# Project Mnemonic

“The first step in naming files and directories is to pick a short mnemonic for your project… When choosing a mnemonic, pick a string that is short because you do not want your names to get too long. Avoid mnemonics that are commonly found in other contexts or as part of words.” (Long 2009 p. 20-21)

TMPWF is an example of a project mnemonic, which stands for Template Workflow (TMPWF).

# Project Title[[1]](#footnote-1)

Descriptive titles typically contain the time period(s) and geographic location(s) the data cover.

# Project Research Question:

# Project Description:

## Project Members (Principal Investigators):

The name and affiliation of the Principal Investigator(s), in order of importance to the study.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Email** | **Affiliation** | **Role** |
| Nathanael Rosenheim | nrosenheim@arch.tamu.edu | Texas A&M University | Project lead |

## Summary

A full description or abstract of the study's subject matter or intellectual content. Your main goal should be to give the user a clear sense of what the study is about. The focus should be on describing the purpose of the study, the major topics covered, and what questions the PIs attempted to answer when they conducted the study. Note that you should avoid attempting to address issues of how the data might be used, who might be interested in the data, or any evaluative comments about the worth or usefulness of the study.

### Example 1

In addition to the standard personal characteristic items, the survey covers items viewed by the NORC staff and an advisory panel of sociologists as "mainstream" interests of modern academic sociology. The interview covers the areas of stratification, the family, race relations, social control, civil liberties, and morale. A major objective of the project was the replication of questions which have appeared in previous national surveys. The data were collected by the National Opinion Research Center as the first in a five year series of general social surveys. The survey was administered in February-April 1972 to a national cross-section sample of adults 18 years of age and older. The data were obtained from the Roper Center for Public Opinion Research.

### Example 2

The ANES 2010 Panel Recontact Study is a reinterview of the ANES 2008-2009 Panel Study panelists. Those who previously completed at least one ANES wave of the Panel Study before November 2008 and who also completed the November 2008 (post-election) wave were invited to complete a follow-up interview in June 2010. Data collection ended in July 2010. The study was conducted entirely on the Internet from a sample selected and recruited by telephone. It represents United States citizens aged 18 years or older as of election day in November 2008. The questions on the recontact survey covered numerous topics. Many questions were previously asked on earlier waves of the ANES 2008-2009 Panel Study. Topics included interest in politics, cosmopolitanism, efficacy, trust in government, divided government, attitudes toward parties, personality, economic peril, race discrimination, numerous policy attitudes, and income inequality. See the questionnaire in the user guide for question wording. Demographic variables include respondent income, political party affiliation, religiosity, employment status, and household income.

### Example 3

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-1995 school year. The Add Health cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when the sample was aged 24-32. Add Health combines longitudinal survey data on respondents' social, economic, psychological, and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships, providing unique opportunities to study how social environments and behaviors in adolescence are linked to health and achievement outcomes in young adulthood. The fourth wave of interviews expanded the collection of biological data in Add Health to understand the social, behavioral, and biological linkages in health trajectories as the Add Health cohort ages through adulthood. The files contained in this component of the Add Health restricted data include the type of medication used by participants during Wave IV.

## Funding Sources:

# The name of the organization(s) that supported the production and/or archiving of these data. You can also add an associated Grant Number(s). Separate multiple Grant Numbers with a comma.

* Example of a Funding Source:
  + United States Department of Justice. Bureau of Justice Statistics
* Examples of Funding Sources with a corresponding Grant Number:
  + National Science Foundation. Social Sciences Division
    - NSF-11789-S4037
  + Robert Wood Foundation
    - 40791
* Example of one Funding Source with multiple grants:
  + National Science Foundation. Social Sciences Division
    - NSF-36984, N3-277-1988

## Scope of Project:

### Subject Terms

Enter relevant social science subject terms that capture the essence of your data collection.

### Geographic Coverage

Enter geographic area(s) covered in the data collection.

### Time Period(s)

The time period to which the data refer. Include the range(s) of dates.

### Collection Dates

Date(s) the data were collected. Provide specific dates and ranges, e.g., YYYY/MM/DD-YYYY/MM/DD.

### Universe

The group of persons or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any of a number of factors may be involved, such as sex, race, income, veteran status, criminal convictions, etc. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, etc. In general, it should be possible to tell from the description of the universe whether a given individual or element (hypothetical or real) is a member of the population under study.

#### Example 1

Adult noninstitutionalized population of the United States living in households.

#### Example 2

Lawyers in the United States in 1984.

#### Example 3

Persons aged 18 and over living in New York with telephones.

### Collection Notes

A description of technical details and other characteristics of the data collection (such as unique authoring, dissemination, or processing information) that cannot be recorded in the other metadata fields but constitute important information for the user.

