Political Science Replication

Twitter: @PolSciReplicate | Facebook: https://www.facebook.com/PoliticalScienceReplicationBlog

MAR 16 2015

2 COMMENTS
BY POLITICALSCIENCEREPLICATION.WORDPRESS.COM REPRODUCIBLE RESEARCH

Good practice in data collection and storing

Research starts with data collection. Before you can do your analysis, you spend hours, weeks, months merging tables and transforming variables. This time is wasted if you don't keep detailed logs about this process. Here's a good practice guide.

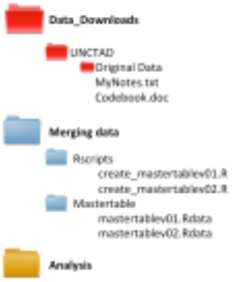
> Basic data management e.g. using syntax to create an audit trail, not creating multiple versions of the same file, etc. is hugely important, but rarely taught, and hence often very badly performed, introducing errors and data structure problems that haunt later analyses… (Chris Stride (www.figureitout.org.uk), statistician, Sheffield University)

Data Collection

- When downloading secondary data, keep a note of the date, version, source incl. URL, and their suggested citation.
- Download all codebooks available for that particular version of the data set.

Tip: I take screenshots when I download data, which contains much of that info.

Storing Data



(https://politicalsciencereplication.files.wordpress.com/2015/03/screen-shot-2015-03-08-at-10-20-53.png)

Create a consistent folder structure for

1. data downloads for raw data
2. data merging, and
3. data analysis

- Put the raw data into a separate folder.
- Never touch the original, raw data files. Ever. Do not change rows, columns, do not delete anything, do not add anything. **Original data are never to be touched again after downloading them.**

Tip: I have a folder called "Data_Downloads". Each sub-folder e.g. UNCTAD, WORLD BANK etc. contains the original excel or .csv files that I downloaded under "original data". Other people have different folder structures but they follow the same principle. See here (https://politicalsciencereplication.wordpress.com/2013/06/04/how-to-make-your-work-reproducible/), or in Gandrud's chapter on file management and data storing [pdf (https://github.com/christophergandrud/RepResR-RStudio/raw/gh-pages/other/ReproducibleResearch_Chapter2.pdf)], or by statistician Jeff Leeks (https://github.com/jtleek/datasharing).

> If research data are well organised, documented, preserved and accessible, and their accuracy and validity is controlled at all times, the results is highly quality data… *(UK Data Archive, 2011).*

Merging and Cleaning

- Merging and cleaning of data happens in a separate step from the data download. Therefore, create a folder called "merging data".
- Within that, I merge my data in an Rscript called "creating_mastertable.R" (you could also call this e.g. "merging.R"). This Rscript loads all original data, cleans them, and compiles them into one large table per country and year (in my case). This table will be saved as my new, tidy data set (http://vita.had.co.nz/papers/tidy-data.pdf) for the analysis (I call that "mastertable"). Never copy paste numbers from one excel file to another to create such a table.
- Whenenver I add a new variable to the data set, I switch from creating_mastertable**v01**.R to creating_mastertable**v02**.R because sometimes adding a new variable can mess things up.
- When you have added all variables together into your master table, save it as a new file e.g. mastertablev01.R or mastertablev01.csv. This can later be loaded into your analysis Rscript.

Tip: When I transform or recode variables (e.g. to take the log or to re-code conflict data), I add the newly created variable as an additional column to my master table. I can then check more easily if the transformation worked and I can also use both versions in my analysis to check for robustness.

> Researchers need to improve, enhance and professionalize their research data management skills (Corti et al, 2014)

Documenting Data

Everything that you do with your variables should be documented from day 1.

- Make notes in a simple text file and keep that in the same folder as your data merging Rcode. I sometimes write down my rationale for kicking out a specific year, or changing a number to "NA" because I happen to know that that particular year we cannot trust that data. I keep going back to these text files whenever I add new data to the mastertable, or when I update the data with more recent years.
- You can even call this file your 'codebook' and do that in pretty table in excel, word or latex. Keeping a codebook from day one will save you a lot of time later.

Sharing Data with Collaborators

All the above will help you to share your data with collaborators. You will make it easy for them to trust your data. Jeff Leek has excellent tips (https://github.com/jtleek/datasharing) on each of the files below that you should make available to your co-authors:

1. The raw data (original data, unchanged!)
2. A tidy data set (your master table)
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 -> 2,3 (Rscript, do-file etc with comments, or a separate document)

> Data management is a big (and sometimes daunting) task (Michael N. Mitchell, 2010)

Tips for R, STATA and SPSS

- Use syntax (http://offbeat.group.shef.ac.uk/FIO/course_syntax.htm) in SPSS.
- Create do-files (http://www.ssc.wisc.edu/sscc/pubs/sfr-dofiles.htm) in STATA for your file management.
- Follow conventions (https://google-styleguide.googlecode.com/svn/trunk/Rguide.xml) on how to comment your code in R.

- Consider using Rmarkdown (http://rmarkdown.rstudio.com/), knitr (http://yihui.name/knitr/), GitHub (https://github.com/features), project template (http://projecttemplate.net/getting_started.html), or an R package (https://github.com/jtleek/rpackages) to present your data collection, analysis and outputs all-in-one.

More Sources

- Hadley Wickham, Tidy Data [pdf (http://vita.had.co.nz/papers/tidy-data.pdf)]
- Svend Juul, Take good care of your data [pdf (http://www.epidata.dk/downloads/takecare.pdf)]
- Jeff Leek, The Elements of Data Analytic Style [I paid 12£ but you can download for free on Leanpub (https://leanpub.com/datastyle)]
- UK Data Archive, Managing and Sharing Data [pdf (www.data-archive.ac.uk/media/2894/managingsharing.pdf)]
- Chris Stride, Data management using SPSS syntax (http://offbeat.group.shef.ac.uk/FIO/course_syntax.htm)
- Michael N. Mitchell (2010). (http://www.stata.com/bookstore/data-management-using-stata/)Data Management Using Stata: A Practical Handbook (http://www.stata.com/bookstore/data-management-using-stata/)
- Corti et al. (2014), Managing and Sharing Research Data. (http://ukdataservice.ac.uk/manage-data/handbook/) A Guide to Good Practice. Sage.
- Ball, Richard J. and Medeiros, Norm, Teaching Students to Document Their Empirical Research (July 12, 2011). Available at SSRN (http://ssrn.com/abstract=1892168).
- BITSS Summer Institute Network, A collection of teaching materials (https://osf.io/8vmr3/) to promote transparency in research

This collections of best practice examples was inspired by a discussion on good practice in data management on the Quantiative Methods Teaching list (http://mailman.ncrm.ac.uk/mailman/listinfo/quantitative_methods_teaching).

Tagged cleaning, data management, reproducibility, secondary data

2 thoughts on “Good practice in data collection and storing”

John F Hall says:

March 18, 2015 at 11:00 am

If you want to see some really good documenmtation, go to
http://www.europeansocialsurvey.org/docs/about/ESS1_end_of_grant_report.pdf
To see later reports, just change EES1 to ESS2 3 4 or 5.

Reply

2. *Towards a more comprehensive replication standard in political science: reproducible data collection | Political Science Replication* says:
   August 5, 2016 at 12:27 pm
   […] is an exellent point. I have discussed earlier that good practice for data collection entails to keep detailed logs about the sources and all […]

   Reply