Data Organization

Structuring your data well enables you to:

-Reproduce results

-Reuse it in the future

-Share it with others

-Gain and retain credibility

-Comply with IRB/funder requirements

> The decisions you make about how you organize and structure your data today will have implications for how you and others can access and make use (or sense!) of that data in the future.

Context and Data Documentation:

Include the following in a readme text file:

1. The data's purpose
2. A list of the files in your data package
3. Data dictionary listing and describing all variables

Data Organization Principles:

1. Use one variable per column
2. Make one observation per row
3. Use human-readable column name
4. Include one table per tab
5. Include an ID or key to indicate any relationship between tables

Whether your data is organized in lists, arrays, hash sets, dictionaries, queues, trees, heaps, or relational databases, it is important to be aware of disciplinary norms, as well as both institutional and funder requirements, that will make its deposit, storage, and long-term support more likely. Increasingly, the path for long-term support involves taking steps to make sure your data is deposited alongside data collected by others in your field or discipline.

Questions to consider for any data project:

1. What are the data organization standards for your field?
2. What are the data export options for your software?
3. What forms of the data will be needed for future access?

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle. https://www.dataone.org/best-practices

Do:

-Consider what your NULL values are and how they are represented

-Consider whether a more robust data dictionary is required

-Use standard data representation (e.g., (YYYYMMDD for dates)

Do Not:

-Use formatting to convey information

-Place comments in cells

-Use special characters in field names

-Use blank spaces or symbols in column names

Discipline-based data repository examples:

-Social Sciences: ICPSR

-Genomics: GenBank

-Earth Sciences: NASA's Earthdata

-Archaeology: tDAR

-Oceanography: NODC

-BioSciences: Dryad

Source - Guidance Briefs: Managing Your ETD Research Files