# Data Organisation in Spreadsheets

This is a summary of: Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989)

See also Carpentries ([Spreadsheets for Ecologists](https://datacarpentry.org/spreadsheet-ecology-lesson/) or [Social Scientists](https://datacarpentry.org/spreadsheets-socialsci/)) for more detailed information.

## Create spreadsheets that are less error-prone, easier for computers to process, and easier to share:

**1. be consistent**

 *Use consistent codes for categorical variables*

*Use a consistent fixed code for any missing values (NA)*

*Use consistent variable names and subject identifiers*

*Use a consistent data layout in multiple files, as well as consistent data formats*

*Use consistent phrases in your notes*

*Avoid spaces in cells*

***Use consistent (file) names***

 *Avoid spaces and special characters, except for underscores \_ and hyphens –*

 *Choose short and meaningful names*

*Use consistent formats for dates:*

2. **write dates like YYYYMMDD** (“ISO 8601” standard)

Use a plain text format for columns in an Excel worksheet that are going to contain dates, so that it does changes the values. (Right click column -> Format -> Cells -> Text)

3. **do not leave any cells empty** (if no value, fill in NA)

4. **put just one thing in a cell and don’t merge cells**

5. **organize the data as a single rectangle**

with subjects as rows and variables as columns

single header row

6. **create a data dictionary**, a separate file that addresses:

 The exact variable name as in the data file

A version of the variable name that might be used in data visualizations

A longer explanation of what the variable means

The measurement units

Expected minimum and maximum values

This is part of the *metadata* that you will want to prepare: information *about* the data. You will also want a ReadMe file that includes an overview of the project and data.

7. **do not include calculations in the raw data files** (write protect it)

 analysis and visualization should happen separately to avoid damaging the raw data

8. **do not use font color or highlighting** as data

Instead, add another column with an indicator variable (e.g., ”trusted” with values TRUE or FALSE).

9. **make backups**

 Multiple locations

 Version control (Subversion or Git)

10. **use data validation** to avoid data entry errors

 “data validation” feature in Excel (see [online tutorial](https://support.office.com/en-gb/article/apply-data-validation-to-cells-29fecbcc-d1b9-42c1-9d76-eff3ce5f7249?redirectSourcePath=%252fen-us%252farticle%252fApply-data-validation-to-cells-c743a24a-bc48-41f1-bd92-95b6aeeb73c9&ui=en-US&rs=en-GB&ad=GB))

12. **save the data in plain text files** (Save as .csv)