RJB
2014-06-17

TIER Documentation Protocol 2.0-beta

This is a beta version of what we are calling the TIER Protocol 2.0.

We have not previously given version numbers to the revisions of the protocol that we have incrementally introduced on our website (www.haverford.edu/TIER), but we plan to do so from now on to avoid confusion.

We will think of the version posted on our website today (2014-06-17) as TIER Protocol 1.0, though it is not labeled as such.

The version of the protocol described in this document is very similar in spirit to version 1.0. The main change is a slight rearrangement of the structure of folders (directories) in which the various documents prescribed by the protocol are stored. We have found this slightly rearranged structure to be a little clearer and more convenient than the structure given in version 1.0. We have also simplified and streamlined the instructions for the protocol in a few ways.

We are calling this version a beta-version because there are several things we want to add before calling it the real TIER Protocol 2.0. The following are some of the things we want to add: explanations for why the Protocol calls for certain documents and why it specifies that they should be organized in certain ways, practical tips on constructing various components of the documentation, examples of tricky situations that arise when constructing this kind of documentation and suggestions on how to deal with them, and comments on the attitude with which one should approach the task of documenting an empirical research project. For people with a bit of understanding of the TIER Protocol, however, this bare-bones version should not be difficult to understand and implement.

This version of the protocol is written in a software-neutral way. There are occasional references to specific packages, but these are just as examples, and all the principles and procedures should be applicable for users of any of the major statistical packages (including Stata, SPSS, SAS and R). We are developing separate versions of the protocol that are written specially for particular kinds of software. Work on versions for Stata and R are in progress, and we plan to produce a version for SPSS as well.

Section I of this version of the protocol describes the components of the documentation that should be assembled, and Section II describes how these components should be formatted and organized.

I. The components of the documentation

The documentation of your paper should consist of a set of electronic documents that include the following components:

--a set of original data files

--metadata describing your original data files

--a set of importable data files

--a set of computer command files

--a data appendix

--a read-me file

Each of these components is described below.

A. Original data. Your original data should include every document from which you extracted any of the statistical data used for your project. Each original data file should be preserved in exactly the format in which you first obtained it, with no modifications.

B. Metadata for original data files. Metadata should be provided for each of your original data files. The metadata for an original data file should include (i) information on the source of the file, and (ii) any additional information an independent researcher would need to understand and use the data contained in the file.

(i) The information on the source of the file should include:

--a bibliographic citation for the data file, in the style (e.g., APA or Chicago) you have chosen to follow throughout your paper

--a Digitial Object Identifier (DOI), if one has been assigned to the file

--the date you downloaded or otherwise obtained the file

--a verbal description of how you obtained the data file (e.g., the URL from which you accessed the file and the steps you needed to take to download it to your computer)

(ii) The additional information an independent researcher would need to understand and use the data can vary a great deal depending on the nature of the particular data file in question. 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.

You should create one document that serves as a guide to all your metadata. When completed, this document should be saved in .pdf format, with the name *metadata-guide.pdf.* One section, or entry, of this document should be devoted to each of your original data files. The entry on a particular original data file should begin with a presentation of the source information described above in item (i). If the additional information described in item (ii) can be conveniently presented in the *metadata-guide.pdf* document, then this information should also be included. If the additional information is voluminous, or if you choose to include any additional documents that contain some of the information that was available with the original data (such as a codebook or users' guide), then this supplementary information can be stored in additional documents. In this case, the entry in the *metadata-guide.pdf* document for the original data file should mention the relevant supplementary documents(s) and describe the information that can be found in them.

C. Importable data. For each of your original data files, you should create a corresponding version that we will call an importable data file.

The importable data file will usually be a slightly modified version of the original, but in some cases the importable version will be identical to the original. Whether and how the importable version differs from the original will depend on whether the original version is preserved in a format that the statistical software you use for your project can open or import.

For every original data file that is preserved in a format that your software can open or import, the importable version should be an exact copy of the original.

For every original data file that is preserved in a format that your software can not open or import, the importable version of the file should be modified to the minimal extent necessary to allow your software to open or import it. The necessary modifications will vary depending on the format of the original file. For example, if the original data file is a spreadsheet that contains explanatory notes at the bottom of the sheet, it will probably be necessary to remove those notes

from the importable version of the spreadsheet. As a second example, if an original data file is formatted for use with a particular type of software (e.g., SPSS) and you are using different software (e.g., Stata), it may be necessary to convert the SPSS-formatted file either to some kind of delimited text format or to Stata's proprietary format. Many other cases may arise in practice, but the principle is always the same: 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 your software to read the data it contains.

Explanations of the modifications made to the original data files (if any) when the importable files were constructed should be given in your read-me file (which is described in more detail below).

D. Command files. Your documentation should include a set of files, written in the syntax of the software you are using for the project, containing commands that execute every step of data management and analysis required to replicate the results you report in your paper.

