**DRAFT**

## TIER Protocol (version 2.1):  SPECIFICATIONS

The TIER Protocol specifies the content and organization of a set of electronic files that serve as replication documentation for a research paper that reports results based on the analysis of statistical data.

The documentation specified by the TIER Protocol meets (in fact exceeds) professional standards, such as journal policies on sharing data and code, but it has been designed primarily for use in classes and independent projects in which students are learning to conduct empirical research.  Certain features of the Protocol serve pedagogical purposes, but would probably not be appropriate prescriptions for professional practice.

The figure below illustrates the organization and contents of replication documentation for a completed paper that meets the specifications of the TIER Protocol.  The remainder of this document describes each of the folders and documents shown in the illustration.
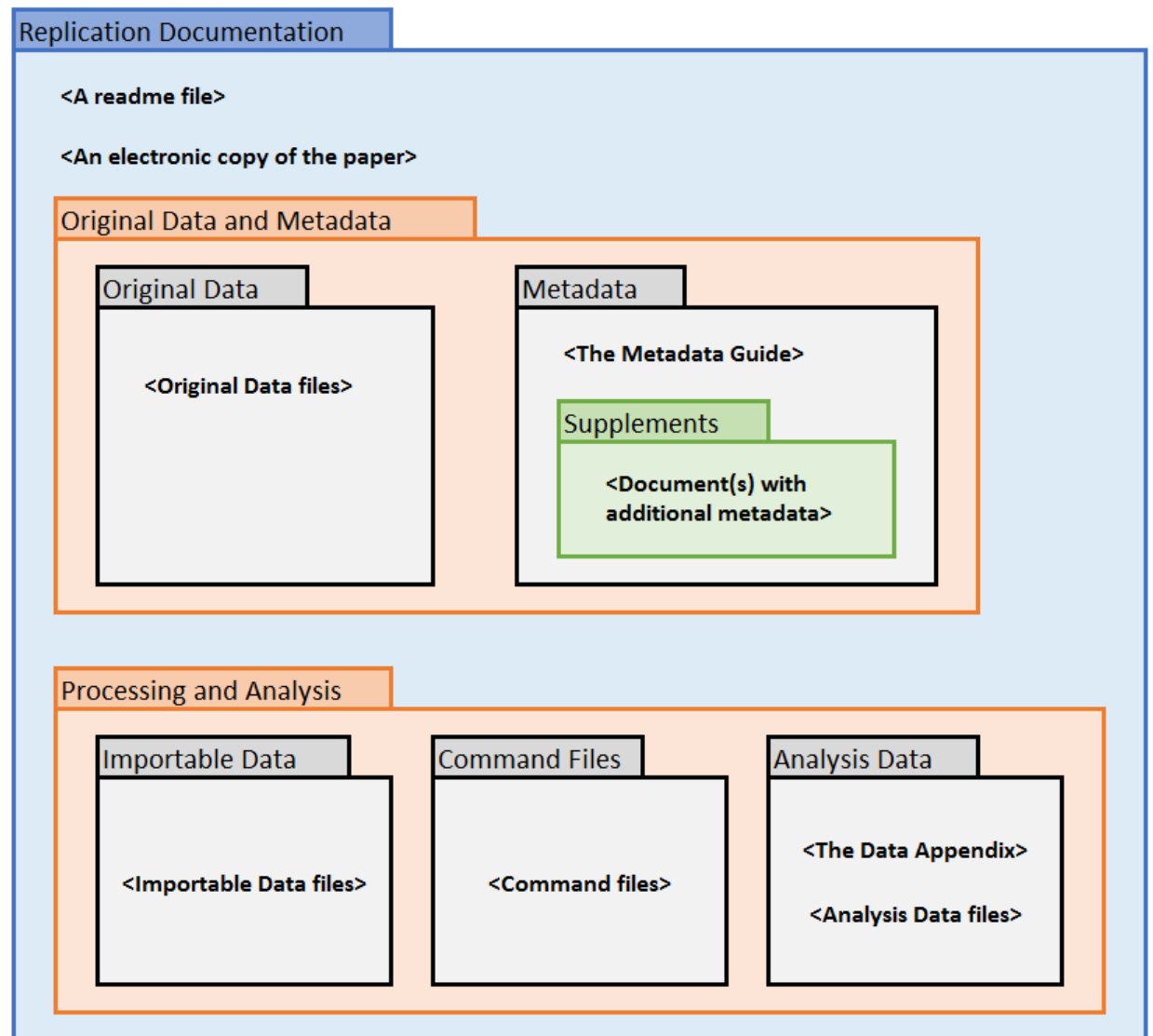
# The Replication Documentation Folder

The Replication Documentation folder contains two documents and two sub-folders.

### Two documents:

#### *The paper.*

An electronic copy of the research paper for which the replication documentation was prepared.  Usually in *PDF* format.

#### *A Read Me file.*

The Read Me file describes the files contained in the Replication Documentation folder and explains how they can be used to replicate the study.

