Getting Started Creating Data Dictionaries: How to Create a Shareable Dataset

Erin M. Buchanan12abc

Sarah E. Crain1abc

Arielle Cunningham1abc

Hannah R. Johnson1abc

Hannah Stash1abc

1Missouri State University

2Harrisburg University of Science and Technology

aDrafted the manuscript

bContributed to the application development

cCommented or edited on applications or manuscript

Corresponding author: Erin M. Buchanan ([erinbuchanan@missouristate.edu](mailto:erinbuchanan@missouristate.edu) may use [ebuchanan@harrisburgu.edu](mailto:ebuchanan@harrisburgu.edu))

THIS PAGE WILL BE DELETED LATER

Tutorial  
Tutorials provide hands-on, practical guidance for researchers. Any topic that could enhance research practices or methods for researchers might be suitable as a tutorial. For example, AMPPS welcomes tutorials that focus on helping researchers learn to use statistical tools, improve their statistical practices and intuitions, better their data management and lab practices, enhance the reliability and reproducibility of their research, or facilitate transparent and open practices. Tutorials typically include dynamic, interactive content and should provide concrete guidance rather than abstract principles. Some tutorials will be solicited by the editorial team, and the team welcomes both suggestions for tutorial topics and proposals for tutorials, which can be submitted by emailing the editors a short, 1-2 paragraph summary.

• Most tutorials should be brief (<3,000 words), but they may be longer if necessary to explain the content fully and make it accessible and usable to readers.  
• Introductions typically should be no more than one to two paragraphs long (<500 words) and should not include extensive literature reviews. The introduction should explain the need for the tutorial and highlight how learning the contents will benefit readers.  
• Rather than a General Discussion section, Tutorials should have a one-paragraph summary of their contents.  
• Tutorials should be accompanied by publicly available code and all resources necessary for researchers (and reviewers) to follow them.  
• Tutorials should include a list of additional resources for readers who would like to learn more. The list can include links to online sources as well as citations to other articles.

Getting Started Creating Data Dictionaries: How to Create A Shareable Dataset

The Open Science Framework, designed by the Center for Open Science, was created to aid scientists by creating an online platform in which they could openly record, report, and share data, thus, demystifying the research process from beginning hypotheses to final journal publication [(Nelson, Simmons, & Simonsohn, 2018; Nosek et al., 2015)](https://paperpile.com/c/M0ehP6/JmdO+FuOm). This tutorial focuses on open data which is beneficial for both individual researchers and science, as it facilitates the spread of knowledge and improvements in research. Open data allows scientists to develop new hypotheses, see multiple perspectives on different research, and identify errors (Piwowar & Vision, 2013). In fields where *p*-hacking, false positives, and questionable research practices are a concern, such as psychology, open data discourages fraud and makes replication more likely [(Piwowar & Vision, 2013; Simmons, Nelson, & Simonsohn, 2011)](https://paperpile.com/c/M0ehP6/SoCY+jU3P). However, open data must be clear and understandable to successfully replicate and be further explored, which is a prerequisite for verifiable research [(Stodden, 2011)](https://paperpile.com/c/M0ehP6/Zovq). While one may suggest that open data is not necessary because scientists can share upon request, it appears that only a small percentage of those scientists end up sharing their data [(Houtkoop et al., 2018; Rouder, 2016; Wicherts, Borsboom, Kats, & Molenaar, 2006)](https://paperpile.com/c/M0ehP6/S4yU+yG93+Fwut).

Practically, creating open data requires the sharing researcher to pursue the avenues to share it, and one of the most common concerns is not having set standards for making data public [(Hardwicke et al., 2018; Houtkoop et al., 2018)](https://paperpile.com/c/M0ehP6/Fwut+ao31). For open data to be useful, information about the contents of data, often called metadata, should also be shared. Metadata includes items such as descriptions of the variables collected, citations, and data collection information. While there is no current standard for the structure of metadata, this tutorial will demonstrate three simple options for creating metadata documents. These applications allow the researcher to enter their metadata and share the files on a platform, such as OSF or Github, for others to read. Table 1 indicates the practical benefits to each of the applications described below.

Table 1.

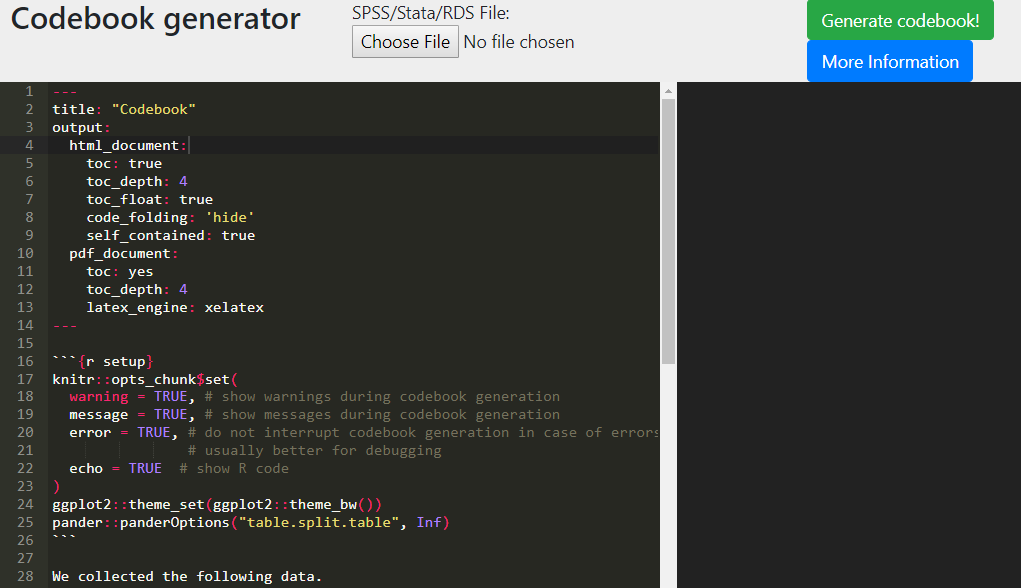
*Information about the Benefits for each Application*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Codebook | DataSchema | DD Creator |
| Citation | Arslan [(2018)](https://paperpile.com/c/M0ehP6/eQCc/?noauthor=1) | Inspired by Data Spice [(Boettiger et al., 2018)](https://paperpile.com/c/M0ehP6/1Sff) | [(DeBruine, Buchanan, & Mohr, 2018)](https://paperpile.com/c/M0ehP6/dFMJ) [(2018)](https://paperpile.com/c/M0ehP6/dFMJ/?noauthor=1) |
| Input | CSV, SPSS, Stata, RDS | CSV, Text, Excel, SPSS, SAS | CSV, Text, Excel, SPSS, SAS |
| Output | HTML report from Markdown | CSV files of meta-data, JSON, and HTML report | CSV files of meta-data, JSON, Rdata |
| Benefits | Easiest to use  Quick metadata generation  Generates a summary for each variable in a readable format | Follows schema.org for output  Metadata entry is medium | Specifies a separate section for category labels  Rdata output  More detailed descriptions, depending on the data |