In all of the command files you create, it is very important to include comments that are detailed and clear enough to make it possible for someone not familiar with your project to understand every step of data management and analysis being executed.

For the purpose of constructing and organizing your command files, you should think of the work you do for your project in terms of three phases, which we will call (i) importing, (2) processing, and (3) analysis. Your command files will include one or more files that execute each of these phases of research. (One additional command file that should be included in your documentation is described below in the section on the Data Appendix.)

(i) For the importing phase, the command file(s) should instruct your software to read the data in each of your importable data files, and then save them in the format of the statistical software you are using. For example, if you are using Stata and you have an importable data file in .csv format, your command file(s) will include an `insheet` or `infix` command that reads the data and a `save` command that saves the data in a .dta-formatted file. (If an importable data file is already in your software's 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 the format of your statistical software, containing all the data you will use for your project.

(ii) The command file(s) for the processing phase should include instructions for all the steps of data management required to transform the data files you created in the importing phase 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 missing values. The command files for the processing phase end by saving the final data file(s) upon which the analysis will be conducted. We will refer to the final data files(s) that you use for your analysis as your "analysis data file(s)."

(iii) The command file(s) for the analysis phase contain instructions for opening the analysis data file(s) you created in the processing phase, and then for generating the results reported in the paper. Each command 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. (E.g., "The following command produces Table 6," "The following command produces Figure 12," or "The following command calculates the correlation of -0.54 between variables *X* and *Y* reported on page16 of the paper.")

E. Data Appendix. Your data appendix is a document that serves as a users' guide to your final data file(s).

If the results presented in your paper were derived from a 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 your analysis data file, so deciding which aspects you will describe in the data appendix will require thoughtful consideration and judgment.

After this brief description of the analysis data file, the data appendix should present information about every variable in the analysis data file. For categorical variables, this information should include (i) a frequency table and (ii) a bar chart showing the proportion of observations in each of the possible categories. For quantitative variables, the information should include (i) basic summary statistics (mean, standard deviation, minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum) and (ii) 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 for each of the analysis data files that was used.

In addition to the data appendix itself, you should also preserve a command file that generates all the output (frequency tables, bar charts, summary statistics and histograms) presented in the data appendix.

F. Read-me file. Your documentation should include a read-me file that gives information about all the other files included in the documentation. In particular, the read-me file should:

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

--explain the structure of the directories in which the documentation is stored (instructions on how the directories should be organized are given below in Section II), and briefly describe each of the files included in the documentation

--for every importable data file, this brief description should include an explanation of the modifications that were made to the original data file

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

II. Organizing and formatting the documentation

To begin, create a directory with a name like (but more informative than) "My Project." Create this directory somewhere secure and easily accessible (ideally on the hard disk of your computer).

Inside your "My Project" directory, create a template of (empty) subdirectories with structure illustrated below:

My Project

Original data and metadata [sub-directory of "My Project"]

Original data [sub-directory of "Original data and metadata"]

Metadata [sub-directory of "Original data and metadata"]

Metadata supplements [sub-directory of "Metadata"]

Processing and analysis [sub-directory of "My Project"]

Importable data [sub-directory of "Processing and analysis"]

Command files [sub-directory of "Processing and analysis"]

The components of your documentation should be stored in these directories as follows:

A. Original data. All of your original data files should be stored in the "Original data" directory.

B. Metadata for original data files. You should create one document that serves as a guide to all your metadata. When completed, this document should be saved in .pdf format, with the name *metadata-guide.pdf*. This document should be stored in the top level of the "Metadata" directory.

The *metadata-guide.pdf* should consist of one section, or "entry," for each of your original data files.

> The source information [item (i) described in Section I.B. above] should appear at the beginning of the entry for each of your original data files.
>
> If possible, the additional information [item (ii) described in Section I.B. above] should be presented immediately after the source information. In some cases, however, the most convenient way to incorporate some or all of this additional information in your documentation may be to include copies of documents (such as codebooks or users' guides) that the data distributor made available with the original data. If you include any such additional documents in your documentation, they should be stored in the "Metadata supplements" directory; in the entry in *metadata-guide.pdf* for the original data file in question, you should refer to each of the additional documents you have stored in the "Metadata supplements" folder, and indicate the relevant information that can be found in them.

C. Importable data. All of your importable data files should be stored in the "Importable data" directory.

D. Command files. All of your command files (those described above in Section I.D. as well as the one described in Section I.E.) should be stored in the "Command files" directory. It is helpful to give the command files names that indicate which part of the project each is used for: importing, processing, analysis, or the data appendix.

E. Data appendix. The data appendix should be saved in .pdf format, with the name data-appendix.pdf. It should be stored in the top level of the "My Project" directory.

F. Read-me file. The read-me file should be saved in .pdf format, with the name read-me.pdf. It should be stored in the top level of the "My Project" directory.