### Two sub-folders:

#### *The Original Data and Metadata folder.*

The Original Data and Metadata folder serves as a record of the data files with which the author began the paper.  It contains copies of all files from which any data used for the project were initially extracted, as well as metadata for those original data files.

#### *The Processing and Analysis folder.*

The Processing and Analysis folder serves as a record of everything the author did with the original data files to generate the quantitative results reported in the paper.  It contains all the data and command files an interested reader would need to replicate all the data processing and analysis conducted for the study, and thereby reproduce the results reported in the paper.

# The Read Me File

The Read Me file is a document that describes the files included in the replication documentation, and explains how they can be used to replicate the study and reproduce the results.

The Read Me file is composed by the author of the paper, and consists of three main sections:

### An overview of the contents of the replication documentation.

This section of the Read Me file briefly describes all the files included in the replication documentation, and outlines the structure of the folders in which they are stored.

### A description of any modifications made to the original data files to create the corresponding importable data files.

This section of the Read Me file should identify all the original data files that had to be modified to make it possible for the software used for the study to read them, and describe precisely the modifications that were made to create the importable versions.

### Step-by-step instructions for using the replication documentation to replicate the study.

The section of the Read Me file gives clear and concrete instructions an interested reader can follow to use the replication documentation to replicate the data processing and analysis conducted for the study and thereby reproduce the reported results.

Typically, these instructions:

- State what kinds of statistical software (including version number and required add-ons) are required to run the command files.

- Explain which files included in the replication documentation need to be copied onto the replicator's computer, the structure of folders and sub-folders in which the files should be copied, and which of the folders each file should be saved in.

- Indicate which of the folders should be set as the working directory when the statistical software that executes the command files is run.

- Indicate the order in which the command files need to be run to carry out the replication. And, for each command file, indicate what other files it uses (e.g., what data files it opens and what other command files it calls) and what output it produces (e.g., new data files it saves, old data files it deletes, and new directories it creates).

4

# The Original Data and Metadata Folder

The Original Data and Metadata folder provides a record of the format and content of the data files used in the study when they were first obtained by the author.

This folder contains two sub-folders:

### The Original Data folder

The Original Data folder contains copies of all the original data files from which any of the data for the study were extracted.

### The Metadata Folder

The Metadata folder contains information about the content and format of the original data files and about the sources from which they were obtained.

# The Original Data Folder

An *original data file* is a file from which any of the data used for a study were extracted.

**A copy of every original data file used for the paper should be saved in the Original Data folder.**

Every piece of data necessary for the processing and/or analysis by which the results of the paper were generated should be traceable to the original data files in one of two possible senses:

- The piece of data can be found in one of the original data files.

*or*

- One or more pieces of data that were used to construct the piece of data in question can be found in the original data files, and code that combines and/or transforms the data from the original data files as necessary to create the piece of data in question can be found in the command files included in the replication documentation.

**The original data files saved in the Original Data folder should not be modified in any way.**

The content and format of every original data file should be identical to the content and format of the file when it was initially obtained by the author.

# The Metadata Folder

The Metadata Folder provides documentation for the original data files.  It contains:

**One document:  The Metadata Guide**

For every original data file, the Metadata Guide contains the kinds of information that are found in a codebook.

**One sub-folder:  The Supplements folder**

The Supplements folder contains additional documents with information about the original data files.

# The Metadata Guide

The Metadata Guide is a document that serves as a codebook for the original data files.

The Metadata Guide is composed by the author of the paper, and consists of one or more sections.  Each section provides information about one of the original data files.

> **For every original data file, the information in the corresponding section of the Metadata Guide should include:**
>
> > *A bibliographic citation for the original data file.*
> >
> > This citation should be in a format consistent with the editorial style (e.g., APA or Chicago) used in the main paper or report on the study.
> >
> > If the bibliographic citation does not include the date on which the author first downloaded, or obtained in some other way, the original data file, that date should also be provided.
> >
> > If the original data file has been assigned a unique identifier such as a Digital Object Identifier or a Universal Numeric Fingerprint (UNF), and that identifier is not included in the bibliographic citation, that identifier should also be provided.
> >
> > *A verbal explanation of how an interested reader can obtain a copy of the original data file.*
> >
> > In many cases, this explanation will give the URL of a webpage from which the data can be accessed, along with instructions for downloading from that webpage a file identical to the original data file used in the study.
> >
> > In all cases, this explanation should be complete and precise enough to allow an independent researcher to locate and obtain an exact copy of the original data file without any additional information or assistance.
> >
> > *Whatever additional information an independent researcher would need to understand and use the data in the original data file.*

The particular information required can vary a great deal depending on the nature of the original data file in question. This is therefore a place where the comment in the introduction about exercising empathy and judgment really comes into play.

In many cases, the additional information that should be provided about an original data file is often similar to the kind of information found in a codebook or users' guide for a data set, such as variable names and definitions, coding schemes and units of measurement, details of the sampling method and weight variables, and descriptions of how any imputed variables were constructed.

In some cases, it is also necessary to include information about the file structure (e.g., the delimiters used to separate variables, or, in rectangular files without delimiters, the columns in which the variables are stored).

