Understand the data (or try to)

In this step, examine the dataset closely to understand what it is, how the files interrelate, and what information is needed to reuse. Common UNDERSTAND steps include:

- Check for quality assurance and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns
- Try to detect and extract any "hidden documentation" inherent to the data files that may facilitate reuse or expose unintended information
- Determine if the documentation of the data is sufficient for a user with similar qualifications to the researcher's to understand and reuse the data. If not, recommend or create additional documentation (e.g., a readme.txt template)

Key Ethical Considerations

- If working with human data, is this research done *with* and not *on* communities and populations involved?
- Are there labels or other descriptive indicators that could be applied to better represent or protect an identified group of people impacted by this dataset? (Example: TK labels)

Essential Tasks

- ☐ Examine files, organization, and documentation more thoroughly. Are there changes that could enhance the dataset?
  - ☐ Are there missing data?
  - ☐ Could a user with similar qualifications to the author's understand and reuse these data and reproduce the results?
  - ☐ Are the data, documentation and/or metadata presented in a way that aids in interpretation? (e.g., readme Example)
  - ☐ Is the context of the data explained? (Methodology information, relevant citations, file relationships, etc.)
  - ☐ Is the content of the data explained? (variable and value labels, units of measurement, etc.)
  - ☐ Are there file references/links to other files in the package (are all the files

there? Correctly referenced?)

- ☐ Is there additional documentation that may be helpful based on the data type? (e.g., codebook, data dictionary, study protocol, survey questionnaires (s), etc.)

Tasks vary based on file formats and subject domain. Sample tasks based on format:

Tabular and text data questions:

- ☐ Check the organization of the data–is it well-structured?
- ☐ Are headers/codes clearly defined?
- ☐ Is quality control clearly defined?
- ☐ Is methodology clear and sufficient?

Scientific image(s) questions:

- ☐ Is proprietary software needed to view or manipulate the images?
- ☐ Is resolution high enough to be interpreted or useful?
- ☐ How do the image files relate to each other?
- ☐ How do the image files relate to other files in this dataset?

To view additional UNDERSTAND steps based on format, view the following primers:

- ISO Images Primer
- Confocal Microscopy Image Primer
- GeoTIFF Primer
- netCDF Primer and Tutorial using NCAR dataset
- Neuroimagining and NICOM NIfTI Primer