INNOVATION TO IMPACT

Good Data Practice SOP

November 2023

Version 1

Title	Good Data Practices
Document number	I2I-SOP-041
Version number	1
Date first published	
Date last revised	

Prepared by

Name	Role	Institution
Jack Gillespie	Research Technician	LSTM, I2I
Katherine Gleave	PDRA	LSTM, I2I
LITE Technical Team	Research Technician	LSTM, LITE

Timeline

Version	Date	Reviewed by	Institution
1	07/11/2023	Jack Gillespie and Katherine Gleave	I2I, LSTM
2			

Version Control[1]

Version	Date	Updated by	Description of update(s)
1	4th October 2023	Jack Gillespie	SOP Created

Related documents

- I2I Best Practice SOP Library, June 2023 (https://innovationtoimpact.org/)

Contents

Acronym List

SOP – Standard Operating Procedures

Purpose

This standard operating procedure (SOP) provides guidance on how to document projects to support reproducible research, and good practice in data visualisation.

Background

Good documentation is essential to complete projects and creates a more efficient working environment where coworkers do not have to spend time deciphering the method or results of previous experiments. Data should be recorded with the aim of allowing people who are not involved in the project to understand and reproduce the experiments. Reproducible research is achieved through the open sharing of the complete raw data and the code used to analyse the data[1]. Expanding upon the idea of reproducible research, both the data and the analysis should be fully documented to allow anyone without a background on the project to understand what data is being recorded, and how and why it has been analysed. Data visualisation can be unintentionally misleading, and it is essential to present data that is truthful, and easily understood to allow the audience to draw conclusions on the results of any project.

Computers and humans read spreadsheet differently and a spreadsheet that is appealing to a human may not be readable by a computer. As data is processed by computers, spreadsheets should be created to allow a statistical programme to easily read the data, while remaining readable to the audience. When entering data into a spreadsheet, the data inputted should be consistent throughout[2] and the following recommendations on Tidy Data by Hadley Wickham[3].

Statistical programmes such as Python, R, and STATA allow the use of notebooks like RMarkdown and Juypter. The notebooks allow the author to annotate and describe the data analysis. Annotations should be made to clearly describe how and why the data analysis has been performed with the intention that someone without training in code can understand.

Data visualisation aims to present complex datasets in a way that is both visually appealing and informative. There are multiple types of graphs available to use, however not all graphs are suitable for your data. For example, representing the distribution and sample size of the data is essential to drawing conclusions from the graph. Bar charts and similar plots hide the data's distribution meaning it can lead to false conclusions on how variable the data is, or when comparing groups of data. Box plots are more suitable but presenting the individual data points is best practise. Similarly, the sample size should also be displayed since data based on a small sample size may indicate differences between groups but may not be powered enough to reliably detect a true difference between the groups.

Raw Data Recording

Raw Data Sheets

Raw data should be recorded on templates specifically created for the experiments being conducted. The raw data sheet should aim to capture all the relevant data for bioassays, including a section to record meta data. Raw data can be captured on paper or electronically. An example of a raw data sheet is in the appendix to show good practice. The table below describes best practices in how to complete the raw data sheets.

Description	Good Document Practices
Documents / Records	Documents should be to a standard format to ensure that they are consistent and easy to follow. • **Legible**: Everyone should be able to read and understand what is written regardless of who, where or what has been written. • **Concise**: The document must tell the entire story, but be clear, unambiguous and include information in a way that is understood by personnel. • **Accurate**: Be error free • **Traceable**: Who recorded it, where and why. • **Contemporaneous**: The information should be documented at the correct time frame along with flow of events i.e., as soon as practically possible after the task / event has been completed. • **Enduring**: Long lasting and durable • **Accessible**: Easily available for review at point of use • **Adequate space**: Enough space is provided on documents and worksheets for anticipated handwritten entries. • **Critical entries**: Independently checked by a second person. • **Complete**: All pages must be present. Removal of a page is not permitted, as this would obscure/ omit data present. Reference documents such as SOPs or RAs should be reviewed on a two-year cycle to ensure the content is still accurate, current, and relevant to the facilities, equipment, and personnel.
Using Indelible Ink	All records must be filled out in black indelible ink for long term legibility. Do not use pencil or ink that can be erased or is likely to fade.
Handwritten entries	**Legible**: A document is unusable if it cannot be read (paper/ electronic documents/ records), so care must be taken when completing the entries to ensure handwriting is legible and entries signed and dated. All entries must be made at the time the tasks are performed. Any change made to raw data should be made as not to obscure the original entry: a single line crossed through the error with an indication of the reason for change, dated and signed or initialled by the individual making the change.
Transcription	Transcription of data is ONLY permitted where the original copy is physically damaged, and the writing is obscured. The 'original page' should be clearly marked as 'Original 'and the copy marked as 'COPY'. Where possible, the Copy should be retained by with the Original by being stapled, and a written justification should be recorded on the 'COPY' In the rare occasion the original has been damaged by the chemicals, e.g. solvent or insecticide spillage, and the original has been deemed unsafe, the original must be disposed via chemical waste. A note explaining what occurred and a copy of the chemical waste form must be attached to the transcribed copy.
Use of Approved documents	Only approved documents and worksheets are permitted for use.
Reviewing and Approving	Data/ records or documents should be checked by a second person who did not perform the task/ process to ensure that the information is correct and accurate. A signature and date by the reviewer/approver confirm that a review has taken place. Unsigned documents or records, 'Draft procedures', are classed as incomplete and should not be used to perform any task or considered as evidence that a completed task has taken place. The only exception to this when a procedure has been drafted to aid implementation of new equipment or methodologies as part of validation activities. Data entered on spreadsheets should be checked by a second person to avoid typing errors being overlooked.