Any other unique or idiosyncratic aspects of the data that an independent user of the data would need to understand should be explained as well.

***Information about an original data file that is available in a publicly available document or at a stable website may be omitted from the Metadata Guide, provided the Metadata Guide includes a note explaining where the information can be found.***

If the information is available in a document such as an existing codebook or users' guide for an original data file, a copy of that document should be stored in the Supplements folder.

In this case, the Metadata Guide should include a note indicating the information is available in the document stored in the Supplements folder, and stating where in the document the information can be found.

The Metadata Guide should also provide an appropriate citation for every document in the Supplements folder.

If the information is available only in an interactive online interface, and not as a static document that can be preserved in the documentation, the Metadata Guide should include a note indicating how an interested reader can access that information.

# The Supplements Folder

The Supplements folder contains existing documents that provide information about the original data files that the author chose to omit from the Metadata Guide.

As described in the specifications for the Metadata Guide, when some of the required information for an original data file is available in an existing public document, the author may choose to omit the information from the Metadata Guide, and instead include a copy of the document that contains the information in the replication documentation for the project.

If any such documents are included in the replication documentation, they should be stored in the Supplements folder.

# The Processing and Analysis Folder

This folder documents all the steps of data processing and analysis by which the author generated the results reported in the paper. It contains all the data and code an independent researcher would need to replicate the study and reproduce all the results.

The Processing and Analysis folder contains three sub-folders:

### The Importable Data Folder

The Importable Data folder contains copies of all the original data files, modified as needed so that they can be read by the statistical software used for the study.

### The Command Files Folder

This Command Files folder contains command files that execute all the data processing and analysis necessary to replicate the study.

### The Analysis Data Folder

The Analysis Data folder contains all the fully cleaned and processed data files that are used to generate the results reported in the paper.

The Analysis Data folder also contains the Data Appendix, which serves as a codebook for the cleaned and processed data files.

# The Importable Data Folder

An importable data file is a version of an <span style="color:green">original data file</span> that has been modified as necessary to make it possible for the statistical software used for the study to read it.

**For every original data file in the Original Data folder, there should be a corresponding importable data file in the Importable Data folder.**

*If an original data file is in a format that can be opened or read by the statistical software that was used for the study:* the corresponding importable data file should be an identical copy of the original.

*If an original data file is not in a format that can be opened or read by the statistical software that was used for the study:* the corresponding importable data file should be copy of the original that has been modified in a way that makes it possible for the software used for the study to open or read it.

**Importable Data files should be modified to the minimal extent possible.**

Importable data files should be created by modifying the original data files only to the minimal extent necessary to make it possible for the statistical software used for the study to read them.

**Modifications made to original data files should be recorded in the Read Me file.**

All the modifications made to the original data files to create the importable data files should be described in the <span style="color:green">Read Me file</span>.

# The Command Files Folder

The Command Files folder contains one or more files containing code written in the syntax of the statistical software used for the study.  The code in these command files executes all the data processing and analysis necessary to replicate the study and reproduce the reported results.

The way in which the code that executes the various steps of data processing and analysis may vary depending on the nature of the project and the preferences of the author.  In many cases, however, the steps can be grouped into three phases, with one or more command files executing the steps in each phase.

### Processing the data

The command files for this phase transform the importable data files into the analysis data files, which contain the fully cleaned and processed data that are used to generate the results reported in the paper.

### Constructing the Data Appendix

The command files for this phase generate the descriptive statistics, tables, and figures presented in the Data Appendix, a document that serves as a codebook for the analysis data files.

### Generating the results

Using the data in the analysis data files, the command files for this stage conduct the procedures that generate the results reported in the paper.

13

# The Analysis Data Folder

The Analysis Data folder contains:

### One or more analysis data files.

The analysis data files are the cleaned and processed data files that are used to generate the results reported in the paper.

They are constructed by the command files containing the code that executes the processing phase of the project.

A copy of every analysis data file should be saved in the Analysis Data folder.

### The Data Appendix.

The Data Appendix is a document that serves as a codebook for the analysis data files.

# The Data Appendix

The Data Appendix is a document that serves as a codebook for the analysis data files.

The Data Appendix is composed by the author of the paper.

One section of the Data Appendix is devoted to every analysis data file.

The section devoted to a particular analysis data file contains a sub-section about each of the variables contained in the file.

The information presented in each sub-section depends on the type of variable the sub-section is about.

### For every variable.

The information should include:

- The name of the variable and a complete definition (including as appropriate, for example, coding and/or units of measurement, the wording of a survey question the variable is based on, or adjustments made for inflation or PPP).

- The names of the original data file from which the variable was extracted, or from which the variables used to construct it were extracted, and the names of the variables extracted from the original data files.

- The number of observations with valid values for the variable, and the number of observations with missing values.

### For quantitative variables.

The information should include:

- Basic summary statistics: the mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum.

- A histogram.

15

## For categorical variables.

The information should include:

- A frequency table.

- A bar chart illustrating the frequency distribution.