Toward a More Open, Trusted, and Efficient Research Environment

Robert K. Olendorf

Open science is becoming increasingly popular. Both funders and publishers require data be made public. The goal is to make research easier to validate, more trusted, and to hasten the speed of discovery. However, due to lack of training, lack of resources and lack of time, researchers often fail to make much of the content they generate public, and they also fail to adequately document and organize it. Here I make an argument that researchers should try to make all their research content public. I briefly describe best practices that should both result in a better product and be less burdensome on the researcher. I also argue that if done properly, opening up their research can have multiple benefits for the research and their career.

Keywords: Open Science; Reproducible Research; Open Data

Contact information: University Libraries, North Carolina State University, Campus Box 7114, 2800 Faucette Drive, Raleigh, NC 27695 USA; email: rkolendo@ncsu.edu

Open data and open research in general are becoming increasingly popular. Many funding agencies now require that the products of research, such as data, analysis code, and other artifacts of the funded research be made available after completion of the project. Many journals also require that the data used to reach the conclusions made in a manuscript be made available. Researchers are also becoming increasingly more open as well. Despite this, many researchers still resist being open. Additionally, much of the data that is deposited in data repositories lacks the appropriate organization, documentation, and metadata to be useful to others. Here I describe the advantages of open research, what open research should look like, and some strategies for incorporating open data into a researcher's daily workflows.

There are multiple motivations behind making research more open. Funders want greater accountability, transparency, and validation of the research they fund. Journals hope to ensure that the conclusions they publish are as accurate and reliable as possible. The motivations for researchers however are more nebulous, usually revolving around open research being good for science, or fulfilling the mandates of funders and publishers.

The FAIR principle suggests that data should be Findable, Accessible, Interoperable, and Reusable. Many funders and journals require that data be FAIR, and many data repositories claim that by depositing data in them makes them FAIR. However, there are no objective criteria by which to judge adherence to these principles. A quick review of the contents of many repositories, however, results in many datasets that are clearly not interoperable or reusable due to lack of documentation and organization. This is not to say that the data are bad, just that they cannot be easily

understood without significant input from the creator. Also, most scientists lack the training and skills needed to do this easily, and they are often overwhelmed and under motivated to spend significant amount of time documenting their data.

While repositories can certainly make some changes to facilitate better documentation and organization, the primary onus falls on the researchers themselves. This leaves researchers needing to know both what to save, and how to manage and document their research to best adhere to FAIR principles. Much of what follows is borrowed from common practices in the realm of open source software development, where standards and best practices are fairly well developed and have been successful in promoting a culture of relatively easy sharing of even complex development projects.

Deciding what to save can be an art; however, the default stance should be everything, including analysis code, raw data, even information about physical samples where possible. This may seem excessive, but to allow full validation of research, it is important for everything to be available. This also builds trust in the research in that clearly there is nothing to hide and shows the researcher's care and skill in managing the data. There are, however, often things that should not or cannot be shared. This includes personally identifying information and ecologically or culturally sensitive locations. Also, some things may need to be embargoed until the research is published or patents are obtained, for instance. Versioning tools such as Git can also be used to track and document changes to the data and other content.

Documenting and organizing the research often gets left to the end of the project. However, it should be built into the workflows from the start. Before any research is started, the team should start a README file, make plans for other documentation such as data dictionaries, and also decide on naming conventions and file structure. The team should regularly review the data and other content as well to ensure that it is well organized and documented.

Finally, we need to ask where is the benefit to researchers for doing this? Unfortunately, most researchers do not get credit for making their data available like they get credit for publishing their research in journals, proceedings, or books. However, there are still direct benefits to practicing open science in this way. First, treating one's data and other research products as if they are going to be open typically results in cleaner, more organized, and documented products. Also, this will help researchers spend less time with their data in the long run, better dealing with personnel changes as grad students and post-docs come and go. Additionally, for large projects with many collaborators, such techniques will facilitate easy sharing of content among personnel. It should also be noted that studies have documented that researchers who make their data open, are cited more often.

While the general trend towards openness in academic research is a good thing, the culture and practices around open research are still developing. Many researchers do make their data available, but typically they do not provide enough of the other products of research to fully realize their potential. By providing most if not all the products of their research, carefully organizing, documenting and versioning from the start, researchers can both get the most impact from their data, and also be more efficient in their own daily practice.