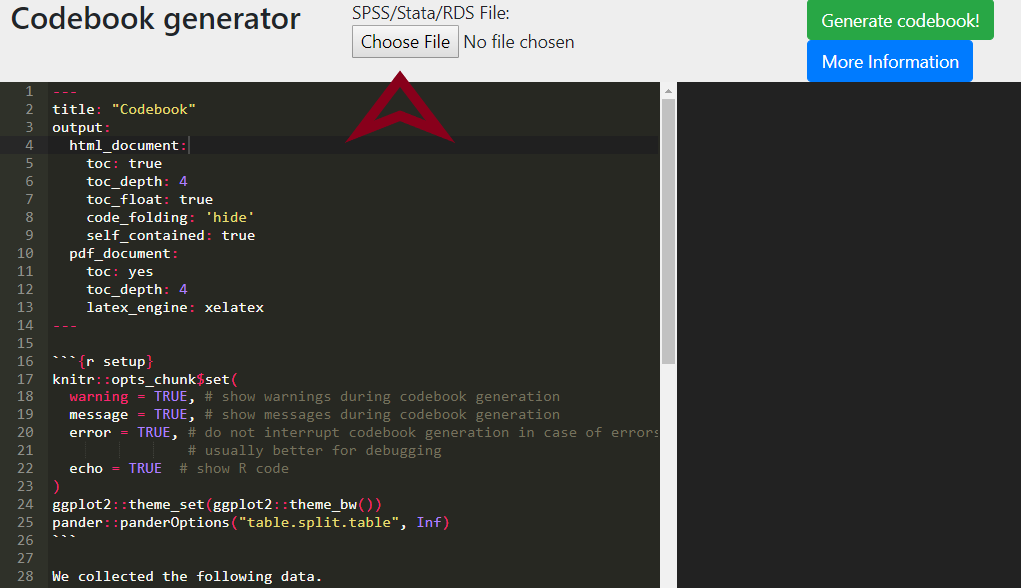
**Codebook Tutorial**

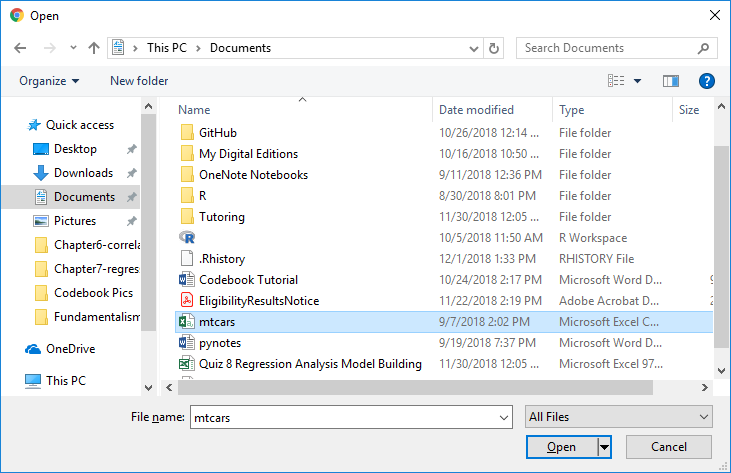
Codebook [(Arslan, 2018)](https://paperpile.com/c/M0ehP6/eQCc) is an *R* package with a corresponding website that allows researchers to create reports of their data, including reliabilities, summaries of items (histograms, descriptive statistics). The embedded metadata (such as item labels) is automatically combined into that report. Of the three available options, codebook is the quickest and easiest to implement; however, non computer savvy users would have trouble editing the automatically produced output if they wished to add more information. The tutorial below covers the web app version of codebook, and Arslan (2018) provides more information on using codebook in *R*.

Step 1: Go to the codebook website: <https://opencpu.formr.org/ocpu/library/codebook/www/>.

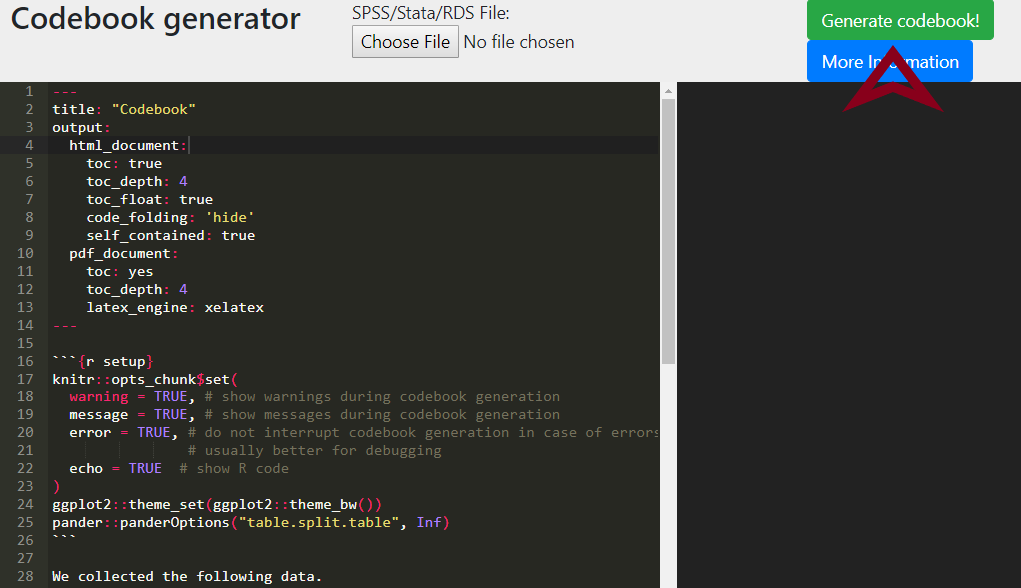


Step 2: Click the **Browse...** button at the top of the page. Select the data file you wish to upload, then click **Open**. You may upload CSV, SPSS, Stata, or RDS files.

