

Good practices for data spreadsheets

This is a document describes good practice using spreadsheets (e.g., Excel) files containing measured data. The basis of these practices is experience biometricians and bioinformaticians have when dealing with data. The aim is definitely not to make life difficult, but to reduce time spent on cleaning data by you as well as others (e.g., biometricians). Time is valuable and so are you. A lot of these suggestions also apply to files you use for analysing your data.

CONTENTS

1	Guiding Principles	1
2	Data Stewardship	1
3	Storage	2
4	File Name Conventions	2
5	Text Files, Paper Datasheets, and Machine-generated Files	2
6	Data Entry	3
7	Data Checking	4
8	Data Protection	4
9	See Also	5

1 GUIDING PRINCIPLES

- It should be clear what the data are and who should be credited for it
- Data should be easy to use by anyone
- Data should be easy to search for, find, and be understandable by anyone in PFR at any time
- Your raw data is your most precious resource. This is what you can't reproduce again. All other steps in the analysis pipeline can be repeated
- There should be a clear trail from the raw data to the final analyses and figures to enable reproducibility of the results

2 DATA STEWARDSHIP

A Data Steward works within a science team or science project, ensuring that the team or project's data are appropriately managed. The developing [PFR Data Management Framework](#) proposes that this role be formalised.

3 STORAGE

- Data should be in an accessible location for collaboration and should be backed up and version-tracked. In general, for Excel files this will be on iPlant. Do **not** use your C: drive (not backed up) or H: drive (inaccessible by other PFR staff).

4 FILE NAME CONVENTIONS

- **Define (and adhere to) a naming convention for your project. This will be specific to your research group but should be readily understood by all in your group.**
- If possible, avoid spaces in filenames as these can cause difficulties when accessing files using some command-line tools
- Use either [camel case](#) (i.e., UpperAndLowerCase) or underlines to separate words (e.g., TrialXYZ2009LabData.xlsx or trialXYZ_2009_LabData.xlsx). Note: if using camel case, when searching for a file in iPlant, it is necessary to use wildcards (e.g., to find 'camelCase' one needs to search for 'camel*' not just 'camel')
- Use dates in reversed order (e.g., YYYY_MM_DD, 2010_03_15) which will give you a better [alphanumeric](#) sorting order when listing directories
- It is safest to use just letters and numbers - in particular avoid [metacharacters](#); i.e., . (except to delimit the extension) [] { } () ^ \$ | ? * and +.
- Use filenames that will be meaningful to colleagues.
- Windows limits filenames to be less than ~255 characters.

5 TEXT FILES, PAPER DATASHEETS, AND MACHINE-GENERATED FILES

Metadata for data stored in files other than spreadsheets will generally need to be kept in additional files and cannot be added into the raw data file. The metadata that is required is the same as is required for Excel files, as described below (ProjectDescription, VariableNames, TrialPlan, Diary, etc.).

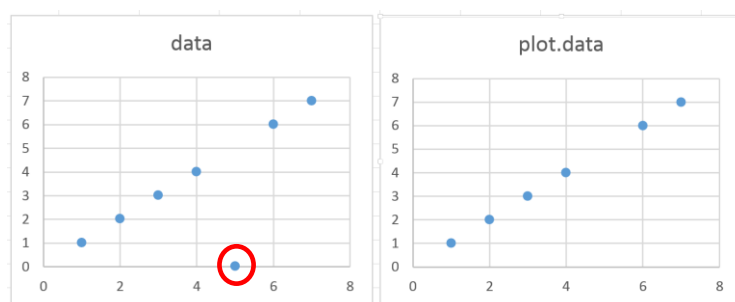
- Files generated by instruments (e.g., temperature loggers, qPCR output, other analytical machines) should be stored as generated, and not be modified. The source of the files (type, e.g., SoftMax plate reader, manufacturer etc.) and any machine setup or settings information should be recorded as metadata
- Export both the machine-generated file and a text file (either as comma delimited .csv or space/tab delimited .txt). Consider these files *raw data* and do not edit them directly. These text files are easily read by analysis software
- If data from machines or in text files needs to be modified, modifications should ideally be done with code (such as in R, Genstat etc.)
- Paper data sheets should be scanned and saved on iPlant. The paper originals should be stored in a safe location, and this location noted in the metadata (e.g., in the ProjectDescription sheet of an Excel file).

