DATA MANAGEMENT FOR RESEARCHERS: ORGANIZE, MAINTAIN AND SHARE YOUR DATA FOR RESEARCH SUCCESS

Briney, K. 2015. Exeter, UK: Pelagic Publishing. 191 pp., 12 black and white illustrations. Hardcover: ISBN 978-178-78427-012-4, £49.99. Paperback: ISBN 978-1-78427-011-7, £24.99. E-book: ISBN 9781784270131, C$51.99

In light of recent reduced access to data from government agencies in various countries, this book is quite pertinent. It is a roadmap of how to preserve data forever, in a form that can be accessed in the future and shared by all. It begins by listing some striking examples of data loss: the original tapes from the Apollo 11 moon mission, with high-quality video footage of the moon landing and the first high-quality image of Earth from the moon. These were housed on magnetic tapes that were wiped and reused for data storage in the 1970s. Other NASA images had not been updated so that they could run on modern equipment. Specialized obsolete hardware had to be found to read the original magnetic tapes and decode the labeling scheme. Examples like this are enough to send any marine ornithologist into panic mode.

This book should be required reading for every graduate and undergraduate student, as well as for anyone doing research, no matter how far along you are in your career. I certainly wish that I had had this book at the outset of my career—it would have saved me countless hours of searching for or reorganizing data. The subtitle tells us to organize, maintain and share our data for research success. I would honestly say that most of us have succeeded despite not having followed the very clear steps for best data-handling and -storage for future retrieval outlined in this book.

I wrote notes in the margins alongside helpful passages that would remind me just what to do with data I already had regarding filing, storage for later use and accessibility to others. Everything that Dr. Briney said was applicable to my research, except for the sections meant for social or medical scientists who work with human subjects.

What Dr. Briney writes is so very obvious once she states it. But much of what she says about creating data in a form that can be easily accessed or shared simply is not considered by most researchers at the outset of their studies. She begins with an excellent roadmap to the lifecycle of data, starting with project planning and ending at long-term storage. In between, she covers data acquisition, analysis, and publication. To this familiar progression she adds data sharing, preservation, and reuse—more recent concepts that are becoming repeatedly common and necessary.

There is a good section on laboratory notebooks (read "field notebooks") and a detailed comparison between paper and electronic documentation. Field notebooks are fine if they are transcribed at the end of each day, but electronic notebooks obviously have advantages: they are searchable, can have data files embedded in them, and can be shared with team members. Drawbacks are obvious, too—you need a computer and a network to share, and electronic files are susceptible to bugs and to upgrades of software, and therefore, need constantly to be updated (remember the NASA data). This lack of updating was true for the very large data set gathered by scientists with the US Fish and Wildlife Service—and later by the US Geological Survey—in Alaska during years of the Outer Continental Shelf Environmental Assessment Program. The data manager left, and nobody took charge of the data that had been entered using a program that was obsolete by the 1990s. The original paper datasheets were put into a federal storage office and never seen again. These were valuable pre-Exxon Valdez oil-spill data that would have been very useful after the spill. To date, they are in electronic limbo. The bottom line for data entry and retention is to constantly upgrade the data so that it can run on the newest software upgrades. This task should become as routine as typing in your field notes daily. Those of us who still have 3¼ inch or 5½ inch floppies, take heed! University computer tech offices hoard hardware, as in the movie WALL-E; if you unfortunately have old floppies, they will be able to save your data on your newest computer, using ancient floppy readers from their shelves. However, you will be highly embarrassed, so to avoid this, Dr. Briney states, update everything now.

Detailed methods and README files are at the top of the list of how to begin data storage, including definitions of fields and what your abbreviations or codes mean. Not performing this simple task may make you kick yourself when you don't recall codes in your notes like "which bird species was 10100101" (Tufted Puffin in the 1980s). The documentation chapter gets into much more detail regarding metadata syntax and other aspects, and should be very helpful for those with large data sets.

The organization chapter is one of the best. It leads with an example of a paper by some eminent Swiss scientists that was retracted because they came to an erroneous conclusion by analyzing the wrong data set. They had accidentally misnamed the file of that set. Dr. Briney lists several ways folders can be organized, but leaves it up to the reader how to do it. She compares indexes with tables of contents (indexes are better for data retrieval), and raises points of what kinds of things to consider in the ordering of information. Again, she emphasizes updating the index at the end of each day. So often we let this part slide and then must figure out weeks later what we did before we can start to index. This takes up more time than if we had indexed our work immediately. Dr. Briney is always pointing out the need to maintain your data such that collaborators or other future researchers could easily access it if you were not there to help them.

The short but precise chapter on data analysis is a stepwise model of how to go about it; it is not how to analyze your data. Steps like "Documenting the analysis process," "Preparing for data analysis," "Data quality control," and "Error checking" are sometimes overlooked in the rush to get on with analyzing one's data. Dr. Briney's advice is to slow down. The one point that is left out in this chapter is the use of relational databases and why they are better than spreadsheets. Having used both, I can unequivocally say that a relational database is more flexible and will give researchers output they want quickly and with ease. Relational databases also have input error checking, so making transcript mistakes are no longer a problem.

The final two chapters—Storage and Backup, and Long-Term Storage Solutions—are perhaps the last thought in eager researchers' minds, when all they want to do is gather the data and then analyze it to see if it supports the hypothesis. Not thinking this through is a dangerous thing, and Dr. Briney gives another horror story of a University of Cambridge PhD student whose laptop and external hard backup drives were stolen and who thus lost all hope of completing his degree. At my university, theft of laptops has become a regular thing. We get regular reminders from the IT department to secure our doors and desks, and to use a backup system. Dr. Briney reviews various systems as well as Cloud storage options, and urges researchers to compare thoroughly their differences, especially the weak points such as capacities, ease of restoration, data compression and integrity, and the like.

The book finishes with a section on data sharing and what kind of repositories are best. If you think your data are worthwhile for others to have and use, now and in the future, then this chapter is definitely worth the read.

Grab this book immediately and start to clean up your files. You will not regret it.

Pat Baird, Simon Fraser University, Burnaby, BC, Canada, pab7@sfu.ca