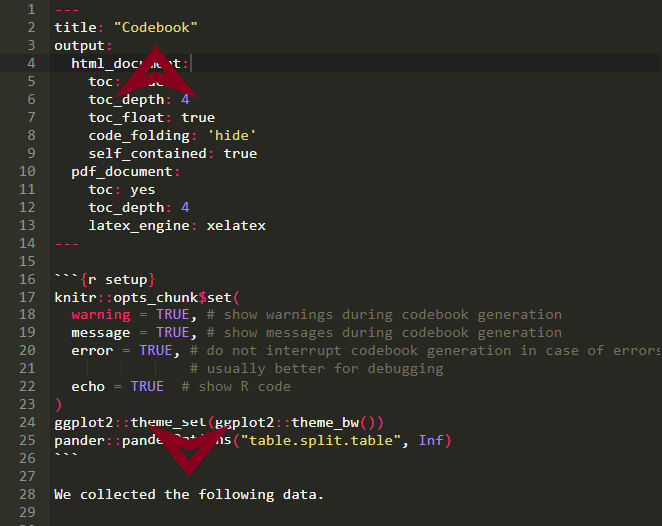




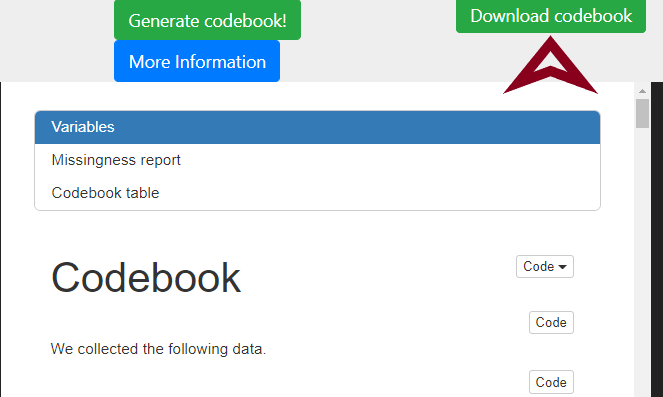
Step 3: Click **Generate codebook!**

****

Step 4: Make changes to the created document by editing the code on the left side of the screen. You can safely edit the title or the text that says, “We collected the following data.” Those more familiar with R Markdown may edit elsewhere. To save changes, click **Generate codebook!** again.



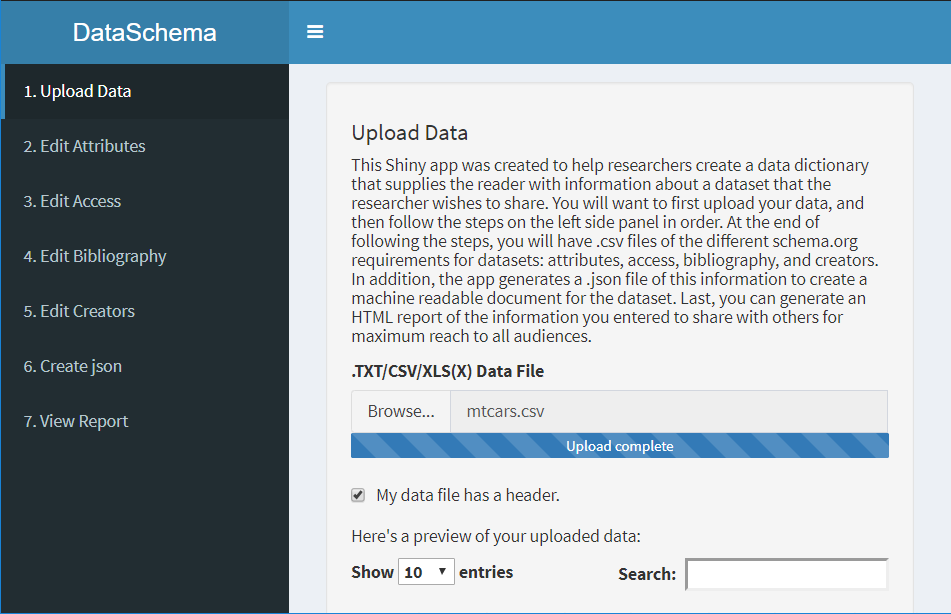
Step 5: When ready, click **Download codebook** to save the file to your computer.

