RJB
First version: 2015-12-21
This version: 2016-03-30

The TIER Documentation Protocol *v2.0*

Version 2.0 for R [*.pdf* format]

I. Overview

The TIER Documentation Protocol provides instructions for assembling a set of electronic files that document all the steps of data processing and analysis you conduct for an empirical research paper.

The documentation specified by the Protocol contains all the data, computer programs, and explanatory information an independent researcher would need to be able to replicate the data processing and analysis you conducted for the project and to reproduce exactly all the results reported in your paper.

The instructions presented here are written for users of R. In a few places, they use R-specific terminology. For example, we refer to command files as scripts, and their names are followed by the *.R* extension. But the R-specific terminology that appears in these instructions can be easily translated to any of the major statistical packages (such as SPSS, SAS, Stata or Matlab).

Two other versions of the Protocol are available for download in *.pdf* format:

> A version written for Stata users.
>
> A software-neutral version that avoids the use of terminology specific to any particular software package.

An on-line version of the Stata-specific Protocol is also available.

II. Create the folders in which you will store your work

To begin, you need to choose a place to keep your files as you create and modify them throughout the course of your research, and to preserve your complete replication documentation when you have finished your paper.

Here we present two options:

> (i) You can store your files on the hard disk of a personal computer—either your own, or some other computer you will be using for the project.
>
> (ii) You can store your files on an online platform for managing and sharing research documents known as the Open Science Framework (OSF).

If you are interested in using the OSF platform, please see our demo, which illustrates the use of OSF for organizing replication documentation that meets the specifications of the TIER Protocol.

If you choose to store your files on a personal computer, you must begin by constructing the hierarchy of folders and sub-folders in which the various components of your documentation will be stored.

> Instructions for constructing this hierarchy of folders and sub-folders yourself can be found below in Section II.A.
>
> If you prefer, you may simply download a ready-made set of the complete hierarchy of folders.
>
> Whether you download a ready-made hierarchy of folders or construct it yourself, you should save it in a secure location on the hard disk of the computer you will be working on.

II.A. Constructing the complete hierarchy of folders

To construct the hierarchy of folders and sub-folders in which the various components of your documentation will be stored, first create a new (empty) folder and give it the name "Replication Documentation." Save this folder in a secure location on the hard disk of the computer you will be using for your project.

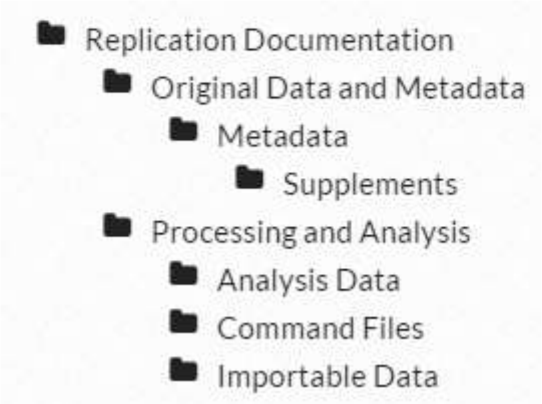
Then create two more new folders, give them the names "Original Data and Metadata" and "Processing and Analysis." Save both of these folders in the "Replication Documentation" folder.

Create two additional new folders, and give them the names "Original Data" and "Metadata." Save both of these in the "Original Data and Metadata" folder.

Create one more new folder, give it the name "Supplements," and save it in the "Metadata" folder.

Create three new folders, and give them the names "Importable Data," "Command Files," and "Analysis Data." Save all three of these folders in the "Processing and Analysis" folder.

When you are finished constructing this hierarchy of folders, it should look like this:



- Replication Documentation
  - Original Data and Metadata
    - Metadata
      - Supplements
  - Processing and Analysis
    - Analysis Data
    - Command Files
    - Importable Data

III. The “Replication Documentation” folder

This is the folder in which you will store all the documentation for your paper, as well as the final paper itself.