## Methodology

### Response Rate

The percentage of respondents in the sample who participated in the study.

#### Example 1

For Part 1 (Law Enforcement Interview Quantitative Data), 292 individuals were contacted for the telephone survey. Of these contacts, 82 were ineligible to participate because they reported having no familiarity or experience working on trafficking cases. There were a total of 121 completed surveys and 89 non-responses (11 refusals and 78 non-contacts). The response rate for the telephone surveys was 58 percent. For Part 2 (Case File Review Qualitative Data) response rates are not applicable.

### Example 2

Of the 2,265 programs initially contacted, 543 responded (24 percent). The final sample of 260 communities was then selected from this pool of 543. In each of these 260 communities, surveys were administered to a batterer program, court, and victim assistance agency. Repeat followup efforts culminated in high final response rates of 75 percent for the batterer programs (n = 195), 53 percent for the courts (n = 139), and 62 percent for the victim assistance agencies (n = 162). There was at least one response from 94 percent of the target communities, and at least one response from either the batterer program or the court in 88 percent of the communities.

## Sampling

Describes how the cases that appear in the study were selected (e.g., random, convenience, snowball). The sample is a selection out of the universe of all possible relevant cases (e.g. adults in the United States, housing units in three counties of Michigan, etc.) that could have been included in the study. A detailed discussion of such things as sampling error or other limitations of the sampling methodology is not required here. If the sampling description is available in a printed publication, provide the complete citation. Note that some studies, such as censuses, do not utilize samples but include all members of the universe.

### Example 1

The probability sample selected to represent the universe consists of approximately 71,000 households.

### Example 2

The data collection is a pooled cross-sectional time-series of bank robberies in 50 states over a period of 6 years (1970-1975), resulting in 300 observations.

### Example 3

Multistage, stratified, clustered sampling design.

## Data Source

The source of the data. This can include administrative records, agency-sponsored surveys, machine-readable files, or books and other publications.

### Examples:

United States Bureau of the Census Economic Surveys, 1998-2000

United States Congressional Record, 1989

Annual Company Organization Survey, 2003

"Voting Scores." Congressional Quarterly Almanac 33 (1977), 487-498

## Collection Mode(s)

The method(s) used to collect the data.

## Scales

List any commonly known scales used to collect data for the study (e.g., psychological scales such as MMPI and CPI, or occupational scales such as the Census Occupational Codes). The inclusion of a common scale should be identified in the documentation associated with this dataset and confirmed by variable names or labels. If the questionnaire used has a finite list of responses, such as "Always," "Sometimes," "Rarely," and "Never," or "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree," then enter in this field, "A Likert-type scale was used," or "Several Likert-type scales were used." The scales can be cited either as a list or described in full sentences.

## Weight(s)

A description of the criteria for using weights in the analysis of a data collection. If a weighting formula or coefficient was developed, provide this formula, define its elements, and indicate how the formula is applied to the data.

### Example 1

A weight variable with two implied decimal places has been included and must be used in any analysis.

### Example 2

The 1996 NES dataset includes two final person-level analysis weights which incorporate sampling, nonresponse, and post-stratification factors. One weight (variable #4) is for longitudinal micro-level analysis using the 1996 NES Panel. The other weight (variable #3) is for analysis of the 1996 NES combined sample (Panel component cases plus Cross-section supplement cases). In addition, a Time Series Weight (variable #5) which corrects for Panel attrition was constructed. This weight should be used in analyses which compare the 1996 NES to earlier unweighted National Election Study data collections.

## Unit(s) of Observation

The major unit being analyzed in the data collection. Use a brief phrase, for example: Individuals, Arrest event, Incident, Agency, Year and quarter, County within state.

## Geographic Unit

The smallest geographic unit represented in the data that can be analyzed.

# Related Publications

The following publications relate in an unspecified way to the data in this project.

# How to write a research question:

The summary of your project should include your research question. The research question acts as a guide for analyzing your data. The research question helps to focus each part of a replicable workflow – what data will be needed, how the data will be cleaned, what elements of the data need to be explored, what types of models are appropriate for data analysis and what outlets will be interested in the results. Formulating your research question should be an iterative process, first knowing that it is important provides motivation to start writing, but as a research project develops it is ok to revise and clarify the guiding question. As a note of caution, avoid changing your research question to fit the model results. To help with writing a good research question Foss and Williams (2015) provide six criteria[[2]](#footnote-2):

Assessing Research Questions: A research question formulated according to the following criteria ensures that your study has a solid center that can hold all of the pieces of the study together:

1. The research question clearly identifies the theoretical construct you are studying.
2. The research question contains some suggestion of recognizability of the theoretical construct.
   1. Will you be able to locate it and distinguish it easily from other constructs that appear in your data?
3. The research question (usually) transcends your data.
4. The research question identifies your study’s contribution to an understanding of the theoretical construct.
5. The research question has the capacity to surprise.
   1. Do you already know the answer?
6. The research question can produce robust results.

