Published online June 8, 2017

# DATA MANAGEMENT

## in the agricultural sciences

by Martín Battaglia, Wade Thomason, and John Fike

## Members Forum

Research environments are becoming increasingly more complex in the 21st century, with a concomitant upsurge in the amount of data produced. Along with collecting and managing large, complex data sets, graduate students must be aware of the most common reasons for data loss and the strategies to prevent them.

### The State of Scientific Data Management

In November 2012, Doucette and Fyfe (2013) conducted an online survey of more than 350 master's and doctoral students in six subject areas (i.e., Geography, Psychology, Sociology, Chemistry, Physics, and Earth Sciences) from nine different research universities in Canada. This work assessed graduate students' behavior, attitudes, and education related to managing research data. Some of the most remarkable outcomes from the survey were as follows:

- 14% had "re-collected data that had been previously collected because [they] could not find or open the file."
- 17% had "lost a file and been unable to re-collect the data."
- 40% were unsure, disagreed, or strongly disagreed with the statement, "I have provided enough documentation that a research peer or future grad student could use my data."
- The majority of students agreed or strongly agreed that: (a) there is a value in reusing or repurposing their research data in the future (74%); (b) management of research data is important for a research group (83%); and (c) they were confident in their abilities to manage research data (90%). However, almost 38% of the students did not have written or verbal policies related to research data management (RDM) within their research group. .
- The majority of respondents who had re-collected data or lost a file also agreed or strongly agreed with the statements related to importance (81%) and confidence in abilities (77%) related to RDM.

Though this study did not include students in the agronomic sciences, we may speculate that these values are similar, or even greater in our disciplines, given the typical need to collect data in both field and laboratory settings. Re-collecting data is expensive and time consuming and requires duplication of effort by the research group. For field-based science, the consequence may be greater since many times we are unable to re-collect the data at the right time in the crop growing season.

### Data Management Plans

Michener (2015) defines a data management plan (DMP) as "a document that describes how you will treat your data during a project and what happens with the data after the project ends." Doucette and Fyfe (2013) define RDM as the "activities that fall outside of the work of creation and analysis of data." Jahnke et al. (2012) employs the term "life cycle data management" to include activities ranging from RDM planning, through collection, identification, processing, and accession of data sets, to the final archival preservation and sharing of data in an appropriate repository. Different titles aside, these definitions meet at one common point: they all reinforce the need for a holistic data management framework from project inception stage until the end of a research project.

A good research DMP has several benefits, including:

- Complete and reliable records of previous research allow researchers to focus on the advancement of future work by building upon earlier efforts. As a consequence, an "exponential growth in knowledge" is possible (Howe et al., 2008).
- Avoiding digital archeology makes researchers more efficient (Valentino and Boock, 2015).
- Appropriate data curation enables processing of a greater amount of more complex data more quickly and efficiently.
- Organized work makes searching (and retrieving) files easier anytime the data are needed.
- Research data reproducibility is improved, enabling the matching of outputs with the exact inputs and transformations that generated them.
- Organized data can be used as an efficient quality control source.
- Frequent backup copies of research data ensures that data remain available even after data/file loss.

## Good Data Management Practices

- Organize data (using version control along with naming conventions that are consistent from the beginning through the duration of the research).
- Choose data and file formats that are not likely to become obsolete in the in the short term. Nonproprietary software sources that include formats like Comma Separated Values (CSV) over Excel, widely used and accepted by the scientific community, represent good and reliable sources to this end. Moreover, data are more likely to be accessible for the long term if they are uncompressed, unencrypted, and stored using standard character encodings such as UTF-16 (Michener, 2015). This system provides an established standard method for encoding more than 1.1 million possible characters into their equivalent binary values.
- Create multiple backup copies of your data as soon as you can and in at least two geographically distributed locations throughout all stages of the data life cycle. For example, you can decide to back up and store your data in one computer, one external drive, and an online server (the cloud) (Michener, 2015). At the same time, it is always wise to adopt a regular schedule for backing up your data (e.g., like Friday before leaving the lab!). You may also decide to back up your data by photocopying or taking photographs of datasheets. Old school vs. new school: all systems are valid and always represent a better option than doing nothing.
- Do not alter your raw data: if changes are needed, create a copy and alter that copy.
- Let a librarian be part of your research. Spend a few hours, at least once, with a librarian to talk about best practices to manage your research data. In the study conducted by Doucette and Fyfe (2013), less than 5% of the surveyed students had done so. You may be surprised by the insights received from a professional whose work is not directly related to your field of study.
- Take at least one research methods course.

## Summary

The correct organization, protection, preservation, and sharing of research (a.k.a. "proper data management") is essential. Proper data management ensures high research productivity, is often a requirement to secure grant funding, enables collaboration, and facilitates future data use. Good data management is a habit to cultivate and will result in greater efficiency and fewer frustrating hours searching through datasheets or field notes.

*M.L. Battaglia, W.E. Thomason, and J.H. Fike, Department of Crop & Soil Environmental Sciences, Virginia Tech, Blacksburg*

## References

Doucette, L., and B. Fyfe. 2013. Drowning in research data: addressing data management literacy of graduate students. Paper presented at ACRL 2013, Indianapolis, IN, 10–13 April.

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide et al. 2008. Big data: The future of biocuration. Nature 455:47–50.

Jahnke, L., A. Asher, and S.D.C. Keralis. 2012. The problem of data. Council on Library and Information Resources, Pub. 154, August 2012. http://bit.ly/2rReqZu.

Michener, W.K. 2015. Ten simple rules for creating a good data management plan. PLoS Comput. Biol. 11(10):e1004525. doi:10.1371/journal.pcbi.1004525

Valentino, M., and M. Boock. 2015. Data management for graduate students: a case study at Oregon State University. Practical Academic Librarianship 5:77–91.

doi:10.2134/csa2017.62.0701