Personnel Signatures	Hand-written signatures must be unique to the individual and listed within the facility signature record to ensure that the signature is traceable back to a member of staff (or contractor). Only sign for the task /actions you have physically performed. Personnel are not permitted to sign for another person's role unless this has been formally agreed and documented and approved by the project manager Signatures must never be forged. New staff should sign the signature record during induction, the signature record must indicate the date staff started and when staff exit the department. Electronic signatures must be unique to the individual and listed within the facility and are traceable back to a member of staff (or contractor) and have the same legal consequences as hand-written signatures within the test facility
Page numbering (Refer to attachments)	All documents and worksheets should have page numbers using the following standard '1 of 2, 2 of 2' to indicate the total number of pages and account for all pages being present. Every page or "bundle" (A 'bundle' refer to the number of pages (i.e., a single sheet of paper = 1 page) securely bound together) attached, must clearly cross-reference the protocol/ report/ unique identification number, attachment number (if more than one document), and signature and date. For a 'bundle,' the front page only may be marked up and will be cross-referenced to the document and signed and dated attachment number, number of pages i.e., 1 of 2, 2 of 2 etc. initial and date. i.e., attaching client specific emails to a study protocol, adding additional supporting data to technical reports.
Archive	A history (audit trail) of documents and records are maintained, including changes and deletions to electronic document versions, and archived for a specified period.

Completing Hand Written Data Sheets

<table>
<tr><th>Description</th><th>Good Document Practices</th></tr>
<tr><td>Making a correction</td><td>If a correction is required, the original handwritten record must still be visibly legible: Make one single line through the error. Record the correction close by. If more than one correction has been made, numbering corrections is acceptable when space is limited. Provide a brief comment or use the abbreviation code for why the error occurred (see codes and abbreviations below). Initial the change to identify who made the correction. Record the date of the correction next to the initials to confirm when the change was made.
<table>
<tr><th>Correction Type</th><th>Abbreviation Code</th><th>Description</th></tr>
<tr><td>Recording Error</td><td>RE</td><td>Entry incorrectly made in wrong place or wrong data entered.</td></tr>
<tr><td>Calculation Error</td><td>CE</td><td>When an error on a calculation has been made.</td></tr>
<tr><td>Scoring Error</td><td>SE</td><td>When a data point is incorrect for any reason other than recording error or a rescore is needed. An explanation must be recorded.</td></tr>
<tr><td>Wrong Date</td><td>WD</td><td>Recording the incorrect date.</td></tr>
<tr><td>Late Entry</td><td>LE</td><td>Updating a sheet to complete missing or omitted data after the fact. See omitted data in table below.</td></tr>
<tr><td>Not Done</td><td>ND</td><td>Step or task not competed.</td></tr>
</table></td></tr>
<tr><td>Omitted Data</td><td>If an entry was missed or omitted and requires completion, clearly indicate on the worksheet/document the date the previously missed /omitted activity was completed and mark with LE for Late Entry with time and date of entry. Document an explanation to substantiate the Late Entry and the reason for the delay in recording. Initial and date the change.</td></tr>
<tr><td>Completing all fields on a record i.e. the use of NA</td><td>NA is used to indicate when information/data is not required as it does not apply to a particular task/test. All entries of NA should be signed and dated.
Multiple spaces, e.g. numerous rows, must be clearly marked across the whole number of blank spaces/rows using a horizontal straight line to indicate that no more information/data is required and also to ensure that the record cannot be added to at a later date without appropriate checking or approval. Use ‘NA’ close to the line and sign and date to show that the field/space is not applicable. An explanation may be required for why the field/space is ‘not applicable’.
Marking out a larger space or whole page which is left intentionally blank may be done with a single ‘Z’ line across all the blank spaces/row. Use ‘NA’ close to the Z line and sign and date to shown that the field/space is not applicable. An explanation may be required why the field/space is ‘not applicable’.</td></tr>
<tr><td>Use of Asterisks *</td><td>Where insufficient space permits a fully notated hand-written comment next to the relevant place, an asterisk ‘*’ mark recorded next to the correction can be used, and the same asterisk ‘*’marked by the handwritten comment elsewhere on the document. If more than one asterisk is required on the same page of a document or report a number should</td></tr>
</table>

	be placed after the asterisk etc. *1, *2 by the correction and also next to the related hand-written comment to identify which asterisk refers to which comment.
Checking corrections	Corrections or amendments may be made after a record has already been checked. Any corrections made after the original document/worksheet has been checked/reviewed will require a second check by the same person who originally checked the original document. The person who completed the original check should review the change and ensure it has been made in a compliant manner i.e. clear, legible, accurate and that the original entry is still visible. The change should be reviewed with respect to the impact on the content of the rest of the document/ worksheet. Sign and date the correction.

