

Guide to Using PFR Excel Templates

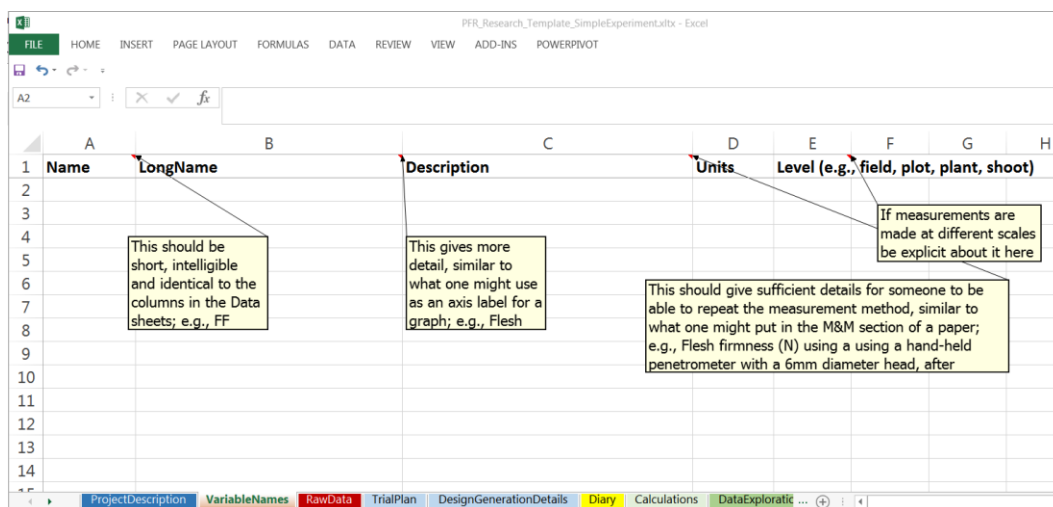
- If you consult a biometrician about your study design, they will usually supply you with a formatted Excel data-entry sheet. In cases where this hasn't been done, use a PFR data template. In Excel, choose File > New > Plant & Food Research when creating a new data file. If this doesn't work, the templates can be found [here on iPlant](#). This can be the basis for creating project-specific templates; for example, if the VariableNames sheet is the same or similar for many projects.
- The templates should be modified as necessary: change the names/colours and insert other sheets as necessary to complete all information about the trial. For example, a diary sheet can keep track of measurement dates. Including a schematic of the trial plan and the code for the trial randomisation is also recommended.
- The three key sheets that should be filled in are the ProjectDescription sheet, the VariableNames sheet and one (or more) sheets with the measured data (DataSheet)
- It is beneficial to later users if calculations and formula are on separate sheets to the raw data (CalculatedData or DataExploration sheets), so the raw data is clearly identified, and modifications to calculations/summaries don't interfere with importing data into other software.
- Changes should be documented, preferably with an audit trail. **Never** password protect an Excel sheet or workbook
- Once a spreadsheet is considered final, lock the sheet containing the raw data. This spreadsheet should be considered the master version. There should be a reproducible trail from this raw data sheet to all downstream analyses and documents.

1 EXCEL TEMPLATE SHEETS

1.1 ProjectDescription Sheet

- This is sometimes called the Cover Sheet or the Narrative Sheet
- The first sheet should provide information on the project description (ProjectDescription sheet), project code(s), name and contact information for the researcher. Including names of people who collected and entered the data will help when people have questions at some future date.
- [Hyperlinks](#) to documentation can be used (e.g., reports, milestone documents), but don't rely exclusively on links as they can break. Generally, it is safer to keep files in the same folder or iPlant library, since they will be kept together if a folder is moved to a different place.
- To facilitate search-ability, the Project Description information should comprise a key name and the value for it (i.e., information about the item type and the second the content; e.g., author: J Smith; year: 1984; crop: Apple) with at least some of the following fields:
 - Title: a brief title summarising the contents of the file
 - Project Number: the project code under which the data was initially collected. If it unofficial exploratory work this can be left blank
 - Project Leader(s): the name(s) of the project leader(s)
 - Data Steward: the person responsible for managing and archiving the data
 - Credits: who should be acknowledged for collecting the data
 - Description: a free form description of the data
 - Links: Links to relevant documents (e.g., project proposal, data analysis report, publications resulting from the data)
 - Keywords: One key word per cell to allow for searching. What information do you think will be useful to help you or others search for your data? Think about what your future self might want to search for, or what would you want to search for in work that you did five years ago. What was the question, study species/system/crop, key effects, methods used, specific machine, site, key methodology..., anything else?