The top level of the “Replication Documentation” folder should contain two documents:

- An electronic copy of your complete final paper
- The ReadMe file for your replication documentation (see Section III.C below)

The “Replication Documentation” folder should also contain two sub-folders:

- The “Original Data and Metadata” folder (Section III.A)
- The “Processing and Analysis” folder (Section III.B)

III.A. The “Original Data and Metadata” folder

The “Original Data and Metadata” folder should contain two sub-folders:

- The "Original Data" folder (Section III.A.1)
- The "Metadata" folder (Section III.A.2)

III.A.1. The “Original Data” folder

We use the term “original data file” to refer to a file from which you took any of the statistical data you used for your project.

Your “Original Data” folder should contain a copies of all your original data files.

Any data that were necessary for any part of the processing and/or analysis you reported in your paper should be contained in one of the original data files in your “Original Data” folder.

Every original data file should be saved in exactly the format it was in when you first obtained it. You may choose to change the name of an original data file when you save it in your “Original Data” folder, but other than that the file should not be modified in any way.

III.A.2. The “Metadata” folder

The top level of your “Metadata” folder should contain one document:

- the Metadata Guide (Section III.A.2.a)

The “Metadata” folder should also contain one sub-folder:

- the “Supplements” folder (Section III.A.2.b)

III.A.2.a. The Metadata Guide

The Metadata Guide is a document that provides information about each of your original data files. For every file in your "Original Data" folder (Section III.A.1), the information in the Metadata Guide should include:

i. A bibliographic citation of the original data file in whatever editorial style (e.g., APA or Chicago) you have chosen to follow throughout your paper.

ii. The date you downloaded, or obtained in some other way, the original data file.

iii. Any unique identifiers, such as a Digital Object Identifier (DOI) or Universal Numeric Fingerprint (UNF), that have been assigned to the original data file.

iv. 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 website from which the data can be accessed, along with instructions for downloading a file identical to the original data file you obtained from that site. In all cases, this explanation should be complete and precise enough to allow an independent researcher to locate and obtain the data file without any additional information or assistance.

v. 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, and deciding what additional information to provide therefore requires thoughtful consideration and judgment. In many cases, the relevant information is similar to what is found in a codebook or users' guide for a dataset: variable names and definitions, coding schemes and units of measurement, and details of the sampling method and weight variables. 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.

If you used data from more than one original data file, the Metadata Guide should be comprised of multiple sections, with one section providing this information for each of your original data files.

The information specified in items (i)-(iv) should be presented in text that you write yourself.

The additional information described in (v) may also be presented in text that you write yourself. In some cases, however, some or all of the relevant additional information can be found in existing supplementary documents, such as users' guides and codebooks that accompany the original data file. When relevant additional information is available in supplementary documents, it is not necessary to include this information in the text of the Metadata Guide. Instead, you may simply include copies of the documents that contain the relevant information in the documentation of your paper. These supplementary documents should be stored in the "Supplements" folder.

For every document you include in the "Supplements" folder, there should be comments in the Metadata Guide that:

i. identify the document.

ii. cite the source from which it was obtained.

iii. indicate which of your original data files it pertains to.

iv. state what relevant information it contains.

You may use any word-processing software you choose to compose the Metadata Guide. It should be named *metadata_guide.EXT*, where *.EXT* represents the extension attached to the names of files created with your word-processing software. (For example, if you are using Microsoft Word, the document would be called *metadata_guide.doc* or *metadata_guide.docx*.)

The Metadata Guide should be stored in the "Metadata" folder (Section III.A.2).

III.A.2.b. The “Supplements” folder

As described in the instructions for the Metadata Guide (Section III.A.2.a), the “Supplements” folder is where you store any existing documents related to your original data files, such as users’ guides or codebooks, that contain relevant information you omitted from the Metadata Guide.

III.B. The “Processing and Analysis” folder

The “Processing and Analysis” folder should contain three sub-folders:

- The "Importable Data" folder (Section III.B.1)
- The “Command Files” folder (Section III.B.2)
- The “Analysis Data” folder (Section III.B.3)

III.B.1. The “Importable Data” folder

For each of the original data files in your “Original Data” folder (Section III.A.1), you should create a corresponding version that we will call an “importable data file.” These importable data files should be stored in the “Importable Data” folder.

In some cases, the importable data file will be a slightly modified version of the original. In other cases the importable version will be identical to the original.