Recording of Numbers

<table>
<tr><th>Description</th><th>Good Document Practices</th></tr>
<tr><td>Decimal numbers</td><td>If a decimal value is a fraction of 1 then a zero must be placed before the decimal point, for example, record 0.98 rather than .98
Unless otherwise stated in an SOP or Study Protocol, record values to 2 decimal places e.g. 0.01 or 1.00. This is not required where only whole numbers are used e.g. number of mosquitoes KD/dead.
If more than 4 decimal places are needed use exponential numbers, e.g. 0.00003 = 3x10^{-5}, and 3,000,000 = 3x10^{6}.</td></tr>
<tr><td>Rounding</td><td>Unless otherwise stated in the SOP, round down numbers ending 1-4 and round up numbers ending 5-9.</td></tr>
<tr><td>Date Format</td><td>The date format used in documentation is DD/MMM/YY:
<table>
<tr><td>DD</td><td>Day of the Month (1-31)</td></tr>
<tr><td>MMM</td><td>First three letters of the month of Jan, Feb, Mar, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec</td></tr>
<tr><td>YY</td><td>Last two digits of the year, 2016 record 16, 2017 record 17</td></tr>
</table></td></tr>
<tr><td>Time Format</td><td>Record time in 24-hour format (00.00 – 23.59). Example: 1pm = 13.00 Record a period of time in hours and minutes. Example: 1 hr 36 min = 1:36 hours.</td></tr>
<tr><td>Attachments to forms (refer also to page numbering)</td><td>Attach one document to another (for example if a protocol is to be attached to a SOP worksheet) by stapling the attachment to the record. Paperclips are not acceptable, as these can become loose and detach. Always cross-reference the document (protocol) number to the worksheet.
Supporting data attached to Protocol, Reports, Deviations or Process changes: Every page or “bundle” (A ‘bundle’ refer to the number of pages (i.e. a single sheet of paper = 1 page) securely bound together) attached, must clearly cross-reference the protocol/ report/ unique identification number, attachment number (If more than one document), and signature and date. For a ‘bundle’, the front page only may be marked up will be cross-referenced to the document and signed and dated attachment number, number of pages i.e. 1 of 2, 2 of 2 etc. initial and date. i.e. attaching client specific emails to a study protocol, adding additional supporting data to technical reports.</td></tr>
<tr><td>Raw data print-outs</td><td>Any raw data printouts generated during a test or task should be signed and dated, reference the associated worksheet No. and be attached to the associated worksheet by staple(s).</td></tr>
<tr><td>Thermal print-outs i.e. Balance printouts</td><td>All printouts made on thermal paper must be copied before attaching to a document or filing, since over time the print can fade. Write ‘copy’ on the copy and ‘Master’ on the original record initial and date. After making a copy, secure the ‘Master’ (by a staple) with the copy to the report/ worksheet</td></tr>
</table>

Filing or Archiving Raw Data

If data is recorded on paper, it should be filed into a folder for the project. First, ensure that a folder for the project has been set up. Second, the folder should contain a cover sheet and a contents sheet. A template is available in Appendix 1 – Cover sheet and Contents sheet. Third, a copy of the protocol should also be present, following the contents sheet. Finally, raw data should be in date order by oldest first. If raw data relates to two projects, for example mosquito or sample characterisation, the original should be copied, and on the copy, state it is a copy, with the data and initials of who photocopied and where the original is filed. Once a project is completed, the data should be filed away in a locked fireproof cabinet.

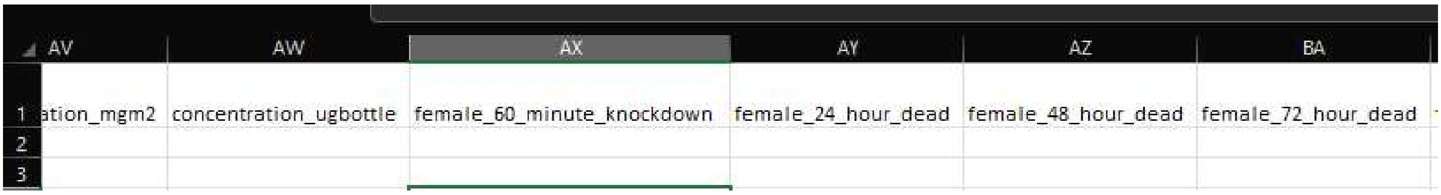
Electronic Data

Spreadsheet

A copy of the unprocessed raw data should be kept separately to the processed version of the data to avoid accidently manipulating data. Values should be reported to two decimal places consistently. Spreadsheets should contain three tabs: A data tab for your data, a data dictionary (called 'readme' here), and a meta data tab.

Data Tab

When creating a spreadsheet to record raw data, the data should be saved as an excel file (.xlsx). If a new template is needed, the author must aim to capture all variables that might impact the results of the bioassay. It might be possible to copy a template for a similar project and adjust the original template to suit your needs. Data should be recorded in a tidyverse style[3]. In the tidyverse style, data is recorded in a rectangle shape. Here each column represents a variable with the values of the variable inside that column. Data should be entered consistently e.g., in Figure 1 the "temperature" column contains only numbers to two decimal places. Where symbols such as percentage is needed, it should not be entered in the rows or column headers as different software may load symbols differently. The column header should read "humidity_percentage" or "humiditypercentange". Spaces in the column name should be avoided and column headers should not begin with numbers. To avoid blank empty spaces in the worksheet, "NA" should be entered where this applies. Numbers should be recorded to two decimal places to avoid ambiguity on the number values. Text should be entered consistently as lower case. In the notes tab, record any observations on the bioassay performed to ensure that changes to the SOP have been followed.

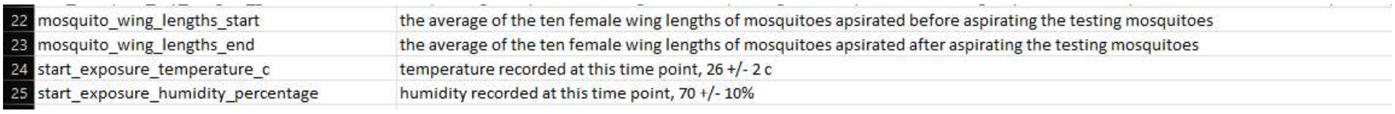


	AV	AW	AX	AY	AZ	BA
1	ation_mgm2	concentration_ugbottle	female_60_minute_knockdown	female_24_hour_dead	female_48_hour_dead	female_72_hour_dead
2						
3						

Figure 1 Tidy data. Here each column contains data relating to a single item and not multiple items. There are no empty blank cells, and the data is structured into a format that is easily readable by most statistical programmes.

Read Me Tab

To create a read me tab, the fastest method is to select the column headers on the data tab and paste them into a new tab. Rename the tab to either Read Me or Data Dictionary. Highlight the column headers then select the cell underneath the first header and select the paste transpose option. This will transpose the column headers into rows. In the cells next to the column headers, define the variable and provide an example of values ranges the audience may find. For columns where the results of the bioassay are being scored such as knockdown and mortality, it will not contain expected values. An example of the read me tab is found in Figure 2.

