

Organizing Data

Organizing data involves ensuring that you can find your data and other research materials (including documentation, code, and physical samples) when you need to, and ensuring that data and materials that go together are connected in a meaningful way.

What does it mean to organize data?

Organizing data means arranging your data and other research materials so they can be found—by yourself and by others—as needed. Here are four factors to consider when organizing data. Remember: you can't use data you can't find.

NAMES

Data should be labeled using a consistent and descriptive file naming system. Your system should allow you to immediately and uniquely identify the contents of your files.

STRUCTURES

Data should be organized with a consistent and easy to navigate file structure. Maintaining such a structure can help reduce the risk of data loss and unnecessary replication.

CONNECTIONS

Connections give context. Data and other materials should be organized in a manner that emphasizes the links between them. This may refer to different versions of the same file or different files related to the same aim or project.

DOCUMENTATION

You should document how you organize your data and other research materials and refer back to and update your documentation often. When thinking through how to organize your files, make sure you also consider how you include all of the related description and documentation (e.g. notes, data dictionaries, metadata).

Requirements and how to meet them

There are specific requirements about how certain types of data should be organized. Under most circumstances, data containing sensitive or potentially identifying information should be stored separately from data that does not. However, whenever possible, you should apply the same organizational principles to both.

Things to think about

- You should document your file naming and structuring schemes. Such documentation may take the form of a data dictionary or README file and should enable somebody other than you to understand how your research materials are organized.
- The size and content of your data will determine the degree of flexibility you have about keeping it organized. It is very likely that your organizational scheme will not be perfect. There may be times when you'll need to rearrange your files.
- Versioning your data may be a good way to keep it organized, as long as it is done in a consistent and descriptive manner. Data_v2.csv may be informative, Data_NewEdits is less so.
- These principles (naming, hierarchies, linking, and documentation) also apply within data files. For example, variable names within a file should be consistent and descriptive. You should maintain documentation about what they refer to.