Whether or not an importable data file differs from the original version will depend on whether the original version is in a format that R can open or import.

There are two cases to consider:

i. **The original data file is in a format that R can open or import.**

   This case obviously applies if the original data file is in R’s *.Rdata* format.

   This case also applies to files that are not in *.Rdata* format, but that can be opened with R. For example, R’s `read.table()` command can be used to import data from a file in *.csv* format. Similarly, if you have loaded the XLConnect package, the `readWorksheetFromFile()` command can be used to import data from an Excel workbook.

   When an original data file is in R’s *.Rdata* format, or another format that can be imported into R without any modification, the corresponding importable data file should be an exact copy of the original. In these cases, the copy of the file in the “Importable Data” folder should have the same name as the copy in the “Original Data” folder.

   Note, however, that in some cases, even when an original data file is in *.csv* or Excel format, it may be convenient or necessary to modify it slightly before using R to import the data it contains. Three examples of cases in which this is true are given below under item (ii).

ii. **The original data file must be modified before it can be imported to R.**

   In some situations, it may be necessary or convenient to modify an original data file before importing it to R. The following examples illustrate a few of the common cases:

   - If the original data file is a spreadsheet that contains explanatory notes as well as data, it may be necessary or convenient to remove those notes from the importable version of the spreadsheet.

   - If a certain variable is measured in dollars, and a dollar sign ($) precedes each value of the variable in the original *.csv* or Excel data file, you may wish to remove the dollar signs so that R recognizes that the variable should be stored in a numeric format.

- If an original data file is formatted for use with a particular type of software other than R (e.g., SPSS, SAS, R or Matlab), it may be necessary to convert the file from its original format to R's .*Rdta* format using a package like Stat/Transfer (https://www.stattransfer.com/).

As these examples illustrate, the particular ways in which an original data file needs to be modified will vary depending on the nature of the original data file. But in every case, the modifications made to an original data file to create the importable version should follow this general principle:

> ***The importable data file should be as nearly identical as possible to the original; no changes should be made to the file other than the minimal modifications required to allow R to read the data it contains.***

When an importable data file is a modified version of the corresponding original data file, the original and importable versions should be given different names.

III.B.2. The "Command Files" folder

This folder should contain one or more R scripts (with the *.R* extension) that contain commands that execute every step of data processing and analysis required to reproduce the results you report in your paper.

In all of the scripts you write, it is important to include comments that are detailed and clear enough to make it possible for someone not familiar with your project to understand the steps of data processing and analysis that are executed by the commands in the script.

For the purpose of constructing and organizing your scripts, you should think of the work on your project in terms of three phases, which we will call (i) importing, (2) processing, and (3) analysis. Your scripts will include one or more files that execute each of these phases of research.

i. For the importing phase, the script(s) should contain commands for R to read the data in each of your importable data files, and then save them in *.Rdata* format. For example, if you have an importable data file in *.csv* format, your script(s) will include a command like `read.table()` that imports the data, and a `save()` command that saves the data in an *.Rdata*-formatted file. (If an importable data file is already in *.Rdata* format, nothing needs to be done to it during the importing phase.)

   At the end of the importing phase, you will have a set of data files, all in R's *.Rdata* format, containing all the data for your project. These data files will serve as inputs to the processing phase described below.

   When you have completed your paper, these files should not be included in the final documentation. Anyone interested in these files can create them simply by executing the do-files you wrote for the importing phase of your project.

ii. The script(s) for the processing phase should include commands that execute all the processing required to transform your importable data files into the final data file(s) that you will use in your analysis. Exactly what these steps will be is highly variable, but they typically include operations such as joining two or more data files, dropping variables or cases, generating new variables, and recoding. At the end of the script(s) for the processing phase, there should be `save()` commands that save (in *.Rdata* format) the final data file(s) upon which your analysis will be conducted. We will refer to the final data files(s) that you use in your analysis as your "analysis data file(s)."

If you have a single analysis data file—i.e, if all of your analysis will be performed using a single data file created during the processing phase—it should be saved with the name *analysis.Rdata*. If you have more than one analysis data file, give them informative names, such as *analysis_euro.Rdata*, *analysis_afri.Rdata*, and *analysis_asia.Rdata*; or *analysis_individual.Rdata* and *analysis.household.Rdata*.

