Tips for effectively managing your data

By Emma Grace Matcham Lauren E. Schwarck

March 19, 2021

The organization and safe storage of data is critical to effective sharing and publication of research. Data organization decisions begin even before data are collected when making decisions on experimental design and the research questions to be addressed. Design your experiments with care, collect the measurements that best answer your research questions, and keep your data organized to save yourself time and frustration during data analysis. Following are three tips to help you approach data management thoughtfully throughout the research process.

3-2-1 Data Backups

Always have a backup! A common recommendation is to have at least three copies of your data on two different types of storage and one physically distant location. You could implement this strategy by keeping one working copy of your data on your laptop, another on an external hard drive at your desk, and a third copy on a cloud

storage account or other remote server. When using online data entry tools (Microsoft Online, Google Sheets, etc.), it's a good practice to store a copy of the file locally at the end of every workday. Even when taking physical notes on datasheets or in a lab notebook, take photos of your work as a backup, and store your photos to two different locations.

File-Naming Protocols

Once you have copies of your data in various places, adopting a smart file-naming protocol enables you to search for documents, keep track of versions, and facilitate document sharing. Descriptive file names that include information such as, the growing season, project name, and a brief description of the data can make it easier to pair files from different sampling types and timepoints, but these file names can get very lengthy. One option is to start each file with a number corresponding to the growing season and project, such as "2105" for project 5 in 2021. Note the use of leading zeros—05 instead of just 5. This helps sort projects by number without project 10 coming before project 5.

When you're working on a file that will have multiple versions over time, you can either number or date your file versions. Leading zeros on version numbers or international date formats, such as YYYY-MM-DD or YYMMDD, help keep your data easy to sort.

Although commonly used in file names, punctuation should be avoided to minimize errors across different operating systems. If punctuation is necessary, underscores generally work well, but periods and other special characters can cause issues.

Investing time in choosing a file-naming system at the beginning of a project can save a lot of time in the long run. But, it's not a permanent decision. If your initial file-naming protocol isn't working for you, you can always try a different system for future projects.

Column Naming and Metadata

The data management step that arguably saves the most time during analysis is a consistent column–naming scheme. A good column name will be readable for both you and your computer, making it easier to call variables when you're analyzing data or making figures in the future. Consistent column naming also improves your ability to merge data stored in different spreadsheets.

Most analysis software is case sensitive, so consistently using either snake case (example: days_after_emergence) or camel case (example: DaysAfterEmergence) is a great way to improve the uniformity and readability of your data. Also consider including units in the column name (example: precipitation_mm) as this can prevent future confusion if your dataset were to be used by someone else. Avoid column names that start with a number and any column names that include spaces or special characters.

If your column names aren't easily interpretable on their own (or even if they are), you may want to make a separate sheet of metadata that defines each column name more fully. Helpful things to include in metadata include units, data source, and the data type (example: string, integer, etc.).

While we're on the topic of column naming, most software makes it easiest to analyze data that is formatted on a single tab of a single spreadsheet with column names in exactly one header row. When in doubt, avoid creating another table within a spreadsheet file or another tab in an Excel or Google Sheets file to prevent yourself from having to later combine the data across all of the tables.

While there are exceptions to every rule, we find the above tips to be a helpful starting point for effectively managing various types of data files. Planning your file and data management early in the research process is a good way to save time and reduce

frustration later on. Managing data takes time and practice to find your system and maintain consistency. Don't overwhelm yourself by implementing all of these strategies at once. Try incorporating a few new strategies into your system, and see what works for you!

More careers & education

Back to issue

Back to home