October 7, 2013

I. Overview

This protocol provides instructions for assembling a set of electronic files that document all the steps of data management and analysis you conduct for an empirical research project. This documentation includes four kinds of files: data, metadata, computer command files, and a readme file. Detailed instructions describing the content, format and organization that the protocol specifies for each of these kinds of documents can be reached from the links at the top of this page.

The instructions presented here are written for students using Stata (version 12), and we therefore use some Stata-specific terminology. (For example, we refer to command files as do-files.) But the references to Stata that appear in the instructions can be easily translated to any of the major statistical packages (such as SPSS, SAS or R). Similarly, although we have written the instructions for students working on Windows operating systems, they can be adapted for Mac operating systems with minor modifications.

To begin, create a new folder and give it the name "Documentation." Save this folder on your personal computer or some other site that you can easily access. This is the folder in which you will be storing documentation files as you create and modify them throughout the course of your research, and it is the folder in which the final versions of those files will be preserved.

II. Data

First create a new folder called "Data," and place it in the top level of your "Documentation" folder. Then create two more new folders with the names "Original" and "Importable," and place them in the top level of your "Data" folder.

Original data files.

In the "Original" data folder, you should save a copy of every file from which you obtained statistical data that you used for your project. These original data files should be preserved in exactly the format they were in when you first obtained them.

If all the data you used were contained in a single file when you first obtained them, then that file will be your only original data file; if your data were originally contained in two or more files, then each individual original data file should be included in the documentation. (If you use data from several different worksheets contained in a single Excel workbook, you should create an individual original data file for each worksheet.)

Your original data files should be given names of the form ***original_'name'.'ext'***, where ***'name'*** is an informal name you choose to refer to the source and/or content of the data, and ***'ext'*** is the extension determined by the format of the file. For example, an original data file with data on US presidential elections that was in Stata's **.dta** format when you first obtained it could be named ***original_elections.dta***; a tab-delimited text file obtained from the Penn World Tables could be named ***original_pwt.txt***; and an Excel-formatted file with data on development aid from the UN Common Database could be named ***original_UNaid.xls***.

Importable data files.

For every file in your "Original" data folder, you should create a corresponding importable version, and store it in your "Importable" data folder. The importable version should be as similar to the original data file as possible. Changes should be made to the original data file only if the original data file is neither in Stata's **.dta** format nor in a format that can be imported into Stata. When such changes are necessary, the file should be modified only in the minimal ways required to make it possible for Stata to open or import the data.

The modifications (if any) that you make to an original data file when you create the corresponding importable version will depend on the format of the of the original data file. Four cases are commonly encountered: (i) original data files that are in Stata's **.dta** format, (ii) original data files that are in a format (such as tab-delimited or comma-delimited text) that can be imported to Stata with a command like `insheet` or `infix`, (iii) original data files that are in the format of one of the major statistical packages other than Stata (such as SPSS, SAS or R), (iv) original data files that are in Excel's **.xls** or **.xlsx** format.

(i) If an original data file is in **.dta** format, the corresponding importable data file you create should simply be an exact copy of the original data file.

(ii) When an original data file is stored in tab- or comma-delimited text or some other format that can be imported to Stata with `insheet, infix` or a similar command, creating the corresponding importable file usually requires little or no modification of the original.

> If the original file contains nothing but rectangular data organized into (delimited) columns representing variables and rows representing cases (with or without variable names in the first row), the importable version should simply be an exact copy of the original data file.
>
> If the original data file contains additional information, like variable definitions, citations of the sources of the data, the URL of the website from which the file was downloaded, or any other explanatory notes or comments, then everything other than the rows and columns of data (perhaps with variable names in the first row of each column) should be

deleted from the importable version. The importable data file should then be saved in the same format (tab- or text-delimited) as the original version.

(iii) For original data files that are stored in the format of one of the major statistical packages other than Stata, we recommend using a program called Stat/Transfer to create the importable version of the file. With Stat/Transfer, you can easily convert a data file stored in the format of any of the major packages (including SPSS, SAS and R) into Stata's **.dta** format. (See www.stattransfer.com for information about Stat/Transfer.)

(iv) There are several possible ways of handling original data files that are in Excel's **.xls** or **.xlsx** format.

