How to design, organize and document data files

Version June 2020

For whom?
- BSc/MSc thesis students, PhD candidates and their supervisors, project leaders

Why?
- To facilitate efficient data analysis and minimize errors.
- The datasheet can be imported/exported to be used straight away in R, SPSS, SAS or other statistical software packages and the structure permits statistical analysis according to a wide range of different models.
- Data files should be self-explanatory such that the data are accessible and understandable to supervisors/ project partners, which facilitates communication and teaching.
- Data are secured for future use.

Type of file
- The data of each study should preferably be presented in a Microsoft Excel Workbook. Datasets that are too big for MS Excel can be presented in other formats that are commonly accessible, such as csv files. Those can be organized per study a file folder.
- For a research project that contains several (sub)experiments the data can be stored in several files/folders.

File (xls workbook) or folder structure
- Each workbook (or folder) contains different worksheets (files):
    - ORIGINAL DATA
    - CALCULATIONS
    - STATS INPUT
    - CODES
    - DOCUMENTATION
- There can be more than one worksheet with original data and with calculated data in one workbook if you work with multiple data sets. Make sure you mark them clearly.
- Additional worksheets (e.g. lab forms) can be added to your workbook, according to personal preferences. However, make sure the structure and content of your work book is clear and manageable in size.

FOR AN EXAMPLE OF A DATA FILE in Excel SEE: **SBL-ExampleDataFile-2014.xls**

Worksheet 1: ORIGINAL DATA (More than one is possible)
- Only original data, no calculations.
- There is only one table, that contains no empty columns or rows.
- Rows (horizontal) contain the samples, columns (vertical) contain the independent (e.g. factors, blocks) and dependent variables (parameters measured).
- Each row (sample) should be labeled with a **unique** ID code.
- When applicable add your own sample code (ID) next to the column with the lab code from the lab report. In the DOCUMENTATION SHEET you include a reference to the lab report.

- Names of variables should not contain symbols or characters such as %, $, & etcetera, and no spaces. Use short variable names (e.g. not more than 12 characters). The short variable names should be fully explained in the CODES sheet.
- Missing values are indicated with a dot (e.g. in SPSS) or NA (in R). Any zero should be a true zero.
- Any changes to the data file (e.g. removal of an extreme outlier) should be recorded and the reason should be indicated, together with the originally obtained value (e.g. by insert comment). Note that extreme outliers can only be removed when there are strong indications that something went wrong (for example a value that is physically impossible, or in hindsight the sample appears not to belong to the target population. Just the fact that the value deviates from your other data is not a good reason to omit it!
- The original data are sacred, do not mess with it. There is only 1 original data set and any changes made to that should automatically translate into changes in calculated values elsewhere in the workbook.

Worksheet 2: CALCULATIONS (MORE THAN ONE IS POSSIBLE)

- Includes all calculations. It should be clear how they are calculated from the raw data so do not "copy -paste to-values"! Instead make direct links to the original values or information by typing "=" followed by the cell address on the raw data sheet. This is also essential for your factor levels, so that correspondence of data and factor levels can be more easily checked.
- As there is only one unique sheet containing the raw data, derived parameters will automatically be recalculated when there is a change in the raw data file. Similarly changes (e.g corrections) to the data input should only be made in the raw data sheet, to prevent that different versions of raw data files circulate.
- Make sure that the unique ID code (e.g. unique sample number) is transferred also to the calculations file through referencing, as this facilitates sorting and data checks.

Worksheet 3: STATS INPUT

This worksheet should contain the following information:

- The STATS INPUT sheet is a sheet with final data (based on the calculated values) that can easily be saved as a csv file that can be opened in R or any other statistical package.

NB: For the worksheets that contain ORIGINAL DATA, CALCULATIONS AND STATS INPUT use the appropriate number of decimals in such a way that the number of significant digits corresponds with the precision of the observation. For example a pH value with 5 decimal places creates false precision; the last 3 decimals do not have any practical relevance. On the other hand a pH of 5 is not precise enough.

Worksheet 4: CODES

This worksheet should contain the following information:

- The explanation of all codes and short variable names used in Worksheet 1 and 2 to indicate independent and dependent variables
- Including the units of expression (e.g. [mg kg^{-1} of dry soil]).
- This assures that the data file is self-explanatory and can be understood by your colleagues/supervisor

Worksheet 5: DOCUMENTATION (RELATED FILES)

This worksheet presents basic information about the project (title, persons involved) and a list of related files with metadata, scripts, original lab files and publications. The exact information required will depend on the project.

Examples for related files are:

- A map with the layout of the experiment, including a description of the statistical design (in terms of factors and levels)
- GPS coordinates of sampling points
- A text file with methods or protocols, such as:
    - Sampling methods (corer type and diameter, sampling depth, single or bulked samples, sampling dates
    - Sample preparation (storage time and temperature of fresh samples, drying temperature, sieving)
    - Experimental conditions (e.g. incubation temperature and time, moisture conditions)
    - Protocols used for laboratory analyses.
- Lab files reporting the results of your analyses. Add information about archiving and tracking codes
- Model scripts, e.g. matlab, python, R, SPSS
- Publications

ADDITIONAL WORKSHEETS?

The core of the database should be in line with the guidelines to allow for use as a reference. Anything else can be added case by case. Additional worksheets can always be added to the same workbook according to personal preferences, or a reference to a separate file can be added in Worksheet 5 DOCUMENTATION (RELATED FILES).

FINAL PRODUCT

- The set of files for your project (e.g. an MSc thesis, PhD thesis chapter or publication) should allow for a complete reconstruction of the data collection and analysis.
- The set of files should be combined into one zip file and handed over to the supervisor or project leader who will make sure it is stored in a secure place on a dedicated network drive.
- For MSc and BSc thesis students, the data files must be submitted to the supervisor before finalization of the thesis project.
- For PhD students the data for each published paper should be submitted to the supervisor as soon as the paper is accepted for publication. At the end of the PhD period, all data presented in the thesis will be stored.

==========

In case of remarks/suggestions please contact Paolo Di Lonardo