SUPPORT YOUR DATA | Refine Your Approach to Managing Data

Planning for Data

A plan detailing how you'll manage your data, code, and other research materials (including documentation, code, and physical samples) over the course of a project will help your research proceed efficiently. Creating a comprehensive, specific, and instructive plan for your data is an important step in developing a new research project, but the best plans also evolve as a project proceeds.

What does it mean to plan for data?

Planning for data means thinking through and documenting how data and other materials will be organized, saved, prepared, analyzed, and shared over the course of your research project.

Requirements and how to meet them

Many funding agencies and institutions now require that researchers compose a short document called a Data Management Plan (DMP). A DMP provides details about the type of data to be collected and managed within a research project. It also documents the individuals responsible for managing the data, how and where data will be archived and shared, and how the financial cost of managing data will be met.

The DMPTool (https://dmptool.org/) is a free tool that provides guidance for creating a data management plan.

Things to think about

- Planning for data is not a one-time activity. You should create a plan as you develop a project, but you should also revisit and revise your plan as your project proceeds.
- Plans should identify the data you intend to collect, as well as how you plan to transform, analyze, and share it. Be as specific as possible.
- A plan is really only useful if people know about it and can follow it. Be sure your plans are communicated to your collaborators.
- Even if you do not have a Data Management Plan (DMP), you may have a document that describes how you plan to handle your data. For example, this information could be included in a study protocol or an IRB proposal.

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.

SUPPORT YOUR DATA | Refine Your Approach to Managing Data

Saving and Backing Up Data

There is more to saving data than ensuring you have appropriate backups. How and where you save your data and other materials depends on their size, format, and content, as well as your intentions about making them available at the conclusion of your research project.

What does it mean to save data?

Saving data means storing research materials so that they can be accessed and used – by yourself or by others – at a later date. Here are three factors to consider when saving your data.

LOCATION

When possible, save multiple copies of your data across a variety of storage mediums. Hard drives, cloud storage, and other options have different levels of reliability, but all will eventually fail or become obsolete.

TIME

Saving data takes time, but losing data wastes more time. Backing up data should be a regular part of your research practice and you should also have a plan for how data will be saved after your research is concluded.

FORMAT

Data should be saved in a format that enables later use. This may involve saving data in open or easily accessible file formats, or simply storing your data alongside the documentation and other research materials needed to make use of it.

Requirements and how to meet them

There are specific requirements about how and where data containing sensitive or personally identifying information can be saved. How you deal with sensitive data will depend on a number of factors including the size and contents of your data as well as the resources available to you.

Things to think about

- The characteristics of your data determine how much flexibility you will have about how and where it can be saved. If you have large quantities of data or data containing sensitive information, it can be challenging to move it from one medium to another.
- Saving data should also involve saving research materials (e.g. documentation, code, etc.) needed to make sense of or use that data.
- There may be a difference between where and how you save your data as you work on it and where and how you save your data over the longer term. Consider the difference between regularly backing up your data and archiving it at the end of a project.

SUPPORT YOUR DATA | Refine Your Approach to Managing Data

Preparing for Analysis

It is very likely that there are several steps between the data you collect and the data you ultimately examine, analyze, and publish. Properly preparing data involves both ensuring that your data exists in a form ready for examination or analysis, and ensuring that you have documented how and why you prepared your data in the manner that you have. This is where you need to think about what you planned and address the reality about what you can or need to do.

What does it mean to prepare data?

Preparing data means cleaning, coding, processing, or otherwise transforming it in some way. While doing this, it is important to document what you've done so that your steps can be re-traced – by yourself or by others – in the future. Remember, documentation about your data is part of your data.

Requirements and how to meet them

Your research community, institution, or research group (e.g. lab) may have specific standards and requirements about how you should prepare your data and document your activities.

If you are unsure about what procedures apply to your data, check against your data management plan, your research group's existing protocols and practices, and any requirements set forth by the places you want to use to share or publish your work.

Things to think about

- Whenever possible, maintain a copy of your data in its original form. The link between the original and prepared data should be clear. If you generate new things, they should fit into your existing schemes for organizing and saving.
- Whenever possible, save any intermediate steps as you prepare your data. This will make it easier to trace back to what you did last. Doing this can be as simple as assigning different file names to different steps or as advanced as incorporating a version control system.
- Preparing data may affect risks related to sensitive data or personally identifying data. You need to be aware of this, but it should not affect the degree to which you document your procedures.
- Don't make assumptions. Even if you automate your preparation, you may still want to do manual quality assurance checks. Even if your decisions seem obvious, you should still document what you did and why.

SUPPORT YOUR DATA | Refine Your Approach to Managing Data

Analyzing Data

There is more to analyzing your data than running statistical tests, summarizing comparisons, and creating visualizations. Analyzing your data also involves ensuring that a future researcher (who may or may not be you) can understand and potentially replicate your analyses.

What does it mean to analyze data?

The methods you use to draw conclusions from your data will, of course, depend on your research questions, your field of research, and the tools you have available to you. However, here are two factors to consider when analyzing your data.

DOCUMENTING ANALYSIS DECISIONS

You should be as transparent as possible about how and why you conducted your specific analyses.

MANAGING ANALYTICAL OUTPUTS

If your analyses generate additional outputs (documents, images, etc.), you should organize and save them as if they were any other research product.

Requirements and how to meet them

Your research group, field of research, or institution may have a set of standards or best practices related to how data should be analyzed and how analytical outputs should be managed. These may be as simple as a set of guidelines about how procedures, parameters, or protocols should be documented.

If you are unsure about the specific requirements that apply to you, try to think about documenting your analyses and managing your outputs as “showing your work.”

Things to think about

- Properly documenting and managing your analyses is important for reasons related to research transparency and reproducibility. However, they will also help prevent you from wasting time and losing data.
- While many best practice recommendations apply mostly to analyses underlying a scholarly publication, you should apply the same procedures to all of the analyses you conduct, no matter the outcome.

Sharing and Publishing Data

Sharing data is more involved than simply uploading files somewhere for other researchers to find. The methods you use to share your data will depend on a number of factors including the size and content of your data, mandates from the entities that fund and publish your research, and any assumptions and requirements related to future use. If you make your data and other materials available, you should make sure that other researchers can find and use them.

What does it mean to share data?

Sharing data means making your data available so that they can be accessed and used—by yourself or by others—in the future. Here are three factors to consider when sharing data.

FORMAT

Data should be shared in a usable format. This may mean sharing raw data instead of prepared data (or vice versa) or ensuring that data is saved in common or open file formats.

COMPLETENESS

Remember that notes, documentation, and other information about your data are part of your data. To ensure that your shared data is useful, make sure these elements are included.

LOCATION

When choosing a method for sharing your data, consider how other researchers will find and use it. The storage options you use to save your data as you work on it will probably be different than the options you use to share it, especially over the longer term.

Requirements and how to meet them

Many research funders, publishers, institutions, and research communities have formal expectations about how data should be shared.

Things to think about

- Though it is very likely that you'll share your data only at the conclusion of a research project, data sharing should be incorporated into your data management practices from the beginning.
- Data sharing is about showing your work. Though many current data sharing requirements focus on the data underlying journal articles and other scholarly works, you should be prepared to share all of your data. All of it has potential value.
- There are limits on how data containing sensitive or personally identifying information can be shared, but you should be prepared to share enough information about your work so that others can evaluate, potentially replicate, and otherwise make use of what you've done.