Your analysis data file(s) should be stored in your "Analysis Data" folder (Section III.B.3).

Strictly speaking, including your analysis data file(s) in the documentation is redundant: anyone interested in your analysis data file(s) files can create them simply by executing the R scripts you wrote for the importing and processing phases of your project. Nonetheless, the TIER Protocol calls for your analysis data file(s) to be included simply because it is sometimes convenient to have a readily accessible copy of the analysis data.

iii. The script(s) for the analysis phase should contain commands that open the analysis data file(s) you created in the processing phase, and then generate the results reported in your paper.

Every command in your analysis script(s) that generates a piece of output or a result reported in your paper should be preceded by a comment that indicates what piece of output or result the command will generate. The following examples illustrate some typical kinds of comments:

```
#The following command produces Table 6.

#The following command produces Figure 12.

#The following command calculates the correlation of  -0.54 between variables X and
#Y reported on page16 of the paper.
```

All of the scripts for importing, processing and analyzing your data should be included in the "Command Files" folder.

One additional script, called *data_appendix.R*, should also be included in your "Command Files" folder. This script is described in the instructions for your Data Appendix (Section III.C).

III.B.3. The “Analysis Data” folder

This folder should contain:

- Your analysis data file(s) [as described in the instructions for your “Command Files” folder (Section III.B.2)]

- Your Data Appendix (Section III.C)

III.C. The Data Appendix

Your Data Appendix is a document that serves as a codebook for your analysis data file(s).

If the data processing phase of your research generated just one analysis data file, and all the results presented in your paper were derived from that single analysis data file, the Data Appendix should begin with a brief description of the analysis data file. Typically, this description will say something about the scope of the sample or population the data represent, specify the unit of analysis, and indicate the number of observations. As in the case of the metadata that accompanies your original data files, however, exactly what information is relevant will depend on the nature of the analysis data file, so deciding which aspects you will describe in the Data Appendix will require judgment.

After the brief description of the analysis data file, the Data Appendix should present information about every variable in the analysis data file. The information presented about each variable should include:

- a complete definition of the variable (including coding and/or units of measurement).
- the name of the original data file(s) from which the variable was extracted or from which the variables used to construct it were extracted.
- the number of valid observations for the variable, and the number of cases with missing values.

For categorical variables, the information should also include:

- a frequency table.
- a bar chart showing the proportion of observations in each of the possible categories.

For quantitative variables, the information should also include:

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

If the results presented in your paper were derived from more than one analysis data file, the Data Appendix should include all of the above information—the brief description of the data file and the information about each of the variables contained in the file—for each of the analysis data files that was used.

You should compose your Data Appendix using whatever word-processing software you choose. It should be named *data_apppendix.EXT*, where *.EXT* represents the extension attached to the names of files created with your word-processing software. (For example, if you are using Microsoft Word, the document would be called *metadata_guide.doc* or *metadata_guide.docx*.)

Your Data Appendix should be stored in your “Analysis Data” folder (Section III.B.3).

In addition to the Data Appendix itself, you should also save an R script that generates all the output presented in the Data Appendix. This script should be named *data_appendix.R*, and it should be stored in your “Command Files” folder (Section III.B.2).

III.D. The ReadMe file

The ReadMe file gives information about all the other files included in the documentation for your paper. In particular, the ReadMe file should:

i. state what statistical software or other computer programs are needed to run the command files.

ii. explain the structure of the hierarchy of folders in which the documentation is stored, and briefly describe each of the files included in the documentation.

iii. describe precisely any changes you made to your original data files to create the corresponding versions saved in your Importable Data folder (Section III.B.1).

iv. give explicit, step-by-step instructions for using your documentation to replicate the statistical results reported in your paper.

You should compose your ReadMe file using whatever word-processing software you choose. It should be named *ReadMe.EXT*, where *.EXT* represents the extension attached to the names of files created with your word-processing software. (For example, if you are using Microsoft Word, the document would be called *metadata_guide.doc* or *metadata_guide.docx*)

Your ReadMe file should be stored in the top-level of your “Replication Documentation” folder (Section III).