> One approach is simply to create a delimited text version of the file using Excel's "save as" function, specifying the storage format as either tab-delimited text or comma-separated values (**.csv**). The only other modification that would then be necessary would be to remove any extraneous notes from the file, as described for (ii) above.
>
> A second approach would be to use Stat/Transfer, which converts Excel files to Stata format exactly as it does for data files in SPSS, SAS or R format.
>
> Finally, if you are using Stata 13, that version of the program is able to import data from Excel-formatted files; provided there are no extraneous notes or comments in the original Excel file, the importable version can simply be an exact copy of the original.

If an importable data file is an exact copy of an original data file, it should have the same name as the original. (For the sake of keeping your work organized, however, you should keep two copies of the file—one in your "Original" data folder and one in your "Importable" data folder.) If you modify an original data file in any way in the process of creating the importable version, the importable file should be given a new name. The new name should be similar to the name of the original data file but the ***original*** prefix in the name of the original file should be changed to ***importable***, and the extension should be changed as appropriate. For example, ***original _pwt.txt*** would be renamed ***importable_pwt.txt***, and ***original_UNaid.xls*** would be renamed ***importable_UNaid.csv***.

II. Metadata

Begin by creating a new folder called "Metadata," and save it in the top level of your "Documentation" folder. Next create another new folder called "Supplementary Metadata," and save it in the top level of your "Metadata" folder.

We use the term metadata to refer to information about or documentation of your original data files. What kinds of metadata are appropriate and necessary will vary a great deal, depending on the nature of your original data file. In some cases, documentation that accompanied the original data file, such as a codebook and/or a users' guide, will contain all or some of the necessary metadata. We will refer to documentation that the producers or distributors made available with a data file as "native" metadata. When no native metadata are available, or when the available metadata need to be supplemented with additional information or explanatory comments, you will need to write some or all of the metadata yourself.

The first step in assembling your metadata is to decide what information or documentation you should provide for each of your original data files. The general principle guiding these decisions is that the metadata for a data file should contain all the information a user would need to understand the contents of the file, such as variable definitions, units of measurement, coding schemes, and sampling methods. In practice, making these decisions requires judgment.

Once you have decided what metadata you will provide for each of your original data files, you should create a document titled ***metadata.pdf***, which should be stored in the top level of your "Metadata" folder. This document should consist of one section, or "entry," for each of your original data files. Each entry should begin with a brief bibliographic citation of the original data file, in a format that would be appropriate for the reference list of a research paper.

The entry for the data file should provide information about any native metadata for the file that is included in your documentation. Give a bibliographic citation for each item of native metadata, and describe briefly the relevant information it contains. If the item is available from a stable and publicly accessible source, explain how a user can obtain it. If the item is not easily available from a public source, save a copy of it in the "Supplementary Metadata" folder you created, and make a note in the ***metadata.pdf*** file indicating that it can be found there.

If you have written metadata for an original data file yourself, that information can usually be included under the entry for the data file that appears in the ***metadata.pdf*** file. If presenting the metadata you have composed for a data file in a separate document would be more effective (which might be the case, for example, if it is very long or consists of tables or figures), then create a separate document containing this metadata, and save it in the "Supplementary Metadata" folder. In the ***metadata.pdf*** file, under the entry for the relevant data file, give the name of the metadata file you created, describe briefly the information contained in that file, and make a note indicating that the metadata file can be found in the "Supplementary Metadata" folder.

III. Do-files

The number of do-files you include in your documentation and how they are organized may vary for a variety of reasons, including the number of data files you have and how they are organized. The following instructions are written for relatively simple situations, in which the number of data files is not too large and no unusual complications arise.

To begin, create a new folder with the name "Do-files," and save it in the top level of your "Documentation" folder.

You will create three do-files to include with your documentation: one (titled ***import.do***) that imports the data from the importable files you created and then saves them in Stata's **.dta** format; one (titled ***cleaning.do***) that combines and processes the data as necessary to create the final data set used in your analysis; and one (titled ***results.do***) containing the commands needed to generate all the results you report in your thesis. All three of these files should be saved in your "Do-files" folder.

It is very important to include thorough comments throughout all of the do-files you create. These comments should be detailed and clear enough to make it possible for someone not familiar with your project to understand every step of data management and analysis executed by the commands in the do-file.