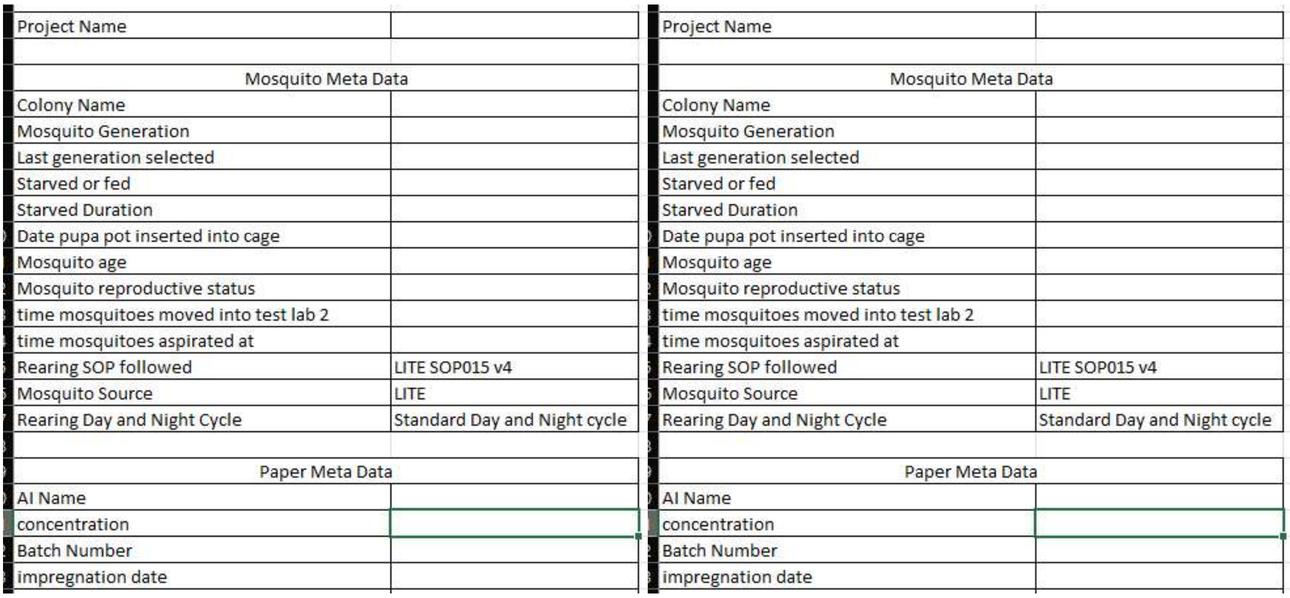


22	mosquito_wing_lengths_start	the average of the ten female wing lengths of mosquitoes apsirated before aspirating the testing mosquitoes
23	mosquito_wing_lengths_end	the average of the ten female wing lengths of mosquitoes apsirated after aspirating the testing mosquitoes
24	start_exposure_temperature_c	temperature recorded at this time point, 26 +/- 2 c
25	start_exposure_humidity_percentage	humidity recorded at this time point, 70 +/- 10%

Figure 2 example of what the read me tab contains. Each column has been defined to clearly state what it contains and where expected values can be given, the spreadsheet provides examples.

Meta Data Tab

The meta data tab should contain information on the SOPs followed, rearing procedures, information about the sample being tested, for example characterisation of the net or filter paper, and contain all information on the insecticides being used. For example, the meta data should clearly state where insecticide treated nets were acquired from, where and what the storage conditions were plus information of acclimatisation to testing environment conditions. Paper meta data should provide information on whether the papers were acquired from WHO or coated in the lab by an operator, coating date, concentration, and storage conditions. Rearing procedures of the mosquitoes should provide information on mosquitoes having access to sugar or being starved prior to the bioassay.



Project Name	
Mosquito Meta Data	
Colony Name	
Mosquito Generation	
Last generation selected	
Starved or fed	
Starved Duration	
Date pupa pot inserted into cage	
Mosquito age	
Mosquito reproductive status	
time mosquitoes moved into test lab 2	
time mosquitoes aspirated at	
Rearing SOP followed	LITE SOP015 v4
Mosquito Source	LITE
Rearing Day and Night Cycle	Standard Day and Night cycle
Paper Meta Data	
AI Name	
concentration	
Batch Number	
impregnation date	

Project Name	
Mosquito Meta Data	
Colony Name	
Mosquito Generation	
Last generation selected	
Starved or fed	
Starved Duration	
Date pupa pot inserted into cage	
Mosquito age	
Mosquito reproductive status	
time mosquitoes moved into test lab 2	
time mosquitoes aspirated at	
Rearing SOP followed	LITE SOP015 v4
Mosquito Source	LITE
Rearing Day and Night Cycle	Standard Day and Night cycle
Paper Meta Data	
AI Name	
concentration	
Batch Number	
impregnation date	

Figure 3 Example of the meta data tab containing a table on the rearing conditions. A full meta data table is available under Appendix 2 – Example of Meta Data Tab.

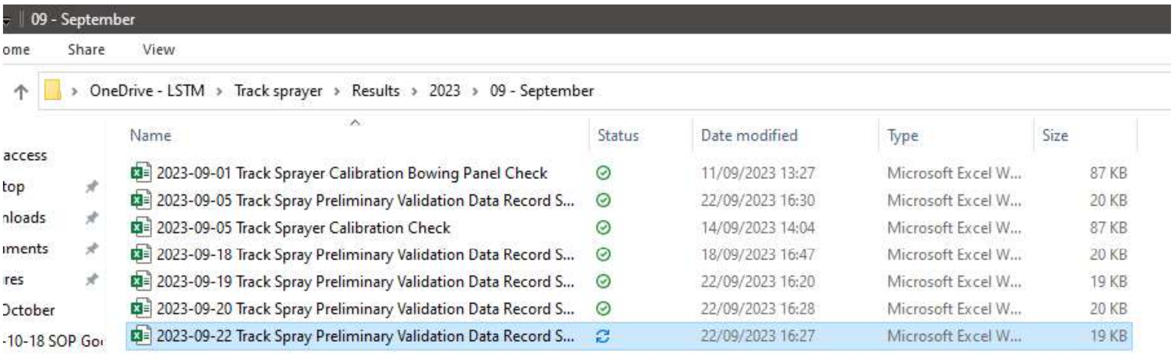
Organising Files

The general principle to saving your data is to be consistent and use a system that allows your data to be easily located. When saving files relating to a project, a specific folder for that project should be made. Within the directory, it should include subdirectories for anything that may be needed, e.g. SOPs, Risk Assessments, and Raw Data. Raw Data should contain subdirectories for each month a project ran, e.g. Track Sprayer/Raw Data/2023/08 - August. Raw Data folder may also require splitting into bioassay types if the project includes multiple bioassays or multiple colonies are tested.

