Datasharing – yes, please! An attempt at a beginner's guide

Posted on July 13, 2016July 14, 2016 by Christina

Science is becoming more and more open and transparent, and I think that's awesome. An important aspect is sharing whatever information is necessary to reproduce results, usually that includes data and scripts. While open science can be bene (https://elifesciences.org/content/5/e16800)ficial (http://whyopenresearch.org/) for a researcher, this practice is still being met with some (justified (http://rsos.royalsocietypublishing.org/content/3/4/160109)) skepticism, but has become more and more accepted and common in research; in fact PLOS One (http://www.plosone.org) for example made it a requirement for publication (how well that's going is a different story (https://twitter.com/hashtag/plosdatapolicy)). Funding (http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm) agencies (https://www.openaire.eu/opendatapilot) across the globe are quickly following suit, so chances are high you either already have to or will in the near future think about data sharing. But what does it entail?

There are several issues that in my view do not receive enough attention, and that add unnecessary hurdles in the sharing and re-use of data. But first, let me get this out of the way: sharing data is great, but you should do it the right way. If you succeed, you will not only help the community, but also yourself by making your work more visible and even citable. This way you get credit for your work, and make it more likely others build in what you have done by making it much easier for them. Everybody wins!

Of course, you cannot share sensitive data that for example make your participants identifiable, I won't discuss this topic in detail here. But there is another side to the ethical coin: By sharing your data, you reduce duplicated efforts, improve the reliability of your whole research field, and return your (probably) publicly funded research output to those who actually made it possible.

Despite all the advantages of data sharing, and the calls for it becoming the norm, however, there is in my view a shocking lack of concrete assistance in doing so. This would include step-by-step guides how and where to post data, which I couldn't find. So as a sort of start, I am listing a few things I learned when adding data to my own (few) publications and as a reviewer of papers.

Where to post data?

There are a lot of repositories (http://www.re3data.org) out there that can host different types of data. But it's easy to get lost in this forest and probably just as easy for your data to be lost there.

One thing that helps your data being discovered is to group it with many similar things. For example, if you have MRI scans, choose a repository that many colleagues working on similar topics use. Or when you have recorded videos of children, Databrary (https://nyu.databrary.org/) seems a great home for those. But such perfectly fitting repositories are not always an option, so you might have to look for the next best thing. When I wanted to post a speech corpus, I was looking first for a place where people typically store their experimental stimuli, but I couldn't find a repository dedicated to that purpose. In the end I went to the language archive (http://tla.mpi.nl) of the Max Planck Institute in Nijmegen. They host a lot of different speech corpora and take donations; in fact they were extremely helpful and patient with me and I had a great experience working with them. (I won't go into detail how sharing this corpus shouldn't have been my job to begin with, as the authors promised sharing in the paper everyone enthusiastically cites…)

Another advantage of the language archive (http://tla.mpi.nl) for me has to do with the second "trick" to make your data more easily discoverable: **metadata**. Now what is this metathingy? I had no clue, but the MPI people talked me through it. Minimally, metadata describe the content and purpose of your files. Why, where, how were recordings made, and in which language? Who was the main responsible person? And so on. The more information you add, the easier it is for others to re-use your data. If you have a good guide to metadata, please share it. I've been looking for one myself…

How to make sure my data stays where I put it?

One thing we should all aim to avoid is **link rot** (it's a real thing). This means you make a website for your work and then 10 years (or sometimes even 1 year) later someone clicks and sees a nice error message. Your hard work is forever lost. This can happen because you changed your mind about which website best reflects you and moved away from babydoc17.edu to drbaby22.org. In this case, the fix is easy: keep the old url and forward it to your new site!