# Argument vs Research Question:[[3]](#footnote-3)

While a research question can provide clarity and guidance for your research, an argument is what is needed to support a journal article. According to Belcher (2019) an argument is “an answered research question or confirmed hypothesis” (p. 67).

|  |  |
| --- | --- |
| **Research Question** | **Possible Arguments** |
| Does x affect y when z is present? | x increases y when z is present. |
|  | x decreases y when z is present. |
|  | x does not affect y when z is present. |
|  | x affects y when z is present if… |

What arguments could you make based on your research question?

For an excellent discussion of how to write a strong argument see Belcher (2019) Week 2 – Advancing Your Argument. p. 60-87

# Data Workflow Planning:

The following sections are meant to help think through the data science process or workflow. As illustrated in figure one the process starts with raw data and moves to data cleaning. The data cleaning process creates a clean dataset or datafile that can be used in data exploration and data modeling. Notice that data exploration may lead to more data cleaning. After data is modeled the results can be communicated in a report and new data products can be sent out to the world.

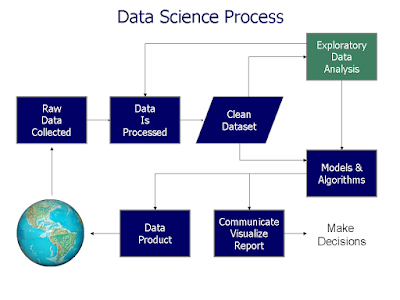


Figure . Data Science Process.[[4]](#footnote-4)

# Data Sources:

Remember “the *analysis* of the data, not the amount of data collected, determines the originality and significance of your study – the analysis is what answers your research question, not your data” (Foss and Williams, 2015, p. 50). When thinking of data sources, consider your research question and try to limit the amount of data that you collect.

For each data source fill out the following list. If an item in the list does not apply type “N/A”.

|  |  |
| --- | --- |
| File Name |  |
| File Type |  |
| Number of files |  |
| File storage size |  |
| Data file dimensions (rows, cols) |  |
| Principal Investigator |  |
| Summary |  |
| Funding Source |  |
| Scope of Data File |  |
| Subject Terms |  |
| Geographic Coverage |  |
| Time Period(s) |  |
| Collection Dates |  |
| Universe |  |
| Methodology |  |
| Response Rate |  |
| Sampling |  |
| Data Source |  |
| Collection Mode(s) |  |
| Scales |  |
| Weight(s) |  |
| Unit(s) of Observation |  |
| Primary Key |  |
| Foreign Keys |  |
| Geographic Unit |  |
| Codebook/Metadata format |  |
| Data Accessibility[[5]](#footnote-5) |  |
| Owner |  |
| Cost |  |
| Time |  |
| Location |  |

# Data Cleaning Process:

Data cleaning includes all of the steps needed to convert the source data into datafiles that are useful for data exploration and modeling. The clean datasets should also be considered data products that could be made public. Data cleaning also includes documenting metadata.

|  |  |
| --- | --- |
| **Typical Data Cleaning Steps** |  |
| Convert file formats |  |
| Extract patterns in text |  |
| Merge Files |  |
| Append Files |  |
| Drop observations |  |
| Drop variables |  |
| Generate new variables |  |
| Creating variables for analysis |  |
| Ordering variables |  |
| Compressing files |  |
| Recoding variables |  |
| Dummy (Binary) Variable creation |  |
| Label variables |  |
| Add variable notes |  |
| Add data file notes |  |
| Adding value labels |  |
| Using multiple languages |  |
| Sorting |  |
| Geocoding |  |
| Ranking |  |
| Flagging |  |
| Filtering Missing Data |  |
| Imputing (filling in missing data) |  |
| Adding unique non-missing id |  |
| Spatial Joins |  |
| Adding foreign keys for relational databases |  |
| Pivoting data (transposing wide to long or long to wide) |  |
| Removing duplicates |  |
| Detecting outliers |  |
| Random sampling |  |
| Working with date-time variables |  |
| Codebook generation |  |

# Cleaned Dataset

Describe the expected cleaned dataset.