File names for individual spreadsheets should also be consistent and contain key words of what the file contains. File names should not contain notes on the bioassay, that information should be recorded in the actual file. An example way of saving files is below:

YYYY-MM-DD_projectname_colonyname_bioassayname_repnumber

An example can be found in Figure 4 of good file name.



09 - September

ome Share View

OneDrive - LSTM › Track sprayer › Results › 2023 › 09 - September

access
top
nloads
iments
res
Dctober
-10-18 SOP Go

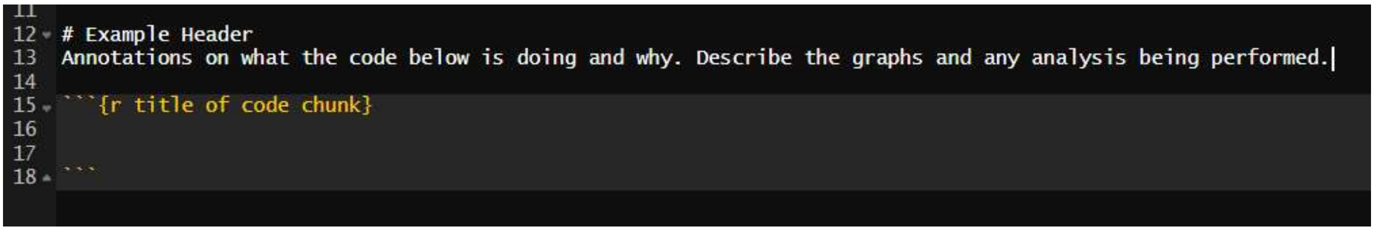
Name	Status	Date modified	Type	Size
2023-09-01 Track Sprayer Calibration Bowing Panel Check	✓	11/09/2023 13:27	Microsoft Excel W...	87 KB
2023-09-05 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:30	Microsoft Excel W...	20 KB
2023-09-05 Track Sprayer Calibration Check	✓	14/09/2023 14:04	Microsoft Excel W...	87 KB
2023-09-18 Track Spray Preliminary Validation Data Record S...	✓	18/09/2023 16:47	Microsoft Excel W...	20 KB
2023-09-19 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:20	Microsoft Excel W...	19 KB
2023-09-20 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:28	Microsoft Excel W...	20 KB
2023-09-22 Track Spray Preliminary Validation Data Record S...	⟳	22/09/2023 16:27	Microsoft Excel W...	19 KB

Figure 4 Saving data in consistent ordered way. The data is automatically saved in date order and any searches for track sprayer will return the data. The data is saved in a results folder and is saved in a folder for that month.

Data Analysis

Annotating, formulas, and software

Annotations may also describe the code, but it is more important to detail the data analysis. Annotations should describe what you are doing. Figure 6 provides an example of good annotation. An equation is being used to calculate value the deposit of an insecticide sprayed, therefore the annotation describes the equation used and all its components. Annotations should explain why the format of a variable has been changed e.g. numeric to a factor. Annotations should also draw attention to key outputs and explain their findings. There are different types of programmes that allow you to annotate code, here are we using R and RMarkdown. There are too many aspects of RMarkdown to include in a single SOP, but generally RMarkdowns allows an author to write text and group code in chunks. To create an R Markdown, open RStudio, select File, and New File, RMarkdown. Enter the details in the pop up and you can begin to start your code and annotations. Any text outside of the code chunk is automatically formatted as markdown text and code chunks are denoted at the start with ``` and ended with ```. Headers to create sections to describe the code are started with a # and text can be placed below.



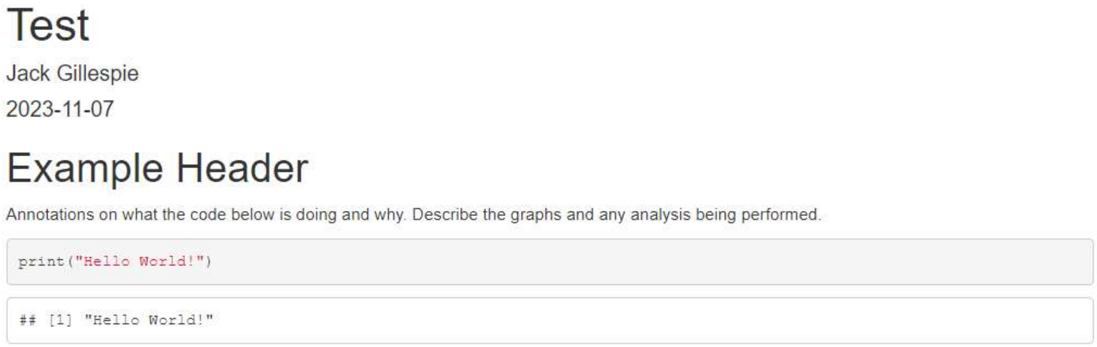
~~~
# Example Header
Annotations on what the code below is doing and why. Describe the graphs and any analysis being performed.

```{r title of code chunk}


```
~~~

Figure 5 - Example of how R Markdown is formatted. By creating headers and titles for your code chunks, it allows for easier navigation of the document.