Another scenario is that you change institutions and your former university kindly deletes the home page they used to host. Yep, that's not uncommon, much like the sister case that your institution changes something about their infrastructure and all links have to be updated. What can we learn from this? Don't use your institution's platform *unless* they have a dedicated file hosting service which gives out **permanent handles**. Even if they change servers, this handle will always lead to your data. Permanent handles can for example come in the form of a DOI (http://www.doi.org) or a HDL (http://www.handle.net) (there are many more, these are just the ones I see most frequently). This makes citing everything also easier! So either ask your library, who are typically in charge of such things, or check the platform of your choosing whether they give out permanent handles of some sort. Some examples are: OSF (http://www.osf.io) (DOI available for public registrations (https://osf.io/faq/) [a frozen version of whatever information you want to put online], and they help you move when you change your mind about the location!); figshare (https://figshare.com); zenodo (https://zenodo.org); and so on. If I missed your favorite, please let me know in the comments!

What about sharing scripts?

Scripts are a separate issue, I think the natural home for them is with some git repository. Git (https://git-scm.com) repositories are extremely useful already when creating and using scripts, because on one hand you have a backup on some git server and a local copy to play with and on the other hand you can travel through time. That means if you accidentally delete or break something, git can save you. It's great! And there are a lot of helpful tools and platforms and so on that make using git quite straightforward. If you are curious now, here is a very, very gentle guide to git (http://rogerdudler.github.io/git-guide/), and you can also check out Page's R course (https://cogtales.wordpress.com/2016/03/18/r-course-lesson-0/), where she starts with data management.

But back to the topic, git platforms such as github (http://www.github.com) can also help you share scripts after your project is done. There are options for you to create permanent links to a specific version of a script (in their own format, so I view this with some caution; here are the details (https://help.github.com/articles/getting-permanent-links-to-files/)). It's probably even easier to collect everything for a project in a folder, called repository, and add a DOI (https://guides.github.com/activities/citable-code/). Zenodo (http://zenodo.org/) is your best friend here.

Regardless of this extant support structure that makes a scientist's life so much easier in the best case and at least is easy to use when just posting script, I still see a thing that makes my toenails curl: **scripts shared as .doc or .docx. No! Please don't!**

This leads us to the next point quite smoothly, look at me mastering the segue for once.

What's a good format?

Did you ever ask a colleague for data that was older than 10 or even 5 years? Chances are you receive a file no existing program can easily read. I am sitting on a kindly shared dataset that I simply don't know how to open, because the software was discontinued and the format abandoned. This problem is easy to ignore for now if you are using software that is currently quite common (for example a specific file output format in your experiment presentation software). But think of future you, who will not be able to use this script after moving labs three times and having recycled all computers currently in use.

So if possible share data in simple formats, a handy list has been compiled by the folks at Datadryad (https://datadryad.org/pages/filetypes). To summarize: make your scripts and data readable and avoid too specialized, proprietary formats.

A case study: Sharing data and scripts with R Markdown on OSF

R (https://www.r-project.org/) is a free, open source software. If you are still using Excel, this course (https://cogtales.wordpress.com/2016/03/18/r-course-lesson-0/) will help you. I use it with R Studio (https://www.rstudio.com/), which (as explained in Page's course (https://cogtales.wordpress.com/2016/03/18/r-course-lesson-0/)) has a very neat integration with github (http://github.com) and bitbucket (https://bitbucket.org/). Once R and R Studio are set up, you're ready to go. You can now write R Markdown (http://rmarkdown.rstudio.com/) documents which combine code and explanatory text, it's quite easy. Compared to latex, an R Markdown document looks much more like "normal" text. The neat thing is that you can explain and show your analysis code, and be sure that your numbers in the text are always right. (Writing papers in R Markdown is also possible, but let's stay on topic.)

A very neat example of a published paper with all underlying scripts and data made available on OSF using R Markdown and .txt files (because they never go out of fashion) was recently shared by fellow blogger Gwilym (https://gwilymlockwood.wordpress.com/): https://osf.io/ema3t/ (https://osf.io/ema3t/)

If you have other examples, add them in the comments!

You see, sharing isn't that hard, it just requires some thinking ahead, in more ways than one (a great way to do so is a "data management plan", which is probably worth its own post). In the end, future you also wins, be it by being able to retrace steps or because you were cited 100 times for that brilliant piece of code you wrote.

Tagged R, ScientificPractices, Statistics   Leave a comment