Running head: BORN-OPEN DATA

The What, Why, and How of Born-Open Data

Jeffrey N. Rouder

University of Missouri

Abstract

Although many researchers agree that scientific data should be open to scrutiny to ferret out poor analyses and outright fraud, most raw data sets are not available on demand. There are many reasons researchers do not open their data, and one is technical. It is often time consuming to prepare and archive data. In response my lab has automated the process such that our data are archived the night they are created without any human approval or action. All data are versioned, logged, time stamped, and uploaded including aborted runs and data from pilot subjects. The archive is GitHub, `github.com`, the world's largest collection of open-source materials. Data archived in this manner are called *born open.* In this paper, I discuss the benefits of born open data and provide a brief technical overview of the process. I also address some of the common concerns about opening data before publication.

The What, Why, and How of Born-Open Data

Psychological science is beset by a methodological crisis in which many researchers believe there are widespread and systemic problems in the way researchers produce, evaluate, and report knowledge (Pashler & Wagenmakers, 2012; Yong, 2012). This crisis is precipitated by the publication of fantastic extra-sensory perception claims in mainstream journals (e.g., Bem, 2011; Storm, Tressoldi, & Di Risio, 2010), by the widespread belief that well-publicized effects may not be replicable (Carpenter, 2012; Kahneman, 2012; Roediger, 2012) and by several cases of outright fraud. Such a crisis is very worrying because if researchers and labs cannot have confidence in one another, then the core of the field is at risk.

This methodological crisis has spurred many proposals for improvement including an increased consideration of replicability (Nosek, Spies, & Motyl, 2012), a focus on the philosophy and statistics underlying inference (Cumming, 2014; Morey, Romeijn, & Rouder, 2013), and an emphasis on what is now termed *open science*, which can be summarized as the practice of making research as transparent as possible. Practitioners of open science make several elements of the research freely available at an archival repository. These elements include the details of data collection, the data themselves, and code for analysis. The rationale behind making science open is clear—as psychological science becomes more open, questionable practices such as fortuitous censoring of subjects becomes more detectable. Moreover and more importantly, researchers who place their methods, data, and analyses in open sources have an incentive to use better judgment in their methodology and more carefully consider the ramifications of their decisions in analysis. Indeed, Wicherts, Bakker, & Molenaar (2011) report that research with open science has stronger evidence and more sound analyses than research not in the open.

One critical part of open science is *open data*, the act of making raw data available on demand to anyone. Recently, there has been a grass-roots movement, The Agenda for Open Research (agendaforopenresearch.org), for researchers to make their data open and for journals to insist they do so.[1] The rationale and context for this call are provided in Morey et al. (2014). A researcher may make her or his data open by publishing them at a sanctioned or curated web repository with some form of institutional support. Examples of such repositories dedicated to archiving open source materials include Github (github.com), Figshare (figshare.com), Dryad (data dryad.com), and Open Science Framework(osf.io). Other curated sites include university-related archives and society-related archives.

The Open-Data Paradox

Open data, unfortunately, seems to be paradox of sorts. On one hand, many researchers I encounter are committed to the concept of open data. Most of us believe that one of the defining features of science is that all aspects of the research endeavor should be open to peer scrutiny. We live this sentiment almost daily in the context of peer review where our scholarship and the logic of our arguments is under intense scrutiny.

On the other hand, surprisingly, very few of our collective data are open! Consider all the data that is behind the corpus of research articles in psychology. Now consider the percentage is available to you right now on demand. It is negligible. This is the open-data paradox—a pervasive intellectual commitment to open data with almost no follow through whatsoever. The open science paradox seems to extend well past psychology and has been documented in other fields including biology and ecology as well (Punewska, 2014). Yet, there are disciplines where data sharing is expected and common, for example, in astronomy (Pepe, Goodman, Muench, Crosas, & Erdmann, 2014)

