Data Management Plan

Samantha Estrada PhD
Assistant Professor & ORSSP Data Analysis Lab Consultant
A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

Stanford University Data Management Plan (n.d.)
Why should I have a DMP?

Why should I care about DMP?

- Funder’s/Grants mandates
- Comply with data protection law and protect data subjects
- Makes research reproducible
- Makes research more efficient
Example Grant
Institute of Education Sciences (IES) RFP

(3) Data Management Plan

Applications under the Efficacy and Follow-up goal must include a Data Management Plan (DMP) placed in Appendix F. Your DMP (recommended length: no more than 5 pages) describes your plans for registering the study and making the final research data accessible to others. Applications that do not contain a DMP will be deemed nonresponsive to the Request for Applications and will not be accepted for review. Resources that may be of interest to researchers in developing a data management plan can be found at http://ies.ed.gov/funding/researchaccess.asp.

DMPs are expected to differ depending on the nature of the project and the data collected. By addressing the items identified below, your DMP describes how you will meet the requirements of the Institute’s policy for data sharing. The DMP should include the following:

- Plan for pre-registering the study in an education repository (e.g., see the SREE Registry of Efficacy and Effectiveness Studies https://www.sree.org/pages/registry.php).
Scenario

Think about the most recent data you’ve collected. How much work would it take for you share it with someone?
Before you start

- Know your target population.
  - Horror stories
    - Filtering out data.
    - Incentives
- The research questions that you’re going answer are dependent on collecting good data.
  - Collecting unnecessary data.
  - Collecting data you already have.
    - We have workshops for that!
  - Not collecting the data you actually need.
    - You can’t answer a question you haven’t measured on your participants.
Before you start

• Be mindful when selecting your measures.
  – Think about length.
  – Ideally, you can pilot your instruments to make sure your instructions are clear and free of errors.
  – Reliability is sample specific.
  – Think about accessibility.
    • Mixed Mode data collection
Before you start

• Important to have an analysis in mind.
  – This analysis should be guided by your research question.
  – The types of variables collected (categorical vs. continuous)
  – You can create pilot data to see how to carry out the analysis.
  – You might want to take workshops to learn more about a specific analysis
    • We have workshops for that!
Open Ended Questions.
Open Ended Questions..

• Be realistic.

• Do you have a research team that can do it?
  – Horror stories
  – Common misconceptions about qualitative software
Research Notebook

• Writing everything down.
• Keeping a “Research Notebook”
  – Can be digital
  – Can be a real notebook.
• Works great when you have many team members
  – Print out code.
  – Write out decisions made.
  – Horror story. Duplicates.
Research Notebook..

• Issue: missing data, duplicates, removing cases.
• Date
• Who took the decision, if there was more than one person involved.
  – If its possible to go back to the data before this step provide this information
• Related protocols
• What was the decision made and what was the rationale for it.
## File Names

<table>
<thead>
<tr>
<th>Best Practice</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 characters</td>
<td>32characters32CharactersLooksExactlyLikeThis.csv</td>
</tr>
<tr>
<td>No spaces, no . or -</td>
<td>✗ My-proj.1.csv</td>
</tr>
<tr>
<td>Use versioning</td>
<td>✓ ProjID_v02.txt</td>
</tr>
</tbody>
</table>
| Don’t use generic data file names that may conflict when moved from one location to another | ✗ MyData.csv  
|                                                          | ✓ ProjID_date.csv                           |
| Use leading zeros in sequential numbering to allow for multi-digit versions: | ✗ ProjID_1.csv ProjID_11.csv  
| • For a sequence of 1-99: 01-99  
| • For a sequence of 1-999: 001-099-999                 | ✓ ProjID_01.csv ProjID_11.csv              |
Variable Names & Subject IDs

• The shortened digital representation of each variable in your dataset
  – Usually columns in the dataset.

• Important that all variable names are UNIQUE.
  – Linking variables requires same subject IDs.
    • Horror stories.

• Qualtrics question numbering
  – We have workshops for that!

• Recommendation. Create a rule for naming convention of variables.
  Document it.
Variable Names Do!

TIME

CSSE1001

SURVEY NAME

UNIQUE IDENTIFIER
Long vs Wide

This is easy to fix, but it's even easier when the naming conventions make sense.
• When collecting online the issue is merging.
Variable Names Don’t!

<table>
<thead>
<tr>
<th>CASE #</th>
<th>Gender</th>
<th>$AMOUNT</th>
<th>10 SAS</th>
<th>SAS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>001-12-PRE</td>
<td>1</td>
<td>120</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>001-12-EXIT</td>
<td>2</td>
<td>100</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>004-12-PRE</td>
<td>3</td>
<td>122</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>004-12-EXIT</td>
<td>1</td>
<td>123</td>
<td>4</td>
<td>NA</td>
</tr>
<tr>
<td>005-12-PRE</td>
<td>2</td>
<td>102</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>005-12-EXIT</td>
<td>1</td>
<td>131</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>001-12-1</td>
<td>2</td>
<td>131</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>001-12-2</td>
<td>2</td>
<td>141</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
• You want to be inclusive in your language.
  – If coding gender always use 0 for the same category.
Code or Syntax

```
title: "Prob Data: Data Log"
author: "Samantha Estrada PhD"
date: "9/17/2021"
output: pdf_document

* I edited the variable names. Don't start with #, or _ it confuses software. This was done in jamovi
* Saved as csv next, to import to R.
* Deleted the -PRE and -EXIT it was confusing software. To keep track of pre and post use TIME variable
* MISSING and N/A changed to NA
* Deleted columns of zeros at the bottom of the data. Not sure why they got calculated as zeros for the totals. Let's check the compute function at some point.
* In TIME column replaced 0 with 1. Again to not confuse the software. This should just leave the TIME 2 variables empty.
* TIME coding changed to 1 = TIME 1 and 2 = TIME 2 to make the new data clear as to what time they were from.
```
• Even SPSS has syntax you can save.
  – This can help you recreate the analysis even if you choose the point-and-click method.
Data Storage and Back-up

• Multiple copies of the data.
  – Have a local working copy (e.g. on your workstation or in shared workspace).
  – Local/external copy in a different location (e.g. external hard drive).
  – Be careful what you store in the “cloud” make sure IRB approves.
Data Sharing

• Decide when you will share your data.
• This may be decided for you by grants though.
• Look for data repositories used by your research community or your host institution (see www.re3data.org for examples).
• Provide metadata that allows others to understand, cite and reuse your data files.
  – Title, creator, identifier, dates, etc
• Best practices for metadata (https://data.research.cornell.edu/content/readme)
Data Sharing...

• Sometimes journals will have their own repositories.
  – https://osf.io
  – https://figshare.com
  – Do you know any field specific?

• A great resource for DMP:
  – https://dmptool.org/ *(DMPTool (n.d.))*
Tell me your horror stories?
Resources

• Research Design & Data Analysis Lab: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/

• Schedule a consultant appointment with me: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-data-analysis-lab-consultants/

• Check out Lab Resources (including recording of this webinar): https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/resources/