Once complete RMarkdown can be outputted as several different formats, a HTML, Word Document or PDF etc. HTML is most versatile and be opened by almost any computer with an internet browser. Figure 6 displays an example of the output based on the earlier example.



Test

Jack Gillespie

2023-11-07

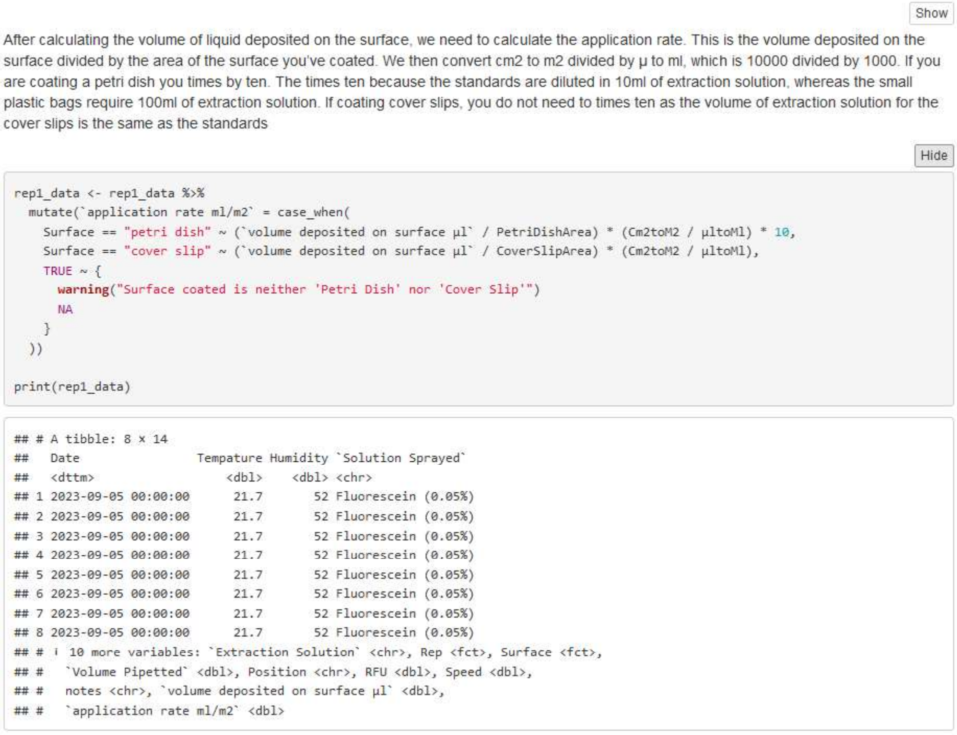
Example Header

Annotations on what the code below is doing and why. Describe the graphs and any analysis being performed.

```
print("Hello World!")
```

```
## [1] "Hello World!"
```

Figure 6 HTML output of the RMarkdown chunk above. The lowest box is the result of the code chunk.



After calculating the volume of liquid deposited on the surface, we need to calculate the application rate. This is the volume deposited on the surface divided by the area of the surface you've coated. We then convert cm2 to m2 divided by μ to ml, which is 10000 divided by 1000. If you are coating a petri dish you times by ten. The times ten because the standards are diluted in 10ml of extraction solution, whereas the small plastic bags require 100ml of extraction solution. If coating cover slips, you do not need to times ten as the volume of extraction solution for the cover slips is the same as the standards

```r
rep1_data <- rep1_data %>%
  mutate(`application rate ml/m2` = case_when(
    Surface == "petri dish" ~ (`volume deposited on surface µl` / PetriDishArea) * (Cm2toM2 / µltoMl) * 10,
    Surface == "cover slip" ~ (`volume deposited on surface µl` / CoverSlipArea) * (Cm2toM2 / µltoMl),
    TRUE ~ {
      warning("Surface coated is neither 'Petri Dish' nor 'Cover Slip'")
      NA
    }
  ))

print(rep1_data)
```

```
## # A tibble: 8 × 14
##   Date                Tempature Humidity `Solution Sprayed`
##   <dttm>                  <dbl>    <dbl> <chr>
## 1 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 2 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 3 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 4 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 5 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 6 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 7 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## 8 2023-09-05 00:00:00      21.7       52 Fluorescein (0.05%)
## # ℹ 10 more variables: `Extraction Solution` <chr>, Rep <fct>, Surface <fct>,
## #   `Volume Pipetted` <dbl>, Position <chr>, RFU <dbl>, Speed <dbl>,
## #   notes <chr>, `volume deposited on surface µl` <dbl>,
## #   `application rate ml/m2` <dbl>
```

Figure 7 HTML output of the RMarkdown. The text above the boxes describes the formula used to process the data and explain the calculations performed. This output allows the author to either show or hide code; here sections of the code are displayed in a grey shaded box and the output of the code is blow. This improves readability for the audience to view the code if they are interested. RMarkdown allows authors to create reports and produce presentations which people may find easier to use when writing up stats.

Data Visualisation

There are different methods to use to plot your graphs. GGplot2 is the most common method used to create plots in R, and similar packages are available for other coding languages[4]. GGplot2 others allow the creator to customise every aspect of their plot, allowing for greater control on how to create graphs than in software such as Excel.

Colours should be used sparingly to highlight different groups or a range of values. Graphs must be colour blind friendly. Deuteronopia, the inability to distinguish red and green, is the most common colour blindness, affecting 1 in 20 men and 3 in 1000 women[5]. The colour combinations of red and green are to be avoided. Appendix 5 - Colour Blindness Simulation contains a link to a website where you can upload your graphs and view them from the point of view of someone who is colour blind.

Conclusion

To summarise, being consistent and thinking about who reads you documents and uses your data is key to making your data accessible to everyone. This enables more transparent and efficient research. Good data practice is an ongoing process, and it is likely that as you work on more projects, you come across different ideas on how to improve your data management.