Many of my colleagues practice what I call *data-on-request.* They claim that if you drop them a line, they will gladly send you their data. Data-on-request should not be confused with open data, which is the availability of data without any request whatsoever. Many of these same colleagues may argue that data-on-request is sufficient, but they are demonstrably wrong. Wicherts, Borsboom, Kats, & Molenaar (2006), asked 149 author teams to release 249 data sets that had appeared then recently in American Psychological Association journals. Only 11% complied with the initial request, and only an additional 16% with repeated requests. A full 73% of author teams never complied. My own view is that investigators who practice data-on-request rarely prepare for the request and may not comply for any number of reasons: The data themselves may be in an arcane format or may be misplaced. Sometimes the raw data is unavailable; other times the code to analyze the data is missing. Sometimes investigators delegate data curation to graduate students and postdoctoral researchers whose tenure at the institution is necessarily limited. Investigators themselves may become ill or leave the institution. Perhaps most alarmingly, investigators may believe their data are not organized nor "clean" enough for release. One colleague, for example, lamented the amount of time it would take to cleanse the E-Prime generated output of all the extraneous information.

Unfortunately, I was a living example of the open data paradox. I was committed to open data. I made this commitment boldly in my National Science Foundation data-management plans. My data were supposed to be archived at my institution's curated, open repository. Yet, sadly, this did not often happen. Why not? Some of it was a lack of effort. It was a pain to document the data; it was a pain to format the data; it was a pain to contact the library personnel; it was a pain to figure out which data were indeed published as part of which experiments. Some of it was forgetfulness. I had no routine and no incentive for archiving data. Even with the best of intentions, it seems that making data open took too much time, effort, and attention. No wonder people revert to making data available upon request. It is so much easier than making data open. Yet, Wicherts et al.'s findings strike me as unnerving—without a field-wide commitment to truly open data, it is difficult to have the utmost confidence in experimental findings in the literature.

Born-Open Data

My personal failures to live up to my commitments in my data management plans has forced me to reconfigure how we collect and process data in my lab. I decided that if we were to make our data open, neither I nor my students could be relied upon to do routine tasks like file uploads. Instead, we needed an automatized system, and here is what we came up with: Behavioral data are uploaded to GitHub, an open web repository where it may be viewed by anyone at any time without any restrictions. This upload occurs nightly, that is, the data are available within 24 hours of their creation. The upload is automatic—no lab personnel is needed to start it or approve it. The upload is comprehensive in that all data files from all experiments are uploaded, even those that correspond to aborted experimental runs or pilot experiments. The data are uploaded with time stamps and with an automatically generated log. The system is versioned so that any changes to data files are logged, and the new and old versions are saved. In summary, if we collect it, it is there, and it is transparent. I call data generated this way as *born-open data* and hope born-open data becomes a standard.

To see a subset of data we have been collecting since the start of 2015, point your browser to `github.com/PerceptionCognitionLab/data1`. One folder is `bayesObserver` which contains the data from a set of experiments designed to assess whether people combine information from the stimuli with base rates in an ideal manner. Here there are a few experiments, and let's explore `be1`. Each file corresponds to a different participant, and when clicked on, the raw data are available. Of course, it is difficult to understand what the data mean without column labels, and these are provided in a separate file, `be1.txt`. Even with the column labels, it is quite difficult to understand the experiment or the meaning of the data without a guiding publication. But with the publication, the data are easily understood and reanalyzed. Hence, we document when we collected the data, and it is open to scrutiny to anyone with the corresponding manuscript.

I have found that there are several advantages to using born-open data:

- The use of born-open data incentivizes me and my students to use the highest level of judiciousness in analysis. I suspect that we will have an increased awareness of our decisions and their consequences. This is the intended effect, and uncomfortable as it may be, I view it as an advantage.

