DIY: FAIR Spreadsheet

Why should I make my spreadsheet data FAIR?

Spreadsheets are widely used in the research community to collect data, visualize data and to analyze the data.
Making sure the data you enter in a spreadsheet is clearly defined, validated, sufficiently described and saved in an widely used format enhances its understandability for others. This makes your spreadsheet **FAIR** (https://www.go-fair.org/fair-principles/) because it benefits data accessibility, interoperability and reusability.

And let's be honest: it also helps you to **keep track** of and **understand** your data!

What are the do's when working with spreadsheets?

- **Be consistent**. Enter and organize your data consistently from the start. Settle on conventions within your team and stick to them!

- **Avoid mixing languages when describing any variables. Best practice is to use English** in the spreadsheet itself, but also in its documentation file: especially if you plan on sharing the data with others!

- **Use meaningful variable names**. The variable name should state clearly what the variable stands for. If using abbreviations, settle upon them with the rest of your team and write them down. Don't just name a variable for example like this: "mv". Instead write out its whole name, for example "mango_velocity".

- **Use meaningful codes for categorical variables**. For example when describing an animal's sex, use "male" and "female" instead of not "0" and "1". If possible use established codes for categorical variables within your project.

- **Use consistent codes for categorical variables**. For example, when describing an animal's sex, use a single common value for both males and females (e.g., "male" and "female"). Don't sometimes use "M", "male" and "Male" etc. to describe the same value.

- **Use consistent variable names**. If in one spreadsheet you have a variable called "Mangoes_eaten_day1", and this variable is also present in other spreadsheets, call it exactly the same in all those files.

Author: Ewa E. Bres, Research Data Service Center at the University of Bonn

- **Settle on a consistent layout of the data**. Put all the variables in columns and observations in rows.
- **Keep your raw data unaltered**. Do calculations and analyses in a copy of the file.
- **Make a README.txt or a Data Dictionary** that describes your dataset. Include a variable legend and enumerate any steps you took to clean up your dataset.
- Take care of your data: **validate** the data, **make regular backups**, **use version control software** to track any changes you make and once the data is complete, **write-protect the file**.
- **Use an open, non-proprietary format like CSV or TSV to export your work**. This ensures the data will be readable and accessible also in the future!

What are the don'ts when working with spreadsheets?

- **Avoid spaces, numbers and special characters in column headings**. Special characters include: $, @, %, #, &, ä, *, (,), !, / etc. These characters should be avoided because they have different meanings in different programming languages. Instead use either underscores or hypens.
- **Don't mix zero and null values**. Instead differentiate between them. Zero is a value meaningful in maths. A null value on the other hand can be seen as a "placeholder" for data values that are not known or not specified.
- **Don't merge cells or leave the cells blank**. Leaving cells blank (for example when a single value is meant to be repeated) makes it harder to determine whether the cell has a value assigned to it or not.
- **Don't put more than one piece of data in a cell**. For example if you have a plate with wells, don't put the variable "plate" and "well" in the same cell. Instead make two separate cells for "plate" and for "well".
- **Don't create several worksheets in one spreadsheet**. Having only one worksheet makes it easier to handle your data and export it to formats like CSV.
- **Don't highlight your cells or use different font colours**. For example, if you need to discriminate between correct values and outliers, create a separate column for that.

Author: Ewa E. Bres, Research Data Service Center at the University of Bonn

What to look out for?

Be especially careful when entering values like gene names and dates. Excel has a tendency of reformatting information into dates! For example the gene name "Oct-4" can be reformatted to October 4.

Scientists can nowadays use different spreadsheet software to analyze their data! In each software type, data is read and handled differently. To limit formatting issues, as well as increase readability and reusability of your data, **save it in an open format** like CSV or TSV.

If you by mistake forget to select a cell in your spreadsheet and just start to type, no text appears. However, what you just typed can automatically get saved to a random cell in the file. Such cells may be hard to find and can alter your raw data.
To avoid this, **make a copy of your raw data file and do the analysis there**! Like this, your raw data stays unaltered and you can always go back to it!

How to clean-up spreadsheets: helpful tools

- **OpenRefine**
  https://openrefine.org/
- **Frictionless Repository**
  https://repository.frictionlessdata.io/
- **Open Data Editor**
  https://opendataeditor.okfn.org/

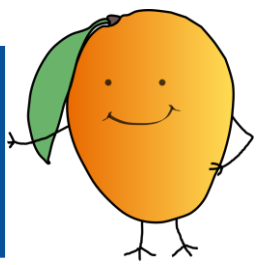
Author: Ewa E. Bres, Research Data Service Center at the University of Bonn

Further reading

- **Data Organization in Spreadsheets**
  https://doi.org/10.1080/00031305.2017.1375989
- OSF Guidelines: **How to make a Data Dictionary**
  https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary
- **Data Carpentry lessons** on Data Organization in Spreadsheets for Ecologists and Social Sciences
  https://datacarpentry.org/lessons/
- **MIT Guidelines to file organization and version control**
  https://libraries.mit.edu/data-management/store/organize/
- **RightField**: tool for adding ontologies in Excel
  https://rightfield.org.uk/about.html
- **Research Data Service Center webpage** & **Confluence space**
  https://www.forschungsdaten.uni-bonn.de/en
  https://confluence.team.uni-bonn.de/x/9EJqBg
- **Excel Archival Tool** to programmatically convert Microsoft Excel files into open source formats suitable for long-term archival
  https://github.com/mcgrory/ExcelArchivalTool/blob/master/UserGuide.pdf
- **Spreadsheet data gone bad**
  https://eusprig.org/research-info/horror-stories/

QUESTIONS? COMMENTS?

GET IN TOUCH: researchdata@uni-bonn.de



Author: Ewa E. Bres, Research Data Service Center at the University of Bonn