What is open science?

Form hypothesis, collect data, publish, repeat. At its most basic form, this describes the cycle researchers in all fields are familiar with. They perpetuate their own cycle of research ideas, data collection, and analyses, and are thus shaded from other scientists’ scrutiny. The steps taken to uncover new breakthroughs are themselves covered. This blinding veil has led to some of the most unethical fraud and false discoveries in science. It’s influenced us to think twice about how we eat a buffet dinner (Lee, 2018), and to believe that we can see into the future (Engber, 2017).

Scientific study can be described as an iceberg, with a majority of the work gone unseen. Beneath the surface, researchers are prone to making mistakes that alter the outcomes of their studies. “P-hacking”, for example, develops faulty analyses that are then published in journals as false positive reports (Nelson et al, 2017). Because these fabricated results are nearly impossible to replicate (a necessity in the fundamentals of the scientific method), researchers have now realized that hiding the cycle of research is detrimental to its accuracy. In the years 2010-2012, the “replication crisis” erupted, and influential studies like Brian Winsink’s deceptive buffet analysis or Daryl Bem’s eyes into the future could not be replicated thenceforth. If experiments lack reliability, their validity then comes into question.

To address these issues, Brian Nosek and Jeffery Spies began the Open Science movement to aid other scientists by creating an online platform in which they could openly record, report, and share data (Nelson et al, 2017). Their foundation, the Center for Open Science, allows collaboration of researchers in any stage of research development. The cycle of open science unfolds the research process for every aspect of research to be shared.

Engber, Daniel. (2017). Daryl Bem proved ESP is real. *Slate.* Retrieved from <https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html>

Lee, Stephanie M. (2018). Here’s how Cornell scientist Brian Winsink turned shoddy data into viral studies about how we eat. *BuzzfeedNews.* Retrieved from <https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking#.mgZ1LqRBK>

Benefits of open data

**Benefits of Open Data**

Open data is beneficial for both individual researchers and science, because it facilitates the spread of knowledge and improvements in research. Piwowar says open data allows scientists to develop new hypotheses, see multiple perspectives on different research, and identify errors. In fields where p-hacking and false positives run rampant, such as psychology, open data discourages fraud and makes replication more likely (Piwowar). There have even been studies that show “papers with publicly available datasets receive a higher number of citations than previous studies without data” (Piowar). However, open data must be clear and understandable to successfully replicate and experiment, which is a prerequisite for verifiable research (TYS). While some may say that open data is not necessary because scientists can just share data when they are requested to, Rouder says that only a small percentage of those scientists end up sharing their data.

Open data already has a history of being successful. In the 1990s, when researchers were trying to decode the human genome, open data made it possible for many individuals to collaborate (TYS). Those involved met in 1996 to discuss how they would go about the decoding process and manage their data (TYS). Considering the massive importance of the Human Genome Project, it can be said that open data has already had a large impact on science.

Fortunately, with the advancement of technology and the widespread use of the internet, open data has come a long way even since the 1990s. There are multiple repositories that researchers can take advantage of when they wish to make their data open. GitHub, Figshare, Dryad, and Open Science Framework are just a few of these (Rouder).

Rouder and his team of researchers use what they call “born-open data.” All of their data is on Github and it updates itself nightly so that the researchers do not have to go in and update everything themselves. This has helped his team avoid much of the hassle associated with open data. Additionally, since Github has institutional support, they do not have to worry about anything happening to their data for decades (Rouder).

Concerns of open data

Concerns of Open Data

Open Data hasn’t been fully embraced by everyone. Before these people commit to Open Data, they see problems that need to be solved first. These problems create barriers to making their data public. Researchers cite legal, practical, and competitive concerns that have kept them from advocating and acting for Open Data.

Legal concerns that cause apprehension for some are about rights and data confidentiality. From the findings in the Tenopir et al. (2011) paper, one of the main concerns of researchers is that they don’t have the rights to make the data public (Houtkoop et al., 2018, pg. 77). Part of this problem can come from sponsored research. Sponsors can require confidentiality in the study, meaning they don’t want the data to be available to others (Leetaru, 2017). Funding is integral in conducting research, and that funding can come from private sponsors trying to use the results for future products. Sponsors don’t want to potentially miss out on profitable results because someone else uses the data first to do so. Researchers don’t want to not be able to conduct the research they want to, so they could be forced to keep the data private since they don’t have the rights to make it public. There is also data confidentiality regarding participants and those involved in the study that cause concern. Anonymity is crucial in conducting sensitive research, and some worry that making the data public would hurt those who participated.