- With born-open data, we do not make data-management mistakes. It is much easier to audit and document all data and all analysis steps. Gone are manually labeling data or wondering if the version on the memory stick or the hard drive is the latest. I load data into the analyses right from GitHub without any intermediary downloads or any proliferation of files. There are no more “raw” files and ”cleaned” files. Instead, there is just data and code. Here is snippet of code that loads up data from Experiment be1. It can be run on any computer with R and the R package “RCurl.”

```r
library(RCurl)
name1="https://raw.githubusercontent.com/PerceptionCognitionLab/data1/master/"
name2="bayesObserver/be1/be1.all"
intext=getURL(paste(name1,name2,sep=""))
dat=read.table(text=intext)
colnames(dat)=c('sub','trl','blk','tBlk','cond','ecc','stim','rt','resp','pts','ptsTot')
```

- With born-open data, backup is automatic. The GitHub copy backups the local copy. Importantly, both copies are versioned. Consequently, there is no reason to worry about weekly incremental backups, monthly backups, restoring from backup and the like.

- Born-open data simplifies sharing of data within the lab and with collaborators. I often email collaborators R code that loads the data fro GitHub and cleans the data, and they can provided analyses as they wish.

Technological Elements

The technologies I use to create born-open data are fairly standard and can be adapted for Windows, Mac-OSX, and Unix/Linux. They are not the easiest technologies to learn, but given that archiving data is a critical enterprise that may affect the researcher for the length of her or his career, it seems that investing in these or comparable technologies is warranted and beneficial. There are four basic elements: shared local storage, git repository software, the use of the GitHub open-source web repository, and a task scheduler. I provide a brief overview of each step. There is good and bad news here for those wishing to implement the system. The good news is that all the protocols are standard and well used. Any skillful IT professional should know them well or be willing to learn them. The bad news is that it will require a bit of tinkering with the specifics dependent on machines, networks, and operating systems.

Shared Local Storage

In my lab behavioral data are collected across several computers. One key problem is coordinating among these computers. Code to run the experiments must be placed on each, and the outputted data must be merged into a master set. These tasks must be done accurately to insure the integrity of the data. In some labs, assistants move files from one machine to another, often via memory sticks. Not only is this approach labor intensive, it may not be reliable. The better approach is to use one shared drive. The lab server shares a drive where both the experimental code and the data are stored. This drive is the master drive, and the other computers that collect the data read and write to it rather than to their own internal drives. When behavioral data are created, then, they are created only on the master drive and there is no need to move data files. Setting up a shared drive is not too difficult and most researchers have either the knowledge to do so in their preferred operating system or institutional support.

Git/GitHub

Once data have been stored in a single location, they must be made into a *repository*. A repository is more than a set of files; it contains also versions, changes, and logs. Because repositories contain logs and versions, they provide for a digital audit trail. For me, the easiest approach is to use a single repository for all experiments in my lab. This repository spans several folders and files. Making a repository, versioning, and logging is performed by a dedicated software application. Over the years there have been several options, but there is a dominant application: *Git*. Git is dominant because in my opinion it is the most flexible and convenient. One advantage of Git is that it interfaces seamlessly with GitHub, a public website that hosts a ridiculously large number of open-source projects. Git is very easy to find and install, and the GitHub clients, such as GitHub for Windows and GitHub for Mac include it. Git is part of the linux operating system and no further installation is needed for linux servers. Help for git is available in the wonderful online book plainly titled, ”Git Book” at `http://git-scm.com/book/en/v2`. Git does take some time and energy to learn, but there is a big payoff. It can be used to version much of the academic process including analysis and manuscript preparation. I find it indispensable in keeping a reliable pipeline from data collection to final manuscript

At the heart of the system is GitHub, the largest web host for sharing open-source projects (`www.github.com`). GitHub may be used at no cost and when used in this mode, the information is freely and publicly available. GitHub was originally designed for the development of open source code, but is now used for a wide range of publicly available projects. Anyone can make a GitHub account and use GitHub software to create a local repository and link it to GitHub. GitHub is designed to be fairly user friendly and provides extensive help and services. Git repositories may be uploaded to GitHub. The GitHub copy is then made available on the web either through the Git system or through the web on demand. Trained IT professionals should be familiar with Git and GitHub because many useful projects are archived on these systems.