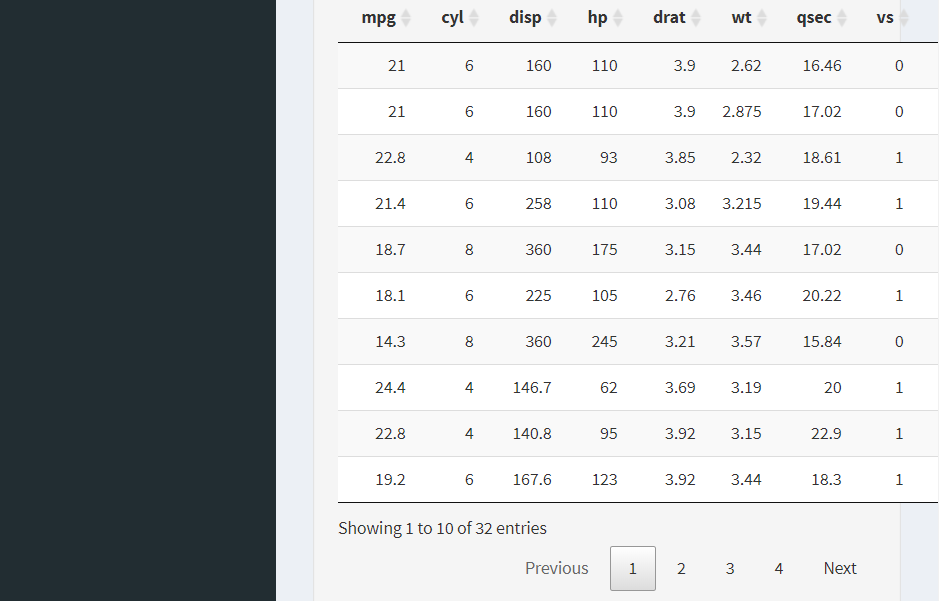


**DataSchema Tutorial**

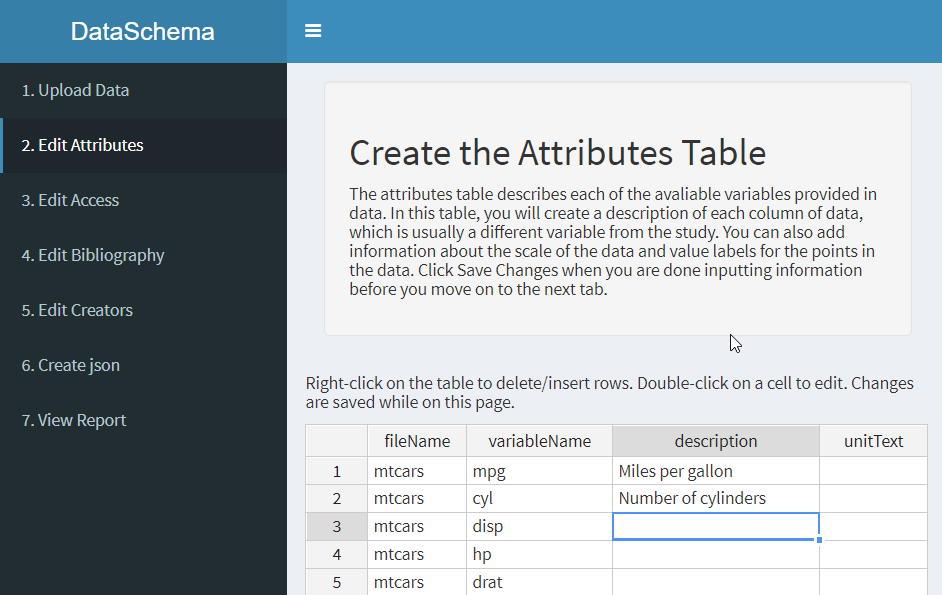
DataSchema was created by forking (copying) from the GitHub repository for Data Spice [(Boettiger et al., 2018)](https://paperpile.com/c/M0ehP6/1Sff). The original *R* code was then rewritten completely as a single Shiny application [(Chang, Cheng, Allaire, Xie, & McPherson, 2018)](https://paperpile.com/c/M0ehP6/tu60) by the lead author with input from the research team on instructions and creating an easy-to-use application. This application allows users to create an HTML report, a JSON file formatted following guidelines for datasets from schema.org, and .csv files of their metadata. JSON files are machine readable formats, which are encouraged for sharing, especially by the new Google Dataset Search [(Noy, 2018)](https://paperpile.com/c/M0ehP6/Ll5T). In this application, descriptions of the dataset properties (e.g., authors, collection dates) and column information should be entered to complete the metadata files.

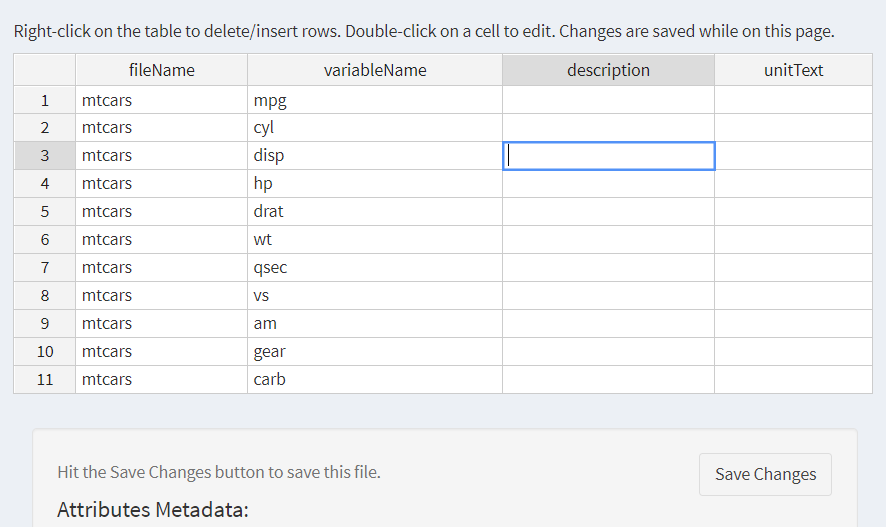
Step 1: Go to <https://doomlab.shinyapps.io/dataschema/> to access DataSchema.

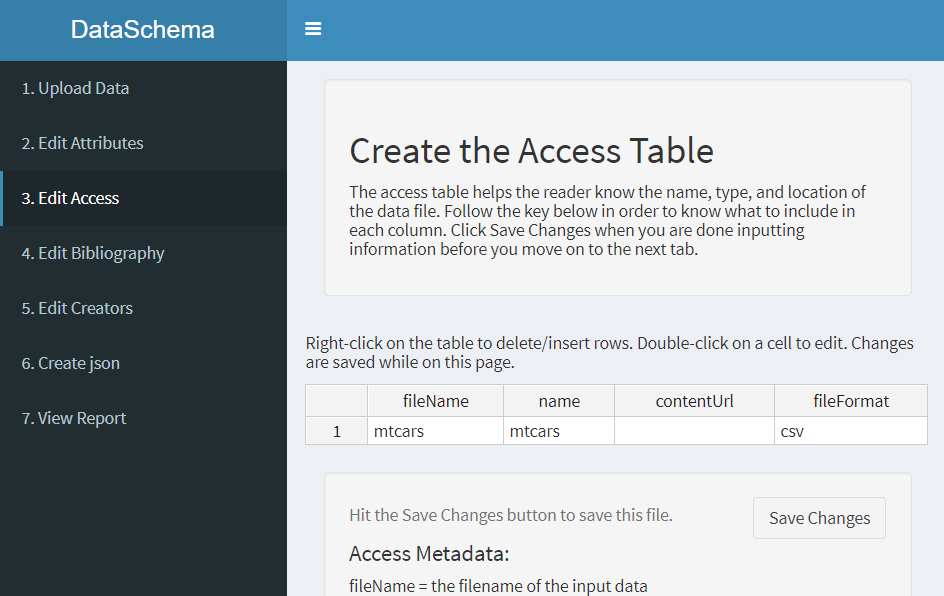


Step 2: Click **Browse...** and select the data file you wish to upload. You may upload a TXT, CSV, XLS(X), SAV, or SAS data file. The table will automatically display your uploaded data.

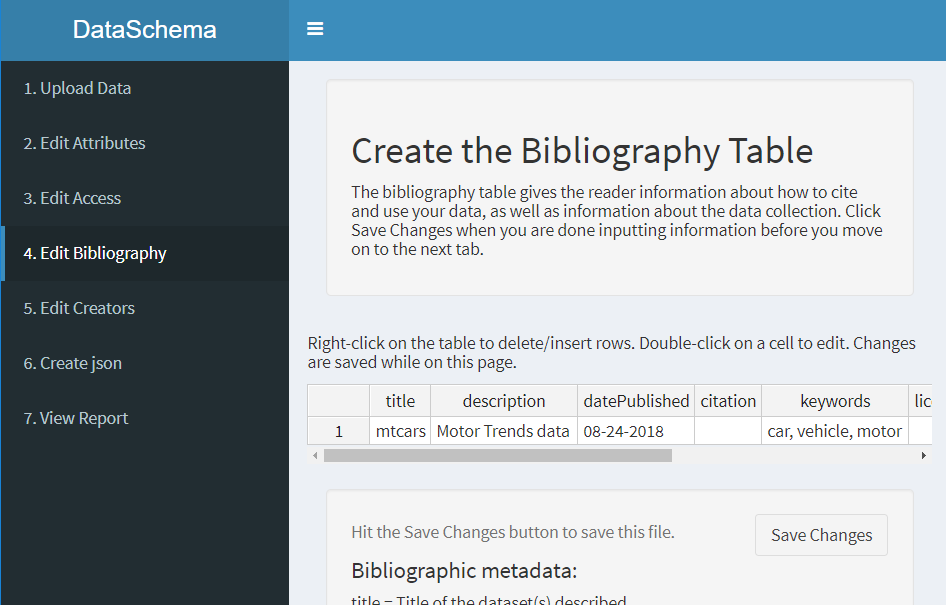
Step 3: Click the **Edit Attributes** page on the left to add description to the variables provided from the data file. The data uploaded will be copied into the table on this page.