|  |  |
| --- | --- |
| File Name |  |
| File Type |  |
| Number of files |  |
| File storage size |  |
| Data file dimensions (rows, cols) |  |
| Principal Investigator |  |
| Summary |  |
| Funding Source |  |
| Scope of Data File |  |
| Subject Terms |  |
| Geographic Coverage |  |
| Time Period(s) |  |
| Collection Dates |  |
| Universe |  |
| Methodology |  |
| Response Rate |  |
| Sampling |  |
| Data Source |  |
| Collection Mode(s) |  |
| Scales |  |
| Weight(s) |  |
| Unit(s) of Observation |  |
| Primary Key |  |
| Foreign Keys |  |
| Geographic Unit |  |
| Codebook/Metadata format |  |
| Data Accessibility[[6]](#footnote-6) |  |
| Owner |  |
| Cost |  |
| Time |  |
| Location |  |

# Data Exploration Plan:

Data exploration is a critical step to understanding your data and for checking data quality, validity, and reliability.

|  |  |
| --- | --- |
| **Typical Data Exploration Steps** |  |
| Summary statistics (count, mean, median, sd, min, max, percentiles) |  |
| Cross-tabs |  |
| Correlation matrix |  |
| Scatterplots |  |
| Boxplots |  |
| Histograms |  |
| T-test |  |
| ANOVA |  |
| Spatial exploration (GIS) mapping |  |
| Temporal exploration – time-series graphs |  |
|  |  |

## Example Table Shells[[7]](#footnote-7)

Include detailed examples of table shells. Please indicate the unit of analysis used in each table.

Table 1. Summary statistics for variables by [unit of analysis].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | count | min | max | p50 | mean | sd |
| Variable Label 1 |  |  |  |  |  |  |
| Variable Label 2 |  |  |  |  |  |  |

Source/Provenance

## Description of Data Exploration Graphs

Include detailed description of planned graphs. Please indicate the unit of analysis used in each graph.

## Description of Data Exploration Maps

Include detailed description of planned maps. Please indicate the unit of analysis used in each map.

# Data Modeling Plan:

Data modeling

|  |  |
| --- | --- |
| **Typical Data Exploration Steps** |  |
| Summary statistics (count, mean, median, sd, min, max, percentiles) |  |
| Cross-tabs |  |
| Correlation matrix |  |
| Scatterplots |  |
| Boxplots |  |
| Histograms |  |
| T-test |  |
| ANOVA |  |
|  |  |

# Example Table Shells[[8]](#footnote-8)

Include detailed examples of table shells. Please indicate the unit of analysis used in each table.

Table 2. Parameter Estimates from Models.

|  |  |
| --- | --- |
|  | Model 1 |
| Variable Label 1 |  |
|  |  |
| Variable Label 2 |  |
|  |  |
| Constant |  |
| Observations |  |
| Adjusted *R*2 |  |

Standard errors in parentheses

Source/Provenance

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

# Data Publication Plan:

1. Much of the text provided in this example outline comes from the data archiving instructions used by ICPSR. For more information visit <https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html> [↑](#footnote-ref-1)
2. Foss, S. K. and William, W. (2015). *Destination dissertation: A traveler's guide to a done dissertation*. Rowman & Littlefield. p. 36-48 – [Available on Google Books](https://books.google.com/books?id=kBmaCgAAQBAJ&lpg=PR5&ots=HvBqyk0XPX&dq=foss%20waters%20destination%20dissertaion&lr&pg=PA38" \l "v=onepage&q&f=false) and [TAMU Library](http://proxy.library.tamu.edu.srv-proxy2.library.tamu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1070638&site=ehost-live&ebv=EB&ppid=pp_35) [↑](#footnote-ref-2)
3. Belcher, W. L. (2019). *Writing your journal article in twelve weeks: A guide to academic publishing success*. University of Chicago Press. [Available on Google Books](https://books.google.com/books?id=xueZDwAAQBAJ&newbks=1&newbks_redir=0&lpg=PP1&dq=Belcher%20Writing%20your%20journal%20article%20in%20twelve%20weeks&pg=PP1#v=onepage&q&f=false) [↑](#footnote-ref-3)
4. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers Retrieved from <https://www.r-bloggers.com/a-data-science-solution-to-the-question-what-is-data-science/> [↑](#footnote-ref-4)
5. Foss and Waters (2015) p. 49 – Try to conceptualize your project so that you can collect the data easily and in a reasonable amount of time. [↑](#footnote-ref-5)
6. Foss and Waters (2015) p. 49 – Try to conceptualize your project so that you can collect the data easily and in a reasonable amount of time. [↑](#footnote-ref-6)
7. The idea behind table shells is to create a plan for analysis and to think through the data exploration process. The CDC National Center for Health Statistics requires this as part of their proposal process. For examples see <https://www.cdc.gov/rdc/b3prosal/PP300.htm> [↑](#footnote-ref-7)
8. The idea behind table shells is to create a plan for anlaysis and to think through the data exploration process. The CDC National Center for Health Statistics requires this as part of their proposal process. For examples see <https://www.cdc.gov/rdc/b3prosal/PP300.htm> [↑](#footnote-ref-8)