.
We also cite a book by @Berelsonetal1954.

Let's also cite a report in parentheses [@Gallup2019].
Here is also some more intricate stuff with prefixes and suffixes
[e.g., @Hargittai2008-fa, pp. 31-57].

## Main Text References
<div id="refs_main"></div>

<!-- appendix split -->
## Appendix

This is the appendix. It cites some other paper [@Friedman2009-gx].

## Appendix References
<div id="refs_app"></div>
```

In the document body, we can then freely include citations as we usually would. When it comes to printing our bibliographies, there are two things that we handle slightly differently from the default way of including bibliographies in Markdown.

1. `<!--- tidy_bib appendix split -->`, supplied as a Markdown comment, defines the split point.
2. We override the default of printing bibliographies at the end of the document by adding `<div id="refs_main"></div>` and `<div id="refs_app"></div>`, respectively. This ensure that the two bibliographies will be printed where we want them to be printed. Note that the names of the arguments must match those defined in the YAML header, just like the file names of the corresponding `.bib` files must match those of the new `.bib` files produced by `tidy_bib`.

On the following page, you can see the knitted PDF output generated from `manuscript.Rmd`.

Main Text

This is the main text. It cites one paper by Athey, Tibshirani, and Wager (2019) inline. We also cite a book by Berelson, Lazarsfeld, and McPhee (1954).

Let's also cite a report in parantheses (Gallup 2019). Here is also some more intricate stuff with prefixes and suffixes (e.g., Hargittai and Walejko 2008, 31–57).

Main Text References

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Ann. Stat.* 47 (2): 1148–78.

Berelson, Bernard R., Paul Felix Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign.* Chicago: University of Chicago Press.

Gallup. 2019. "Gallup Poll Social Series: Governance."

Hargittai, Eszter, and Gina Walejko. 2008. "The Participation Divide: Content Creation and Sharing in." *Inf. Commun. Soc.* 11 (2): 239–56.

Appendix

This is the appendix. It cites some other paper (Friedman, Hastie, and Tibshirani 2009).

Appendix References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning.* Vol. 1. New York: Springer series in statistic.

FAQ: Variations in using `tidy_bib`

What if I have 1 (or many) input `.bib` files?

For a single `.bib` files, simply specify `bib_input = "path/to/my/bibfile.bib"`. For any number of `.bib` files, collected the path names and/or file names in a vector, e.g. `bib_input = c("my/first/bibfile1.bib", "my/other/bibfile2.bib", "my/last/bibfile3.bib")`.

In a document with many sections/chapters, how can I include separate bibliographies for each of them?

Embedding many chapter or section-specific bibliographies is easy, as long as you keep track of

1. the anchors you set for the split points in your `.Rmd` file;
2. the enumerated `bib_output` files that `tidy_bib()` will produce;
3. the specification of your partial bibliographies in your YAML header;
4. and the placement of the partial bibliographies in your `.Rmd file`

Here is the rough structure of an `.Rmd` file for a thesis with five chapters with chapter-specific bibliographies. Note that we only need to specify four anchors to obtain five separate bibliographies:

```
---
title: "My Thesis"
author: "Me"

output:
  bookdown::pdf_document2:
    fig_caption: yes
    number_sections: no
    toc: false
    pandoc_args: --lua-filter=multiple-bibliographies.lua
    keep_tex: no

bibliography_intro:
  - bib_for_chapter_1.bib
bibliography_theory:
  - bib_for_chapter_2.bib
bibliography_methods:
  - bib_for_chapter_3.bib
bibliography_results:
  - bib_for_chapter_4.bib
bibliography_conclusion:
  - bib_for_chapter_5.bib
---
```

```
tidy_bib(
  rmarkdown_file = "my_thesis.Rmd",
  bib_input = "some/folder/my_bib_bibfile.bib",
  bib_output = "bib_for_chapter.bib",
  by_sections = c(
    "<!--- bib-anchor intro -->",
    "<!--- bib-anchor theory -->",
    "<!--- bib-anchor methods -->",
    "<!--- bib-anchor results -->"),
  repair = TRUE,
  removeISSN = TRUE,
  removeISBN = TRUE,
  removeDOI = TRUE,
  removeURL = TRUE
)
```

```
# Introduction

This is my introduction.

<!--- bib-anchor intro -->
## References
<div id="refs_intro"></div>

# Theory

This is my theory chapter.

<!--- bib-anchor theory -->
## References
<div id="refs_theory"></div>

# Methods

This is where I explicate my methods.

<!--- bib-anchor methods -->
## References
<div id="refs_methods"></div>

# Results

These are my really cool findings.

<!--- bib-anchor results -->
## References
<div id="refs_results"></div>
```

```
# Conclusion

This is my conclusion.

## References
<div id="refs_conclusion"></div>
```