References

1. Peng, R. D. Reproducible Research in Computational Science. *Science* **334**, 1226–1227 (2011).
2. Broman, K. W. & Woo, K. H. Data Organization in Spreadsheets. *Am. Stat.* **72**, 2–10 (2018).
3. Wickham, H. Tidy Data. *J. Stat. Softw.* **59**, (2014).
4. Wickham, H. A Layered Grammar of Graphics. *J. Comput. Graph. Stat.* **19**, 3–28 (2009).
5. Neitz, M. & Neitz, J. Molecular Genetics of Color Vision and Color Vision Defects. *Arch. Ophthalmol.* **118**, 691–700 (2000).

Appendix

Appendix 1 – Cover sheet and Contents sheet

Project Name

Date Started:

Date Completed:

INNOVATION TO IMPACT

Bioassay Information	Filed Date	Filed By	Date saved electronically	Saved electronically by

Appendix 2 – Example of Meta Data Tab - fictional example data

Project Name	Insecticide Treated Nets Test

Mosquito Meta Data	
Colony Name	Tiassale
Mosquito Generation	152
Last generation selected	150
Starved or fed	Fed, 10% sugar solution
Starved Duration	NA
Date pupa pot inserted into cage	11 Nov 23
Mosquito age	4 days
Mosquito reproductive status	Mated
Time mosquitoes moved into test lab 2	11:30
Time mosquitoes aspirated at	11:00
Rearing SOP followed	LITE SOP015 v4
Mosquito Source	LITE
Rearing Day and Night Cycle	Standard Day and Night cycle

Paper Meta Data	
AI Name	Permethrin
Concentration	0.75%
Batch Number	1589
Impregnation date	Jun 23
Expiry date	Jun 24
AI Name	PY Control
Concentration	Negative Control – Silicone Oil and Acetone
Batch Number	1872
Impregnation date	Jun 23
Expiry date	Jun 24
AI Name	NA
Concentration	NA
Batch Number	NA
Impregnation date	NA
Expiry date	NA
Paper sources	Coated by JAG
Time papers moved into test lab 2	10:00
Paper storage conditions between bioassays	Papers stored in the fridge at 4°c

ITN Meta Data	
ITN Brand	Name of ITN
ITN source	Manufacturer
ITN batch number	1853
ITN manufacture date	Feb 22
ITN expiry date	NA
ITN condition (visible damage)	Undamaged
ITN time in field	NA – direct from manufacturer
Active ingredient #1	Permethrin
Active ingredient #1 concentration	20 grams of AI per kilogram of net ± 20%
Active ingredient #2	NA – Single AI net
Active ingredient #2 concentration	NA
Negative control net	Untreated net, sourced from Manufacturer
Positive control or comparator net	NA
ITN storage conditions	Stored in chemical cupboard CC3

Bottle Meta Data	
Bottles - date coated	1 Apr 23
Bottles - time coated	16:00
Active ingredient #1	Permethrin
Active ingredient #1 concentration	10µg per bottle
AI purity	97.8
AI batch	AB789
Active ingredient #2	NA
Active ingredient #2 concentration	NA
AI purity	NA
AI batch	NA
Concentration of MERO	NA
Time Bottles moved into test lab 2	9:00 2 Apr 23

Appendix 2 – Example template to record data – fictional example data

Non GLP WHO Tube Bioassay Record Sheet

Test Date:	5 Aug 23	Colony (Species and strain):	*Anopheles gambiae* Kisumu		
Test performed by	JAG	Total Wet weight of 20 female mosquitoes:	0.258 g	Within range Y/N	Y
Mosquitoes aspirated by (initials):	KG	Total Dry weight of 20 female mosquitoes:	0.025 g	Within range Y/N	Y
Mosquito Age:	3 to 4 days old	Time aspirated into loading tubes:	11:00	Removed within 1 hr Y/N	Y
PBO Batch Number	NA	Control Batch Number	754		
Permethrin Batch Number	154				

Lab set temperature at 26ºC +/- 2ºC and Humidity at 70% +/- 10%.

Exposure start		Exposure end		24 hour	
Temp (ºC)	Humidity (%RH)	Temp (ºC)	Humidity (%RH)	Temp (ºC)	Humidity (%RH)
27.1	78.2	27.1	77.9	26.9	75.8

Tube ID	Insecticide	Concentration (%)	Date/Time Papers Coated	Date/Time papers to dry	Number of times papers used	Date papers stored in Fridge	Time papers moved to testing temp	Mosquitoes to testing temp	Exposure start time	Exposure end time	Time papers out of fridge	Time papers returned to fridge	KD	Dead 24 hours	Total Mosquito Number
1	PY Control	Negative	5 Jul 23 16:00	5 Jul 23 17:00	0	5 Jul 23	10:00	11:30	12:30	13:30	1hr	15:00	0	1	25
2	Permethrin	0.75	5 Jul 23 16:30	5 Jul 23 17:30	0	5 Jul 23	10:00	11:30	12:30	13:30	1hr	15:00	25	25	25
3															
4	NA space not required	JAG 5 Aug 23													
5															
6															
7															
8															
9															
10															

11															
12															
13	NA space not required	JAG 5 Aug 23													
14															
15															
16															
17															
18															
19															

<table>
<tr><td rowspan="2">Scoring Details</td><td colspan="2">Date</td><td>Time</td><td>Date</td><td>Time</td><td>Date</td><td>Time</td></tr>
<tr><td colspan="2">5 Aug 23</td><td>14:30</td><td>6 Aug 23</td><td>14:30</td><td>6 Aug 23</td><td>15:30</td></tr>
<tr><td>Signature</td><td colspan="3">JAG</td><td colspan="2">JAG</td><td colspan="2">JAG</td></tr>
<tr><td>Notes</td><td colspan="7">NA</td></tr>
<tr><td>Paperwork checked by (initials):</td><td colspan="4">KG</td><td colspan="2">Date:</td><td>7 Aug 23</td></tr>
<tr><td>Raw data uploaded by (initials):</td><td>JAG</td><td>Date:</td><td>7 Aug 23</td><td>Raw data upload checked by (initials):</td><td>KG</td><td>Date:</td><td>7 Aug 23</td></tr>
</table>