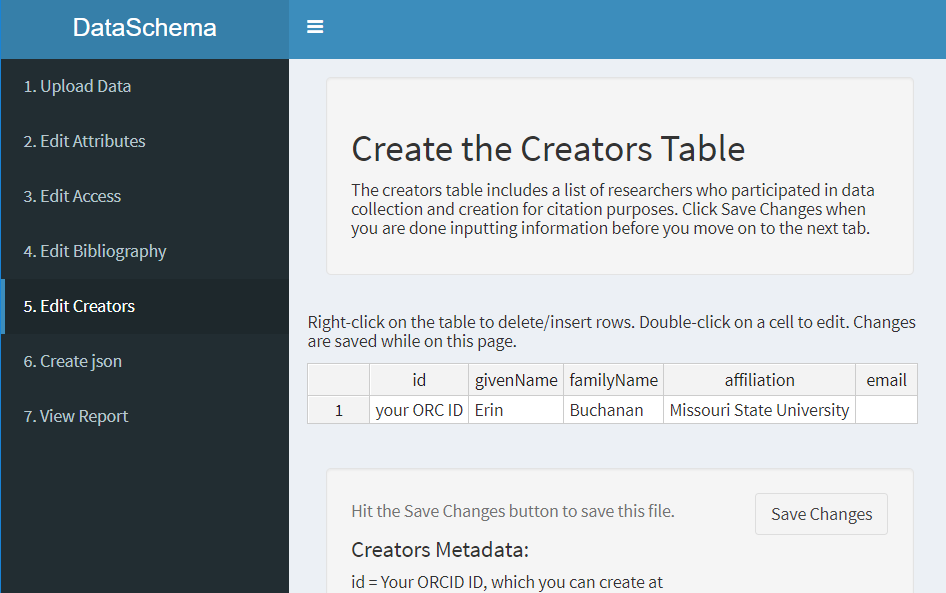
Edit the table by selecting a cell and typing directly into it. Make sure to click **Save Changes** to save the table before moving onto the next step. Step 4: Go to the **Edit Access** page to include the name, type, and location of the data file. Edit the table by directly typing in the desired box, and then click **Save Changes** before moving on to the next step.

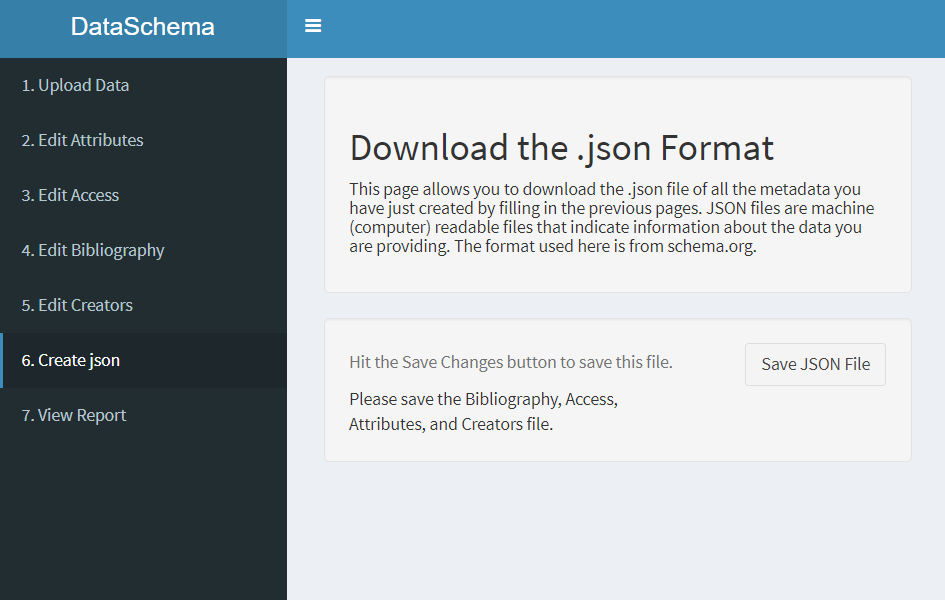


Step 5: Go to the **Edit Bibliography** page to add information about the data collection and data citation. Edit the table by clicking directly on the box, and make sure to click **Save Changes** before moving on to the next step.

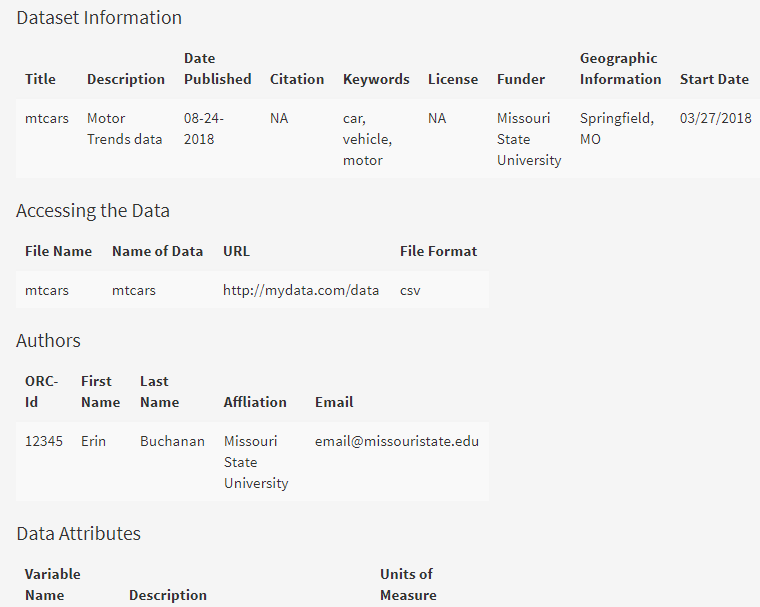


Step 6: Go to the **Edit Creators** page to include citation information about the researchers who were involved in data collection and creation. Edit the table by typing directly in it, and click **Save Changes** before moving on to the next step.



Step 7: Go to the **Create json** page to download the .json file of the metadata created by editing the tables from previous pages. Click **Save JSON File** to save the metadata as a .json file.

