Wye Collaborative Monitoring Network_
Guidance on working with citizen science data

Information sheet 1:
Cleaning raw data downloads from Epicollect

This information sheet describes how to download and clean raw data from each of the four citizen science group's involved in the Wye Catchment Collaborative Monitoring Network in 2021, including Friends of the Upper Wye, Friends of the Lugg, Campaign to Protect Rural England and Wye Salmon Association.

The guidance presented here enables data users to:

(1) Convert raw data downloaded from each of the group's Epicollect accounts so that it is easily understandable and comparable with other citizen science groups across the network;

(2) Prepare raw data downloaded from each of the group's Epicollect accounts so that it is ready for interpretation and analysis.

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD
School of Earth and Environmental Sciences
Ysgol Gwyddorau'r Ddaear a'r Amgylchedd
UKRI Natural Environment Research Council

This document has been prepared by Cardiff University on behalf of the Wye Catchment Collaborative Monitoring Network

Friends of the Upper Wye (FOUW)

- This information sheet is based on raw data downloaded from FOUW's 2021 Epicollect form set up and project status of 'public' (December 2021). If you are reviewing this guidance on more recent data downloads, please beware that the information provided here may be out of date.
- Data can be viewed directly from https://five.epicollect.net/project/fouw. However, you may need to register as a user on Epicollect and be granted access by FOUW to download the raw data set. You can contact science@fouw.org.uk for more information on downloading FOUW's raw citizen science data.

Recommended steps to cleaning FOUW's raw data download from Epicollect

1. Download and save the raw data from Epicollect

- Go to 'view data' on FOUW's Epicollect project page
- Click on the 'Download' button in the options bar at the top of the page
- Select 'custom' timeframe and select the start and end dates for your data download
- Choose CSV as the format to download the data in
- Click on the green 'Download' button
- Open the CSV file and save it to your computer

2. Select and delete Epicollect's automatic metadata and other data not required for analysis contained within the following columns of the spreadsheet:

Column A: ec5_uuid - The unique identifier of each data record entry (row of data)

Column B: updated_at - Timestamp when the entry is uploaded to the server

Column C: created_at - Timestamp when the entry was created

Column D: title - Title of data entry, based on data existing in other columns

Column E: 1_Full_name - Citizen scientist's full name

DO NOT DELETE COLLUM F: 2_Sample_site_ID

Column G: lat_3_Sample_site_locati - Longitude of sampling site, based on GPS

Column H: long_3_Sample_site_locati - Longitude of sampling site, based on GPS

Column I: accuracy_3_Sample_site_locati - Accuracy of GPS location

Column J: UTM_Northing_3_Sample_site_locati -Northing of site location

Column K: UTM_Easting_3_Sample_site_locati - Easting of site location

Column L: UTM_Zone_3_Sample_site_locati - UTM Zone of site location

3. **Copy location data and Unique identifying code from FOUW's master site log to the correct site on the spreadsheet**

- Insert three new columns to the right of column A (now 2_Sample_site_ID)
- Label column B, 'Unique_site_ref', column C, 'Lattitde', and column D. 'Longitude'
- Click on cell A1 (now 2_Sample_site_ID) and select 'Sort A to Z' on the 'Sort & Filter' drop down menu – all data will be sorted by numeric value of the site
- Scroll down and highlight all entries with sample sites IDs >100, cut these entries and paste the at the bottom of the sheet (after site 99)
- Open FOUW's master site log
- Copy and paste the unique site ID, latitude and longitude of the first site from FOUW's master log into cell B2, C2 and DC (headed Unique_site_ref, Lattitude and Longitude) next to the correct sample site ID (in column A) of your new spreadsheet
- Highlight these three cells, then drag and drop data into all the cells in those columns with the same sample site ID, ensuring that the unique site reference stays the same
- Repeats the three steps above for all sites with latitude and longitude data found in the log
- Delete all data entries from your CSV file that are marked as red in the log – these are from training sites locations

! Alert the FOUW group admin if any sample site ID's found in your data download do not exist in the log or do not have a corresponding latitude and longitude data stored in the log. FOUW will need to update the log with the correct site details. **If this is not possible, you will need to delete these entries from your CSV file.**

! Alert the FOUW group admin if any entries have a blank cell in column A, headed 2_Sample_site_ID. FOUW group admin may be able to identify them from the raw data download using the name of the sampler. **If this is not possible, you will need to delete these entries from your CSV file.**

4. **Standardise headings across all columns in line with other citizen science groups in the network, by replacing existing headings with the following titles:**