Execution and Scheduling

The last step is to execute the adding of the day's data files to the Github repository. The steps are as files: 1. All new files with certain filename extensions are added to the local version of the repository. Then the local version is committed, meaning that the state of all files is logged, and the commit is timestamped and labeled. Finally, the new state is uploaded to the GitHub site. I have written these steps in the following script, which is executed nightly.

```
#!/bin/bash
git add *.dat.*
git commit -am "automatic commit"
git push
```

This script is executed nightly by a *task schedueler*. Setting up the a task scheduler is the last step. Programs that execute others at a set time are called task schedulers. I use CRON tables on my linux server. Task scheduling is built into Mac OSX (`http://support.apple.com/en-us/HT2488`) and Windows (`http://windows.microsoft.com/en-US/windows/schedule-task`). Trained IT personnel should be capable of writing Git scripts and automatizing their execution.

Technical Concerns

Although the Git/GitHub system is working well, there are important technical concerns. The first is about the sizes of files and repositories: Git does not manage very large files efficiently. I have read though I cannot find the citation that Git is not recommended for files larger than 100MB. Repositories should not be larger than a gigabyte. These are not hard limits, but Git tends to slow down at these larger sizes. These limits are irrelevant for my lab where we collect behavioral performance data that are produced at a rate of a few kilobytes per week. Yet, they may be relevant for physiological data in which single files may exceed 100MB and the collection of raw data for a single experiment may exceed several gigabytes. There is no limit I know of the number of repositories, and if our behavioral repository becomes too large, we will start a new one. A second concern is about GitHub. GitHub is not a curated archive. In most curated archives, reposited materials cannot be deleted or changes. The material is immutable, and indeed, most University and society archives work this way. Files on GitHub may be changed by the uploader. Fortunately, changes are logged and older versions of the same file are saved. I think functionally GitHub acts as a curated archive, but nonetheless, the distinction should probably be maintained.

Concerns About Born-Open Data

I think the concept of open data in general and born open data in particular are needed in psychology to promote better science. The current culture of closed data and analysis serves us very poorly, and is a partial contributor to the current crisis in confidence.

Unfortunately, many people I talk to do not think opening data is wise or needed. In my conversations I have heard a number of concerns about born-open data which may prevent some from adopting it. I think these concerns are misguided:

Concern: Proprietary Data. Some researchers view their data as their property and worry about the wisdom in sharing proprietary information. Most of us are aware of the amount of time, care, and effort required to obtain good data. Nonetheless, if data are used to make scientific claims, then they must be open to scrutiny. It's fine of course to have proprietary data and there are several contexts in which data are most usefully kept proprietary. None of these contexts, however, is in the scientific literature, at least not in my opinion.

Concern: Privacy. Some researchers cite the privacy of their participants as a primary concern. Privacy concerns are important and legitimate. The solution of course is to archive deidentified data. In my lab we program the experiment-run scripts to generate files with deidentified and identified data. Only the deidentified files are added to the repository. Deidentification will often mean that not all demographic data may be available as some of these data may identify the participant. There are some cases where data cannot ethically be made open, say those involving illegal activities which could put the participant at risk. As a rule, researchers should make open those data that may be ethically shared.

Concern: The Fear of Being Scooped. Some researchers support the notion of opening up their data after publication rather than before. Others support a related view of having control of who sees their data before publication. The downside of this view, however, is that reviewers and journals neither have a method of scrutinizing the data in the peer review process nor have any enforcement mechanism that the data will be made available after publication. Given the results of **?**, it is likely that researchers who insist on sharing after publications are unprepared to share.

The argument for post-publication sharing is that the researchers should be the first to publish their data. I agree, and I do not know anyone who does not. I believe born-open data does not threaten the ability of researchers to publish first because it is exceedingly difficult to understand what data means without a guiding publication. If you think it is easy, take a look at Experiment be1 discussed earlier. The column labels are provided; can you figure out what happened in the experiment? Even if you could, kindly remember that scooping data is fraudulent and representing others data as your own is a form of deciept. Those that practice born-open data can document the production with time stamps. The risk of being scooped is microscopic. By any reasonable standard, the gain to the community of open data on submission more than outweighs the risk of any individual being scooped.