> 1. <u>import.do</u>. The purpose of the ***import.do*** file is to import the data from each of your importable data files that is not in **.dta** format, and then to save the imported data in a new file that is in **.dta** format. After creating and running ***import.do***, you will therefore have a **.dta**-formatted version of each of your raw data files.
>
> If all of your importable data files are in **.dta** format, you do not need to create an ***import.do*** file.
>
> For each of your importable data files that is not in **.dta** format, ***import.do*** should contain commands that import the data from the importable file into Stata (usually with the `insheet` or `infix` command) and then save it in **.dta** format.
>
> Each new **.dta**-formatted file created by ***import.do*** should be given a name that corresponds to the name of the importable version, but with the prefix "***import_***" dropped from the name and with the extension changed to **.dta**. For example, if the name of the importable file is ***import_UNaid.txt***, the file created by ***import.do*** should be saved with the name ***UNaid.dta***.

If two or more of your importable data files were not in **.dta** format, your ***import.do*** file should contain one block of commands for each of your importable data files. Each block of commands should instruct Stata, as described above, to import the data from one of the importable files, and save the file in **.dta**.

Note that any **.dta**-formatted files that were created by commands in the ***import.do*** file should not be included in the electronic documentation that you turn in with your thesis; it is not necessary to include these with your documentation because anyone interested in replicating your analysis can create them simply by running ***import.do***.

2. cleaning.do. The purpose of the ***cleaning.do*** file is to process your data in whatever ways are necessary to create the final data set or sets that you will use for the analysis you present in your thesis.

Since there is a great deal of variation in the number and structure of the data files that students use in their senior theses and in the how the data need to be organized in preparation for analysis, it is impossible to give a comprehensive description of the specific commands that should be contained in ***cleaning.do***. In general terms, however, ***cleaning.do*** should take the data from the **.dta**-formatted versions of your data files, and then clean, merge, manage and organize them as required to create and save the final data set or sets that will be used for your analysis.

If your analysis will be conducted using just one final data file, the command in ***cleaning.do*** that saves the final data file should give it the name ***final.dta***. If you need to use data from more than one processed file to generate your results (e.g., some of your results might be generated using a file with data on individuals from different countries, while other results are generated from a file in which the data has been aggregated to the country level), the commands in your ***cleaning.do*** file that save the processed data files should give them names that include the prefix "***final_***" (e.g., ***clean_individual.dta*** and ***clean_aggregated.dta***).

Your ***cleaning.do*** file should contain just the minimal set of commands necessary to create your final data set or sets. You will almost certainly spend a good deal of time exploring and experimenting with your data before deciding exactly what analyses you want to present in your paper and how your final data need to be organized so that you can conduct the analyses you choose. In the course of that exploration and experimentation, it is likely that many commands will accumulate in your ***cleaning.do*** file, including many that turn out to be dead-ends or unnecessary for what you ultimately decide to do with your data. All such extraneous commands should be deleted, so that the ***cleaning.do*** file you turn in with your documentation contains only commands that do something necessary to prepare your data for analysis.

Note that the clean data file or files created when you run ***cleaning.do*** should not be included in the electronic documentation you turn in with your thesis. An independent researcher could create them by running ***cleaning.do*** (after running ***import.do***).

3. <u>results.do</u>. Your ***results.do*** file should contain commands that generate the results you report in your paper, using data from your final data file or files.

For every numerical result, figure or table presented in your paper, ***results.do*** should contain a command that opens the appropriate clean data file, followed by a command that generates the output showing the result. Each command that generates the output for a result presented in the thesis should be preceded by a comment that indicates where the result appears in the thesis (e.g., the number of the table or figure being produced, or the number of the page of the thesis on which the result is reported).

IV. A readme file. The documentation for your thesis should include a **.pdf** document, titled ***readme.pdf***, that gives an overview of the various files that you have assembled to document your project. In particular, the readme file should:

(i) list all the files included in the documentation, describe the content and format of each file, and outline the organization of the files into folders and subfolders; and

(ii) explain how the files included in the documentation can be used to replicate the results reported in the paper.

The instructions for replicating the results of the thesis [described in item (ii) above] should be precise and detailed. The objective is to ensure the transparency of all the steps required to access, process and analyze your data, so that it would be a straightforward task for an independent researcher to replicate all those steps.