- A1: '2_Sample_site_ID' to 'Sample_site_ID'
- E1: '5_Date_sample_taken to 'Sample_date'
- F1: '6_Time_sample_taken' to 'Sample_time'
- G1: '7_Turbidity_NTU__ple' to 'Turbidity_secchi_NTU'

- H1: ‘8_Electrical_conduct’ to ‘EC_probe_HM3_uS’
- I1: ‘9_Temperature_c’ to ‘Water_temp_probe_HM3_C’
- J1: ‘10_Nitrate_strips_be’ to ‘Nitrate_strips_hach_ppm’
- K1: ‘11_Phosphate_strips_’ to ‘Phosphate_strips_lamotte_ppb’
- L1: ‘12_Hanna_handheld_ph’ to ‘Phosphate_checker_hanna_ppm’
- M1: ‘13_Fixed_point_photo’ to ‘Fixed_point_photo_url’
- N1: ‘14_Any_comments_abou’ to ‘Sample_comments’
- O1: ‘16_Rainfall_in_the_l’ to ‘Rainfall_24hr_description’
- P1: ‘17_Flow_conditions’ to ‘Flow_ description’
- Q1: ‘18_Relative_water_le’ to ‘Water_level_relative_description’
- R1: ‘19_Fixed_point_water to ‘Water_level_fixed_meters’
- S1: ‘21_Can_you_see_an_al’ to ‘Algal_bloom_query’
- T1: ‘22_If_yes_how_big_is’ to ‘Algal_bloom_size’
- U1: ‘23_Can_you_see_any_o’ to ‘Pollution_signs_query’
- V1: ‘24_If_yes_please_des’ to ‘Pollution_signs_description’
- W1: ‘25_Please_take_a_pho’ to ‘Pollution_photo_url’

5. Remove false zeros (for the Hanna Phosphate Checker)

- Open FOUW’s master site log, freeze the column ‘Unique_site_ref’ and view columns ‘Hanna_checker’ and ‘ Hanna_checker_start_date’
- In your CSV file, go to column L (‘Phosphate_checker_hanna_ppm’) and delete all zeros in the column that are linked to sites marked as ‘No’ in the ‘Phosphate_checker’ column of the master site log
- For sites where a date is given in the ‘Phosphate_checker_start_date’ column of the master site log, delete all entries in your CSV file of zero that have been entered for that site *before* the start date given

6. Amend data to the correct instrument range (for the Secchi tube)

- Select ‘Filter’ on the ‘Sort & Filter’ drop down menu
- Click on the drop down on cell G1 (now Turbidity_secchi_NTU) and select to view only those values less than 12, click OK
- Replace all values that appear (less than 12) with the value ‘12’
- Unselect ‘Filter’ on the ‘Sort & Filter’ drop down menu to turn the filter off

Friends of the Lugg (FOL)

- This information sheet is based on raw data downloaded from FOL's 2021 Epicollect form set up and project status of 'public' (December 2021). If you are reviewing this guidance on more recent data downloads, please beware that the information provided here may be out of date.
- Data can be viewed directly from https://five.epicollect.net/project/fol. However, you may need to register as a user on Epicollect and be granted access by FOL to download the raw data set. You can contact Christine Hugh-Jones (christinehughjones@gmail.com) for more information on downloading FOL's raw citizen science data.

Recommended steps to cleaning FOL's raw data download from Epicollect

1. Download and save the raw data from Epicollect

- Go to 'view data' on FOL's Epicollect project page
- Click on the 'Download' button in the options bar at the top of the page
- Select 'custom' timeframe and select the start and end dates for your data download
- Choose CSV as the format to download the data in
- Click on the green 'Download' button
- Open the CSV file and save it to your computer

2. Select and delete Epicollect's automatic metadata and other data not required for analysis contained within the following columns of the spreadsheet:

Column A: ec5_uuid - The unique identifier of each data record entry (row of data)

Column B: updated_at - Timestamp when the entry is uploaded to the server

Column C: created_at - Timestamp when the entry was created

Column D: title - Title of data entry, based on data existing in other columns

DO NOT DELETE COLLUM E: 1_Sample_site_ID

Column F: 2_Full_name - Citizen scientist's full name

Column G: 3_Sample_site_ID - Repeat of sample site ID

Column H: lat_4_Site_location - Longitude of sampling site, based on phone's GPS

Column I: long_4_Site_location - Longitude of sampling site, based on phone's GPS

Column J: accuracy_4_Site_location - Accuracy of GPS location
Column K: UTM_Northing_4_Site_location -Northing of site location, based on GPS
Column L: UTM_Easting_4_Site_location - Easting of site location based on GPS
Column M: UTM_Zone_4_Site_location - UTM Zone of site location based on GPS