Making data open on submission is not the same as born-open data. I prefer born-open data for my lab because of the nightly automatization. The benefits of automatization, detailed above, are important, and I fear that without it, my data would not be made open. For me, the gains in having instantly open data without any thought or additional steps outweigh any fears of being scooped. I suspect that the same holds true for many researchers.

Concern: Professional Vulnerability. I suspect a salient if not often discussed reason people are unenthusiastic about open data is a sense of vulnerability and fear. For me, the worst case scenario is that someone is going to find an indefensible mistake, and, as a consequence I will have to retract a paper, or even worse, look incompetent among my colleagues. And I know that anyone who looks critically at my born-open data does so as a motivated skeptic who does not believe my claims. And against this risk, there is no short-term gain for opening data. Many researchers are aware of this vulnerability. Here are some examples of the arguments I have heard:

- One of my colleagues told me bluntly that she fears open data. She fears that someone will look at her data and her conclusions and find a mistake, and she will look and feel "stupid."

- One of my colleagues told me that his data are not well organized and he does not want the field to see this state of chaos. It would take him much time and effort to get his data to the standards with which he would feel comfortable sharing.

- One of my colleagues has revealed that he doesn't trust the intentions of others. Another told me she distrusts the self-appointed replication police who are viewed in some quarters as trying to shame otherwise good researchers. I do not agree with those who attribute ill intentions to critics. Instead, I attribute these arguments to a fear of scrutiny, though those that make them this way are seemingly less self-awareness than those who understand where their fears come from.

The fear of scrutiny is a deep and real one, and it is compounded for people who do not have the security of tenure. I have great sympathy for people who are scared to open their data because, honestly, I am scared to open my data as well. It is a scary thing. Clearly, the practice of open data, and especially born-open data requires being comfortable with additional vulnerability and scrutiny. Yet ,it is this very vulnerability that makes for better scientific practice. The very act of putting our data out there makes our pipelines more reliable and and our decisions more judicious. And this gain holds even if nobody happens to scrutinize our data. This vulnerability is quite a good thing in the long run.

Summary

Born open data has the potential to lead to a better, more self-reflexive, and more open state of psychological science. I think it is incumbent of the most senior of us to lead the effort to make data open. We have the most security, the most evolved perspective on being critiqued, and the least to fear from increase scrutiny. If we do so, then the younger people may follow. And, as a result, we will all benefit.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. Retrieved from `http://dx.doi.org/10.1037/a0021524`

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558-1561.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Kahneman, D. J. (2012, September). *Open letter: A proposal to deal with questions about priming effects.*

Morey, R. D., Chambers, C. D., Lakens, D., Lewandowsky, S., Wagenmakers, E.-J., & Zwaan, R. A. (2014). *An agenda for open research: Incentivising open research practices through peer review.* Retrieved from `http://agendaforopenresearch.org/agenda10.pdf`

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68-75. Retrieved from `http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x`

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among us astronomers. *PLoS One*, *28*(9). Retrieved from `DOI: 10.1371/journal.pone.0104798`

Punewska, M. (2014, December). Scientists have a sharing problem. *The Atlantic Monthly*. Retrieved from `www.theatlantic.com/health/archive/2014/12/scientists-have-a-sharing-problem/383061/`

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992-2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*, 471–485. Retrieved from `http://dx.doi.org/10.1037/a0019457`

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011, 11). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*(11), e26828. Retrieved from `http://www.plosone.org/annotation/listThread.action?root=19627`

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726-728. Retrieved from `http://wicherts.socsci.uva.nl/datasharing.pdf`

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298-300.

Author Note

Email: rouderj@missouri.edu, Web: pcl.missouri.edu; Twitter: @JeffRouder. I thank April Swagman, Evie Vergauwe, Timothy Ricker, Richard Morey, and Eric-Jan Wagenmakers for helpful discussions about open data and its implications. I am indebted to Igor Izouimine, Richard Winkel, and Mike Watson for technical support and techical resources. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

Footnotes

[1] I am a signatory to this agenda and support it completely.