Step 8: Go to the **View Report** page to view the metadata as an HTML report. Click **Save Report HTML** to save the report as an HTML file. Open the file to view the HTML report of the metadata.

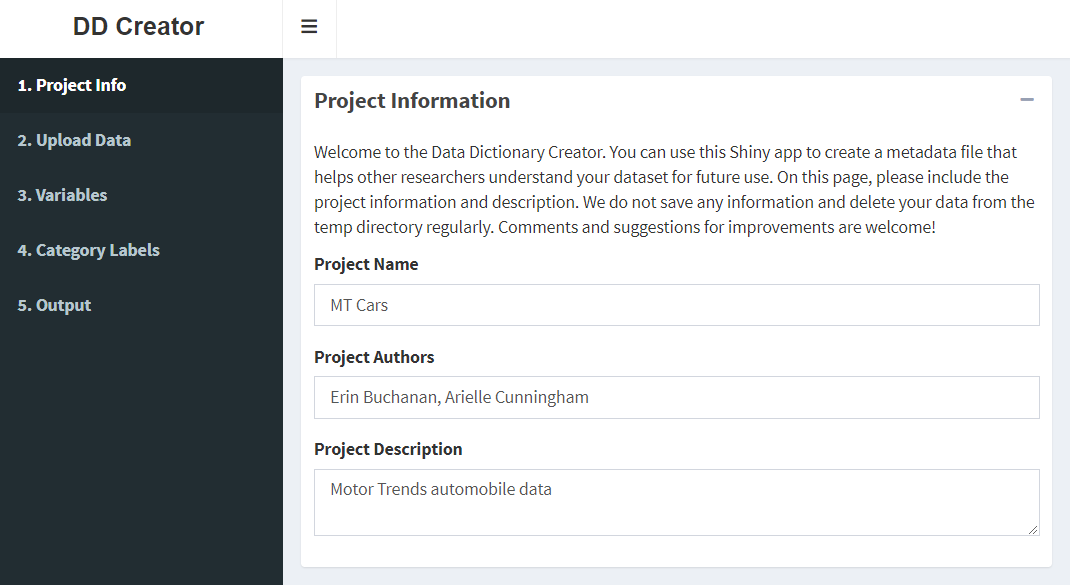


**Data Dictionary Creator**

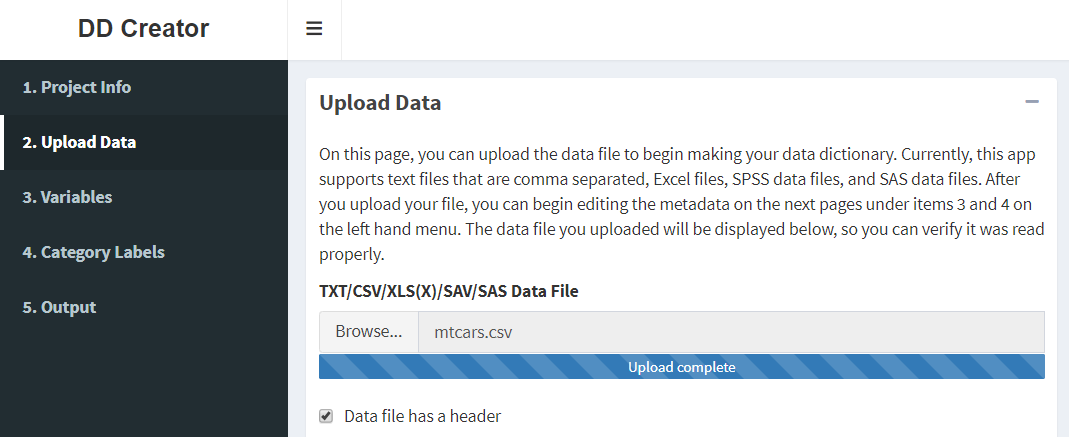
Data Dictionary Creator (DD Creator) was created as part of a Society for the Improvement of the Psychological Sciences Hackathon by Debruine, Buchanan, and Mohr [(2018)](https://paperpile.com/c/M0ehP6/dFMJ/?noauthor=1). The Shiny application was recently updated by the lead author, again with input from the research team for instructions and ease-of-use comments. DD Creator allows a user to enter metadata for each column provided in the dataset, while automatically providing a starting point for the number of unique values, missing values, variable type (i.e., character, numeric), and minimum/maximum values. A description of each column can be added, along with information about the levels/groups in the data and synonyms for the variables. On a separate page, category labels can be provided for both character and numeric data (i.e., Likert-type scales that include labeled numbers). Users can then download CSV files of the metadata, a JSON formatted metadata file, and an Rdata file that includes the dataset and descriptive information integrated together.

Step 1: Go to <https://doomlab.shinyapps.io/ddcreator/> to access DD Creator.

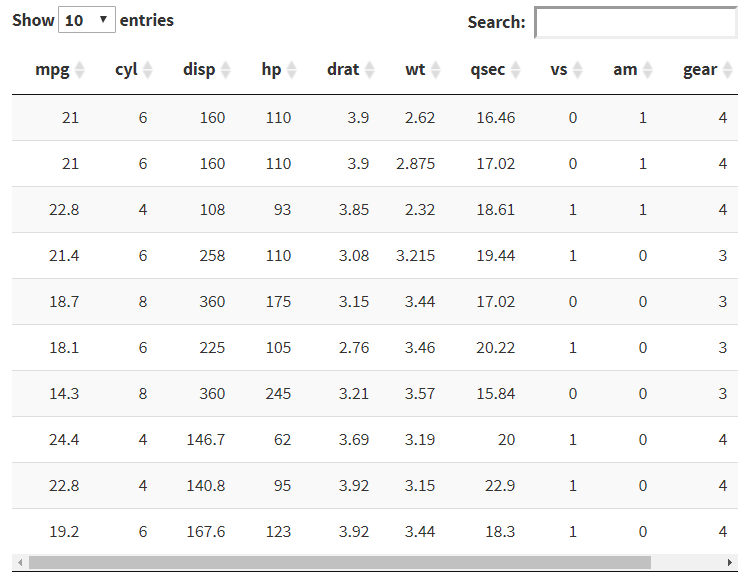
Step 2: On the first page (**Project Info**), type in the appropriate information under the **Project Name**, **Project Authors**, and **Project Description** fields.