In sensitive research studies, there is a risk that the participants could be identified, but there are ways to decrease the risk and secure confidentiality. The most common way to protect identity is removing direct identifiers provided in the datasets (Fraser & Willison, pg. 13, 2009). K-anonymity is used to protect against identifying someone by indirect variables. It involves making the variable into a larger category, such as age to age group or birthdate into birth year (Fraser & Willison, pg. 13, 2009). Pseudonymization is used to create a fake “ID” for a participant across multiple datasets (Fraser & Willison, pg. 15, 2009). Besides direct identifiers, listing specific dates could lead to that person being identified. Many people have access to birthdays and those closer to the individual have a chance of knowing when surgeries or treatments took place. To help protect this individual, specific dates like birthdays or surgeries should be listed in “durations or intervals” (Fraser & Willison, pg. 20, 2009).

Practical concerns that create barriers for some are about the process itself of making data public. Money, time, set standards are seen as three practical complications to pursuing Open Data. In the paper from Tenopir et al. (2011), the two biggest concerns cited by researchers were lack of funding and inefficient time and one of the most common minor concerns were researchers not having set standards for making data public (Houtkoop et al., 2018, pg. 73). Lack of funding leads to researchers fears of wasting time and money because of lack of set standards. For researchers who are new to Open Data, it might be hard finding out all the intricacies of it, such as where and how to make their data public. When researchers don’t know the data sharing standards that means in order to make their data public, they must find out how and where to publish data and then actually publish it. This process could mean less time and money being devoted to aspects of the study that may be required or mandatory.

However, there are set standards created by The Center for Open Science, called the Transparency and Openness Promotion. There are three levels that a journal can adopt into their submission guidelines that researchers can follow and know what and how to utilize data sharing (TOP Guidelines). As more journals start to adapt to the Open Science movement, more researchers will know how to share their data just by going to submit their work, instead of having to look it up specially.

People also are concerned about Open Data due to the competitive nature of conducting and publishing research. Competitive barriers are ways in which making data public would disadvantage the original researcher. Loss of credit is an issue, where researchers don’t want to make their data public before they publish their own results because they don’t want their data “scooped”, or the data is published by another researcher before your own is (Houtkoop et al., 2018, pg. 73- 77). With making data public, researchers may also feel that a loss of control over their own work is a competitive barrier to Open Data. For example, researchers asked in one study said they were concerned about others rejecting their own conclusions because they analyzed the data differently, others findings errors in the data, others misinterpreting the data, and “loss of control over intellectual property” (Houtkoop et al., 2018, pg. 77). Some studies can result in data that can become “patentable intellectual property” but making the data public can hurt the patent process (Barron, 2018). In that case, making the data public can result in both a loss of control and a loss of potential money and funding from sponsors.

Open Data isn’t the end of marketable research. One way for researchers to maintain their competitive edge is requiring secondary researchers to cite the original researchers when using their data (Houtkoop et al., 2018, pg. 73). The fear of “scooping” has also been addressed. In one study, two research labs who both used open science practices were observed as they went through a research project. It found that their fears of “scooping” were reduced when they focused on the research itself, as opposed to publishing being the main motivator (Laine, pg. 12, 2017).

Is there even open data now? (what’s the status of open data)

* Google searchable open data

There’s no one way to do this, so here’s some suggestions.

Best practices for open data

* Machine readable (json)
* Interpreting datasets
  + Connection of variable name to description of what that is
  + Type of variable: factor, character, number, logical
  + Measurement of the variable: repeated or between
  + Level labels
  + Measurement scale

Best practices for options to creating open data

* KISS (easy to use)
* Give examples / tutorials of how to fill out the thing
* One big file instead of multiple files

Thoughts from Data Spice so far:

* Installing R, packages, coding things can be hard
* Files and where files should go and how to connect everything
* What if they wanted to add extra information – no place to go at the moment (labels for levels of a variable, for example)
* Purpose of the map ? why use it? How would you fill that part in if you are using turk – allow for usefulness but maybe not in that way
  + Where/how data was collected but not necessary for mapping
* Definition of what should be in each column
* What you should you include in each of these things – examples of what should go in those columns
* The practical part where the code doesn’t work

Sharing the data

* Json machine readable stacked code
* Not currently what data spice has
* HTML not a bad format – structured better
  + But OSF doesn’t render those – maybe we can ask them to
* PDF formatting is another alternative (copying might be a problem)
* Metadata csv or something importable into other programs

Three examples of data dictionaries

Do the things (closer paragraph about why it’s a good idea).

*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](mailto:ampps.editor@gmail.com) 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.