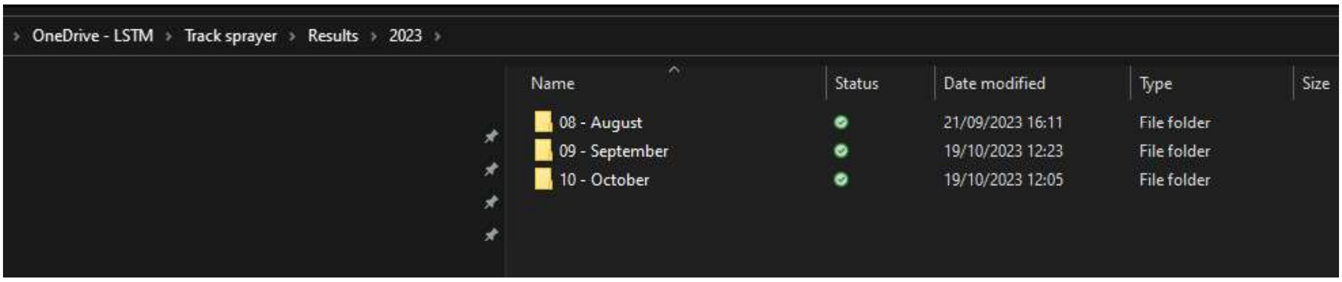
Project Name	

Mosquito Meta Data	
Colony Name	Kisumu
Mosquito Generation	NA
Last generation selected	NA – Susceptible strain – profiled Jan 23
Starved or fed	Fed, 10% sugar solution
Starved Duration	NA
Date pupa pot inserted into cage	2 Aug 23
Mosquito age	3 – 4 days old
Mosquito reproductive status	Mated
Time mosquitoes moved into test lab 2	11:25
Time mosquitoes aspirated at	11:00
Rearing SOP followed	LITE SOP015 v4
Mosquito Source	LITE
Rearing Day and Night Cycle	Standard Day and Night cycle

Paper Meta Data	
AI Name	Permethrin
Concentration	0.75%
Batch Number	1589
Impregnation date	Jun 23
Expiry date	Jun 24
AI Name	PY Control
Concentration	Negative Control – Silicone Oil and Acetone
Batch Number	1872
Impregnation date	Jun 23
Expiry date	Jun 24
AI Name	NA
Concentration	NA
Batch Number	NA
Impregnation date	NA
Expiry date	NA
Paper sources	Coated by JAG
Time papers moved into test lab 2	10:00
Paper storage conditions between bioassays	Papers stored in the fridge at 4°c

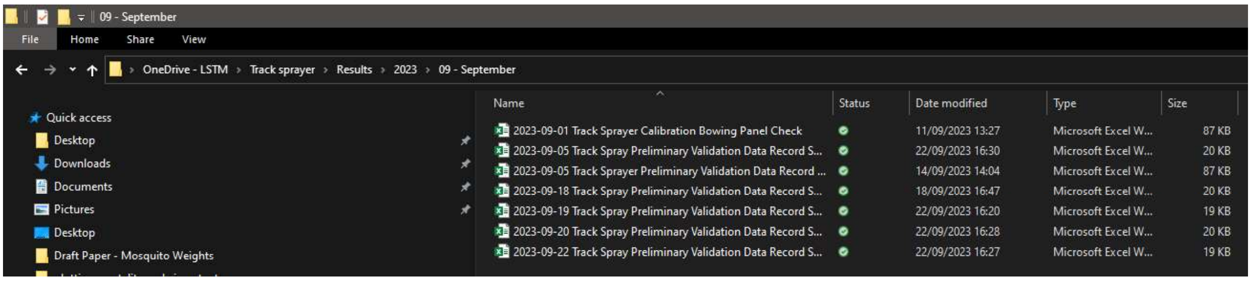
Appendix 4 – Good File Management

Below are screenshots of files stored in a method that allows the users to easily locate and search for files.



OneDrive - LSTM › Track sprayer › Results › 2023 ›

Name	Status	Date modified	Type	Size
08 - August	✓	21/09/2023 16:11	File folder	
09 - September	✓	19/10/2023 12:23	File folder	
10 - October	✓	19/10/2023 12:05	File folder	



OneDrive - LSTM › Track sprayer › Results › 2023 › 09 - September

Name	Status	Date modified	Type	Size
2023-09-01 Track Sprayer Calibration Bowing Panel Check	✓	11/09/2023 13:27	Microsoft Excel W...	87 KB
2023-09-05 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:30	Microsoft Excel W...	20 KB
2023-09-05 Track Sprayer Preliminary Validation Data Record ...	✓	14/09/2023 14:04	Microsoft Excel W...	87 KB
2023-09-18 Track Spray Preliminary Validation Data Record S...	✓	18/09/2023 16:47	Microsoft Excel W...	20 KB
2023-09-19 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:20	Microsoft Excel W...	19 KB
2023-09-20 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:28	Microsoft Excel W...	20 KB
2023-09-22 Track Spray Preliminary Validation Data Record S...	✓	22/09/2023 16:27	Microsoft Excel W...	19 KB

Appendix 5 - Colour Blindness Simulation

https://www.color-blindness.com/coblis-color-blindness-simulator/

Appendix 6 – Useful Links on R

Below is a link to the book R for Data Science by Hadley Wickham. This comprehensive book covers all the basics on how to use R and is a good resource on how to use R.

https://r4ds.hadley.nz/

This link is a book ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen. It covers ggplot2 all aspects of making graphs

https://ggplot2-book.org/

This link is a series of lessons on how to use R Markdown produced by R Studio.

https://rmarkdown.rstudio.com/lesson-1.html