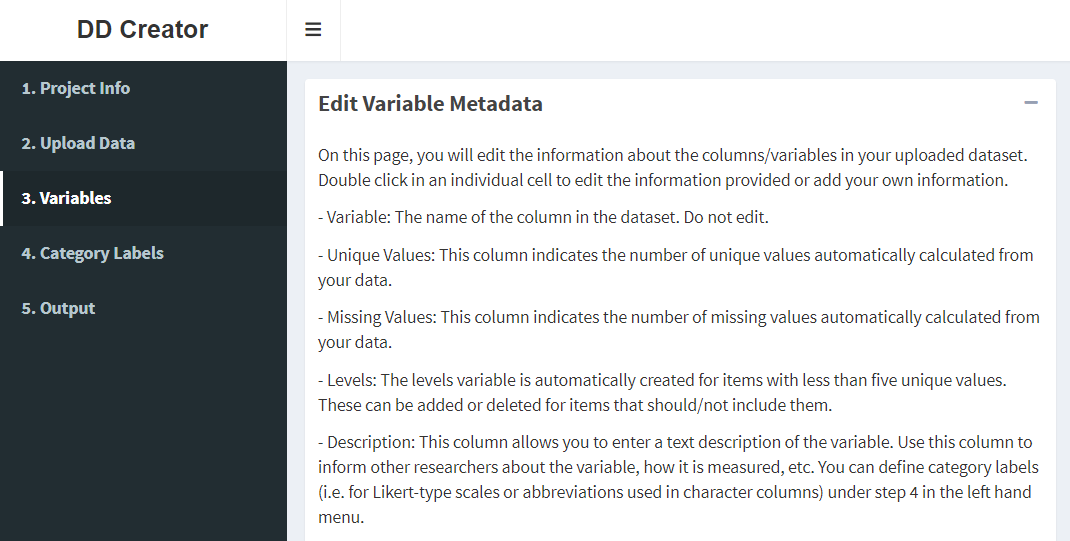
Step 3. Click the **Upload Data** tab on the left side of the page. Click **Browse** and select the data file you would like to upload. If you data file has a header, check the **Data file has a header** box.



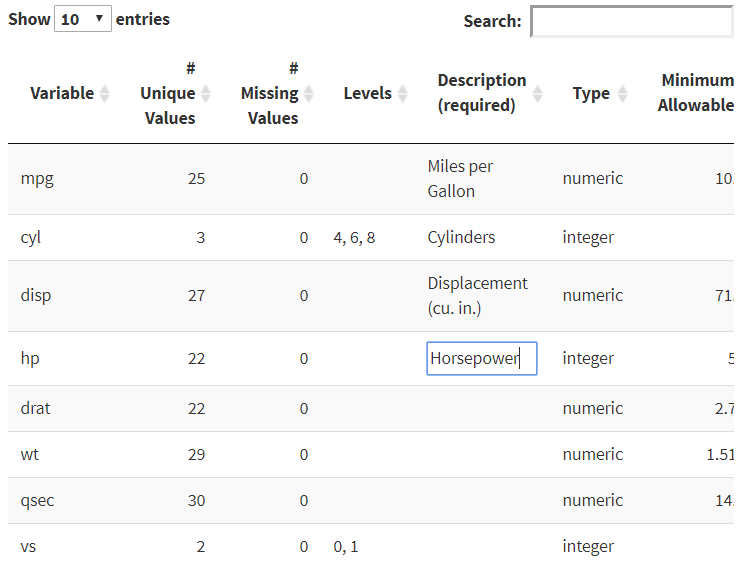
A table of your data will display below. You can change the number of entries you see on the screen as well as search for specific information in the dataset.



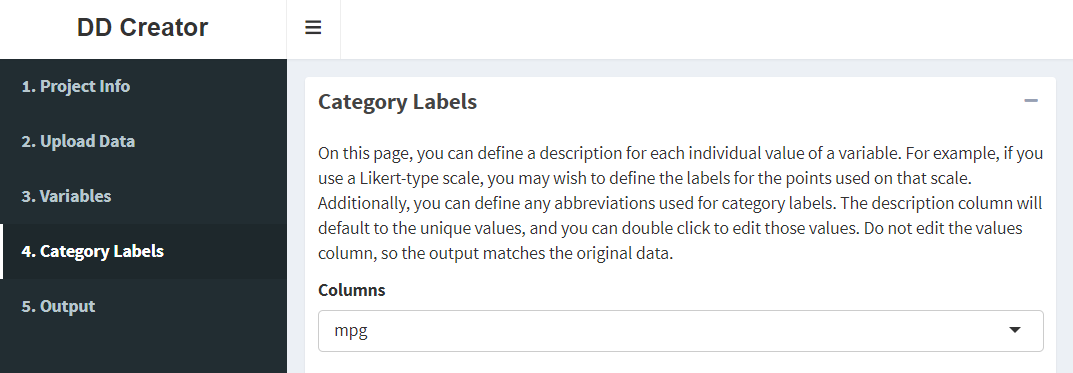
Step 4: Click the **Variables** tab on the left side of the page. The metadata table will populate with information from the dataset you uploaded in the previous step.



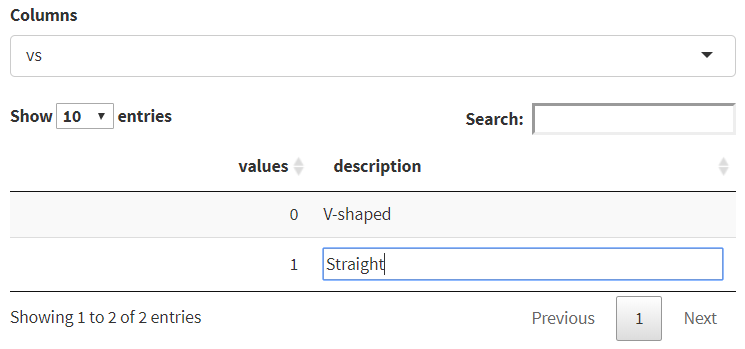
Step 5: Double-click inside a cell to edit or add information about your variables. See the instructions on the page for information about each column. The page will save automatically.