1.2 VariableNames Sheet



- This sheet describes the columns in the DataSheet(s) in more detail. See [example files](#) for some different examples.
- It should contain at least three columns: the first “Name” column exactly matches the column headers in the DataSheet(s), the second “LongName” column describes it in more detail as one might in a figure caption, and the third (“Description”) gives the detail that might be in the Methods and Materials section of a manuscript (i.e., sufficient to allow another researcher to replicate the method exactly. For example,

Name	LongName	Description
Rep	Replicate	Three replicates
Treatment	Treatments applied	A: Chemical A
		B: Chemical B
		W: Water control
Yield	Plot yield (g)	Total Yield (g) at harvest, corrected by the combine harvester to 15% moisture

- Two additional columns are “Units” in which to record the units of measurement (if any), and “Level” to record the what may be referred to as the “unit of observation” which is typically the thing that was measured or that the data pertains to (e.g., tree, branch within a tree, leaf, fruit, box of 20 fruit, site, field plot, human subject, etc.).
- The first column can be created by copying the row of variable names from the RawData sheet and pasting them transposed.
- Categorical (e.g., treatment levels) or score variables should be explicit about coding scheme used.
- Every column in the Data sheet(s) should have an entry in the VariableNames Sheet, but it may not be necessary to complete the third (or even second) column in the VariableNames sheet if the term is full self-explanatory.
- Include explanation of any calculated columns: such columns should generally be in a CalculatedData sheet.

- Even 'obvious' terms are not always so for researchers from a different discipline (e.g., 'block' to some is obviously the orchard block, but to others would obviously be a feature of the statistical design of the experiment).

1.3 RawData Sheet

	A	B	C	D	E	F	G
1	ID	Trt	Block	TreeSize	Disease	FruitWgt	Comment
2	PFR_0001	A	1	Tall	Present	156.3	
3	PFR_0001	B	1	Tall	Absent	123.6	
4	PFR_0002	A	1	Medium	Present	236.2	
5	PFR_0002	B	1	Medium	Present	186.3	
6	PFR_0003	A	1	Small	Absent	232.4	
7	PFR_0003	B	1	Small	Present		
8							
9							
10							
11							
12							
13							

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 throughout the process).

Raw data should occur as **a single contiguous block with a single header row followed by the rows of data, with one observation per row**. Each row be uniquely identified ('ID') either with a single column (Plot Number), or multiple columns (Plot, Plants within Plots).

If using a paper-based data collection method, it is best if its format matches that of the RawData sheet (i.e., a single contiguous block with one record per row – data randomisation packages will give you a datasheet in this format) although this may not always be practical and some compromises may need to be made. Design both the paper data sheet and the Excel datasheet before collecting the data.

- The column names used in the RawData sheet should generally be fairly short with extra detail provided in the VariableNames sheet. Some people like to have an explanation of the column names above the single header row (e.g., as recommended in [the Reading University guidelines](#)). Similarly, one might wish to have a couple of rows above the header row giving the minima and maxima for data checking (see below). Calculations should generally be put to the right of the data block, or ideally on separate sheet (which emphasises that it is calculated data).
- Repeated measures (multiple assessment dates; counts for each of several disease scores etc.) can be put into separate columns if this facilitates data collection. However a new row for each measurement will reduce the need for downstream data manipulation.
- The column names in the header row should:

- contain only alpha-numeric characters (i.e., no [metacharacters](#) or other punctuation)
- begin with a letter (not a digit/ number)
- exactly match the names in the Variable Names sheet
- be unique
- be short but decipherable
- be consistent, including [case sensitivity](#), across years and similar trials
- follow a system which is preferably documented (e.g., 'Harvest date' and 'Harvest maturity' should be abbreviated consistently - perhaps harvDate and harvMat, not harvDate and HrvMat)
- Freeze the panes so that the header row and factor (e.g., treatment) columns are always visible.
- Generally there should only be one type (i.e., numbers, dates, text etc.) of data in a column. For example, don't put "Dead" in a column for flowering date, but use additional columns for such data or comments.
- Be explicit about separating present/absent versus numeric data into separate columns (for example germinated or not, if germinated seedling weight). Coding non-germinated plants as missing for weight is clearer for subsequent users/analyses.
- Missing data are traditionally coded as * or NA. Choose one and be consistent. Avoid using missing value symbols to mean anything other than a missing value. See Data Entry section in [Good Practices for Data Spreadsheets](#).
- Potential misinterpretation of data by Excel can be avoided by formatting the cells before entering data e.g., a date is often interpreted as text or text interpreted as a date or a date interpreted as being month/day/year instead of day/month/year).
- Don't put blank rows or columns between rows or columns of data. Instead use colour and other formatting to highlight aspects of the sheet.
- The data must be able to be understood correctly if any formatting is removed – do not use colours, fonts, and other formatting as the sole mean to code for data (this formatting info is not easily readable by other software). Similarly, be sparing with the use of in-cell comments and do not merge cells in the data area. Important information should be entered explicitly into a cell.
- Use [data validation](#) to select from a drop-down list for treatment names, plant names, etc. to prevent inconsistencies and spelling errors.
- Be **very careful** when sorting data in Excel to include all of the relevant data columns – otherwise data rows can become misaligned. **Ideally, do not manually sort data!** Pivot tables and filter can generally be used instead of sorting.
- Avoid use of hidden columns and rows within the data area – these can be easily overlooked by viewers of the spreadsheet, and software that imports Excel data will probably include the "hidden" cells to the surprise of those doing the analysis.
- Columns provided by a biometrician or a randomisation package are there for a reason. Please don't delete them.
- Avoid placing any information **below** the block of data – such things as notes, column means and summary calculations should be placed on a separate sheet from the data.
- Once data collection and checking is complete, protect the RawData sheet so that all data are **read only** within it. This is now the master data sheet and should be used for all

subsequent analyses and summaries (i.e., **refrain from making copies** even for analysis).

1.3.1 Correcting data

- Data in the RawData sheet should match the data as it was recorded. Any modifications to this should be noted, either in a log file or in a separate notes column on the RawData sheet. For example, if the handwritten data value is impossible for your study, any change to the value (e.g., missing or another value) should be noted.

1.3.2 Repeated measures

- Many experiments collect multiple measurements per plot (or plant or other experimental unit). It is often easier to add extra column/s rather than having one row for each of the multiple measurements. Any rearranging data for statistical analysis should be done using a scripted program (e.g., Genstat or R).

1.3.3 Linked experiments

- If data are collected at different scales (e.g., some measurements on a plant, and some measurements about the field or greenhouse conditions), these should be located on separate sheets so that the number of measurements made for each data type is clear for later users.
- If two datasets refer to the same variables (individual, plant, plot ...) use the same names and codes to uniquely identify the variable across both spreadsheets.

1.3.4 Simultaneous data entry

- Data can be entered into the same file simultaneously by two or more users if the file is on iPlant and opened using the Excel Web App. This may not work with more complex files (for example, those with complex formatting).

1.4 Other Sheets

- Figures should not be on the RawData sheet but on a separate sheet (DataExploration sheet).
- Do not embed calculated rows within the raw data (e.g., means), but use pivot tables in a separate sheet instead. Calculated columns are acceptable but preferably on a separate sheet (CalculatedData). This will protect your raw data from being lost/accidentally changed.
- Keeping the diary, the trial design, and data summaries together makes finding these files easier and protects from hyperlinks becoming broken.

2 SEE ALSO

2.1 iPlant

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

2.2 External resources

- [Reading University guidelines to using spreadsheets for data entry](#)
- [Generic advice on spreadsheet design](#)

3 USING THE TEMPLATES WITH STATS AND GRAPHICS PACKAGES

The raw data sheet be easily imported into any stats or graphics package. For example:

Genstat: File > Open, or Use the command `IMPORT`

R: The function `read_excel` in the library `readxl` will read Excel worksheets

Minitab: File > Open

SigmaPlot: Workshop Tab > Import File

4 SHARING THE TEMPLATES

The Excel template files (and accompanying guides) may be given to others outside PFR. The usual confidentiality rules apply for sharing of data.