6 DATA ENTRY

- A specific symbol for missing values will allow later users to distinguish between blank cells due to not entered data and genuinely missing data. The symbols * and NA are defaults for many data analysis packages. Most data analysis packages will enable you to specify the missing value symbol. Unfortunately graphs in Excel treat missing data characters as zero. We have found that the easiest way to deal with this is to:
 - Code missing data for use in analysis as * or NA
 - NB: use only one missing data indicator within a dataset, i.e., * or NA, not both; use a comment column to note any reasons for the missing data
 - Create a separate “calculated” column for your graph data (preferably on a separate sheet)
 - Use a formula to replace a symbol with an NA()

	A	B	C	D
1	data	plot.data		
2		<code>=IF(ISNUMBER(A2),A2,NA())</code>		
3	2	2		
4	3	3		
5	4	4		
6	*	#N/A		
7	6	6		
8	7	7		

- This will cause a blank to be plotted, not a zero



- A lot of time in data cleaning is taken up by having variable names or treatment codes that have inconsistent spellings. For example, “Hot”, “Hot “ (with a trailing space), and “hot” will be considered as separate levels by analysis software. Use [data validation](#) to select treatment levels from a pre-entered list.

7 DATA CHECKING

- Any changes to the raw data must be recorded either in a log file or in a comments column on the raw datasheet. Within-cell comments should be avoided as these are easily overlooked.
- Use [data validation](#) to issue a warning if numbers are outside the expected range. Conditional formatting can also be used for this.
- Use simple checks of minima and maxima, either as a couple of rows along the top (which can be removed once all data collection is complete) or as a pivot table
- Checks for spelling mistakes in category names can be made in Excel or other spreadsheet programmes by using a filter function for that variable. However, this will not pick-up differences in capitalisation, nor distinguish between names that differ only by having trailing/ leading spaces (“Control” vs “ Control”)
- The formula ‘COUNTBLANKS()’ can be useful for detecting missing values
- There is a [guide to useful tools for data management in Excel on the iPlant Data Management site](#)
- Be aware that Excel may automatically apply irreversible formatting to your data.

According to Microsoft support:

- If a number contains a slash mark (/) or hyphen (-), it may be converted to a date format
- If a number contains a colon (:), or is followed by a space and the letter A or P, it may be converted to a time format
- If a number contains the letter E (in upper-case or lower-case letters; for example, 10e5), or the number contains more characters than can be displayed based on the column width and font, the number may be converted to scientific notation, or exponential, format
- If a number contains leading zeros, the leading zeros are dropped.

Certain types of data (e.g., clone identifiers, gene names, plate coordinates) are particularly susceptible to these issues. To avoid the problem, make sure to first select the whole spreadsheet and Format -> Cells -> Number -> Text when pasting data into Excel (the default is “General”). If using this approach, even genuine dates will be regarded as text, and very long data strings (e.g., sequence data) may be converted to hash (#) characters. If this occurs, it is necessary to switch these cells back to “General” format. There are plenty of people within PFR who can help should this be a problem.

8 DATA PROTECTION

- Your raw data is your most precious resource. This is what you can’t collect again. All other steps in the analysis pipeline can be repeated (and should be repeatable). Once you have finished entering and checking data lock the sheet (or the cells in the sheet to prevent any errors).
- Use one file as your data file and don’t make copies. There should be an unbroken and repeatable trail from your raw data to your final analyses, figures and tables.

9 SEE ALSO

On iPlant

- [Examples of good use of Excel templates.](#)
- [Key skills in Excel](#)
- Using [pivot-tables](#) to summarise data.
- [Reading University guidelines for use of spreadsheets](#)

External resources

- [Spreadsheet design](#)