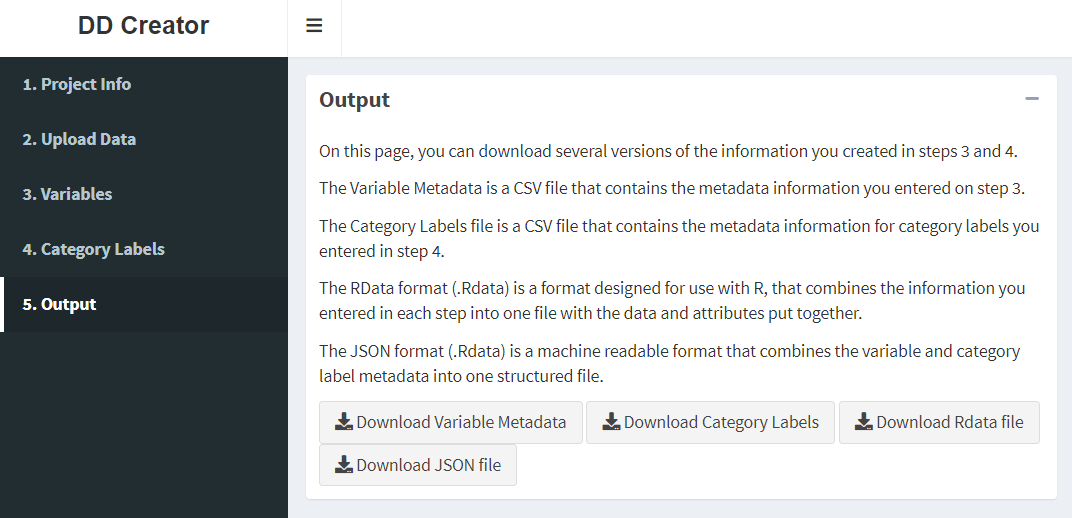
Step 6: Click the **Category Labels** tab on the left side of the page. The **Columns** drop-down list will populate based on your dataset.



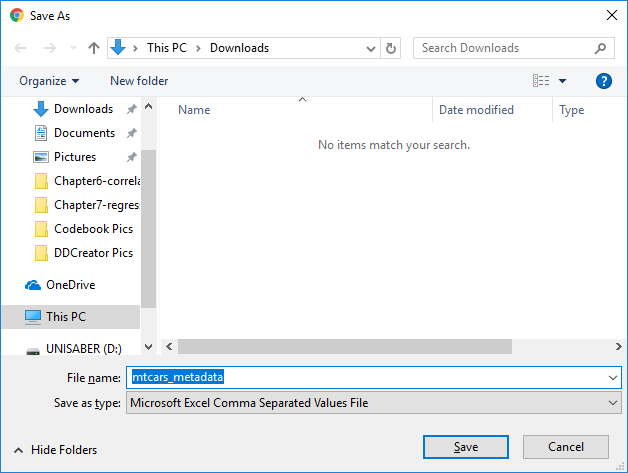
Step 7: Use the **Columns** drop-down box to select which variable description to change. Double-click inside a cell to make changes. The page will save automatically.



Step 8: Click the **Output** tab on the left side of the page. Select the file you would like to download. Variable Metadata and Category Labels will download as CSV files. You may also download an Rdata or JSON file.



Enter a file name and click **Save**.



**Summary**

In this tutorial, we provide three accessible options for researchers to create data dictionaries - shareable metadata information that allows others to read and understand the data provided. This type of tutorial is especially critical as transparency practices become more commonplace and as journals and researchers implement guidelines for sharing information and open data. The availability of large, open neuroimaging datasets led to the development of Brain Imaging Data Structure, which defines standards for neurological data [(BIDS; Gorgolewski et al., 2016)](https://paperpile.com/c/M0ehP6/aKvK/?prefix=BIDS%3B), and a similar movement is occurring in psychology with the Psych-DS project. This tutorial provides a manageable first step toward sharing data for all levels and skills of researchers, and the applications showcased here will continue to evolve as cohesive standards are formed through group discussion.

**Additional Resources**

References

[Arslan, R. C. (2018). How to automatically generate rich codebooks from study metadata. https://doi.org/](http://paperpile.com/b/M0ehP6/eQCc)[10.31234/osf.io/5qc6h](http://dx.doi.org/10.31234/osf.io/5qc6h)

[Boettiger, C., Chamberlain, S., Fournier, A., Hondula, K., Krystalli, A., Mecum, B., … Woo, K. (2018, October 10). dataspice. Retrieved November 24, 2018, from](http://paperpile.com/b/M0ehP6/1Sff) <https://github.com/ropenscilabs/dataspice>

[Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2018). *shiny: Web Application Framework for R* (Version R package shiny version 1.2.0). Comprehensive R Archive Network (CRAN). Retrieved from](http://paperpile.com/b/M0ehP6/tu60) [https://CRAN.R-project.org/package=shiny](https://cran.r-project.org/package=shiny)

[DeBruine, L., Buchanan, E. M., & Mohr, A. H. (2018, July 1). ddcreator. Retrieved November 24, 2018, from](http://paperpile.com/b/M0ehP6/dFMJ) <https://github.com/debruine/ddcreator>

[Gorgolewski, K. J., Auer, T., Calhoun, V. D., Cameron Craddock, R., Das, S., Duff, E. P., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. https://doi.org/](http://paperpile.com/b/M0ehP6/aKvK)[10.1038/sdata.2016.44](http://dx.doi.org/10.1038/sdata.2016.44)

[Hardwicke, T. E., Mathur, M. B., MacDonald, K. E., Nilsonne, G., Banks, G. C., Kidwell, M., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. https://doi.org/](http://paperpile.com/b/M0ehP6/ao31)[10.31222/osf.io/39cfb](http://dx.doi.org/10.31222/osf.io/39cfb)

[Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85. https://doi.org/](http://paperpile.com/b/M0ehP6/Fwut)[10.1177/2515245917751886](http://dx.doi.org/10.1177/2515245917751886)

[Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s Renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/](http://paperpile.com/b/M0ehP6/JmdO)[10.1146/annurev-psych-122216-011836](http://dx.doi.org/10.1146/annurev-psych-122216-011836)

[Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/](http://paperpile.com/b/M0ehP6/FuOm)[10.1126/science.aab2374](http://dx.doi.org/10.1126/science.aab2374)

[Noy, N. (2018, September 5). Making it easier to discover datasets. Retrieved November 24, 2018, from](http://paperpile.com/b/M0ehP6/Ll5T) <https://www.blog.google/products/search/making-it-easier-discover-datasets/>

[Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/](http://paperpile.com/b/M0ehP6/SoCY)[10.7717/peerj.175](http://dx.doi.org/10.7717/peerj.175)

[Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069. https://doi.org/](http://paperpile.com/b/M0ehP6/S4yU)[10.3758/s13428-015-0630-z](http://dx.doi.org/10.3758/s13428-015-0630-z)

[Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/](http://paperpile.com/b/M0ehP6/jU3P)[10.1177/0956797611417632](http://dx.doi.org/10.1177/0956797611417632)

[Stodden, V. C. (2011). Trust Your Science? Open Your Data and Code. *Amstat News*, *409*, 21–22. https://doi.org/](http://paperpile.com/b/M0ehP6/Zovq)[10.7916/D8CJ8Q0P](http://dx.doi.org/10.7916/D8CJ8Q0P)

[Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *The American Psychologist*, *61*(7), 726–728. https://doi.org/](http://paperpile.com/b/M0ehP6/yG93)[10.1037/0003-066X.61.7.726](http://dx.doi.org/10.1037/0003-066X.61.7.726)