3. **Copy location data and Unique identifying code from FOL's master site log to the correct site on the spreadsheet**

- Insert three new columns to the right of column A (now 3_Sample_Site_ID)
- Label column B, 'Unique_site_ref', column C, 'Lattitde', and column D. 'Longitude'
- Click on cell A1 (3_Sample_Site_Ref) and select 'Sort A to Z' on the 'Sort & Filter' drop down menu – all data will be sorted by numeric value
- Open FOL's master site log
- Copy and paste the unique site ID, latitude and longitude of the first site from the master log into cell B2, C2 and DC (headed Unique_Site_Ref, Lattitude and Longitude) next to the correct sample site ID (in column A) of your new spreadsheet
- Highlight these three cells, then drag and drop data into all the cells in those columns with the same sample site ID, ensuring that the unique site reference stays the same
- Repeats the steps above for all sites with latitude and longitude data found in the log

! Alert the FOL group admin if any sample site ID's found in your data download do not exist in the log or do not have a corresponding latitude and longitude data stored in the log. FOL will need to update the log with the correct site details. **If this is not possible, you will need to delete these entries from your CSV file.**

! Alert the FOL group admin if any entries have a blank cell in column A, headed 3_Sample_Site_ID. FOL group admin may be able to identify them from the raw data download using the name of the sampler. **If this is not possible, you will need to delete these entries from your CSV file.**

4. **Standardise headings across all columns in line with other citizen science groups in the network, by replacing existing headings with the following titles:**

- A1: '3_Sample_site_ID' to 'Sample_site_ID'
- E1: '6_Date_sample_taken' to 'Sample_date'

- F1: ‘7_Time_sample_taken’ to ‘Sample_time’
- G1: ‘8_Turbidity_NTU’ to ‘Turbidity_secchi_NTU’
- H1: ‘9_Electrical_conduct’ to ‘EC_probe_HM3_uS’
- I1: ‘10_Temperature_c’ to ‘Water_temp_probe_HM3_C’
- J1: ‘11_Nitrate_test_stri’ to ‘Nitrate_strips_hach_ppm’
- K1: ‘12_Phosphate_test_st’ to ‘Phosphate_strips_lamotte_ppb’
- L1: ‘13_Have_you_used_a_H’ to ‘Phosphate_checker_hanna_query’
- M1: ‘14__Hanna_handheld_P’ to ‘Phosphate_checker_hanna_ppm’
- N1: ‘15_Fixed_point_photo’ to ‘Fixed_point_photo_url’
- O1: ‘16_Any_comments_abou’ to ‘Sample_comments’
- P1: ‘18_Rainfall_in_the_l’ to ‘Rainfall_24hr_description’
- Q1: ‘19_Flow_conditions’ to ‘Flow_ description’
- R1: ‘20_Relative_water_le’ to ‘Water_level_relative_description’
- S1: ‘21_Fixed_point_water’ to ‘Water_level_fixed_meters’
- T1: ‘23_Can_you_see_an_al’ to ‘Algal_bloom_query’
- U1: ‘24_If_YES_how_big_is’ to ‘Algal_bloom_size’
- V1: ‘25_Can_you_see_any_o’ to ‘Pollution_signs_query’
- W1: ‘26_If_YES_please_des’ to ‘Pollution_signs_description’
- X1: ‘27_Please_take_a_pho’ to ‘Pollution_photo_url’

5. Amend data to the correct instrument range (for the Secchi tube)

- Select ‘Filter’ on the ‘Sort & Filter’ drop down menu
- Click on the drop down on cell G1 (now Turbidity_secchi_NTU) and select to view only those values less than 12, click OK
- Replace all values that appear (less than 12) with the value ‘12’
- Click on the drop down on cell G1 (now Turbidity_secchi_NTU) and select to view only those values more than 240, click OK
- Replace all values that appear (more than 240) with the value ‘240’
- Unselect ‘Filter’ on the ‘Sort & Filter’ drop down menu to turn the filter off

Campaign to Protect Rural England Herefordshire (CPRE)

- This information sheet is based on raw data downloaded from CPRE's 2021 Epicollect form set up and project status of 'private' (December 2021). If you are reviewing this guidance on more recent data downloads, please beware that the information provided here may be out of date.
- Data can be viewed directly from https://five.epicollect.net/project/cpre-herefordshire. However, you may need to register as a user on Epicollect and be granted access by CPRE to download the raw data set. You can contact Andrew McRobb (andrew@pmlgb.com) for more information on downloading CPRE's raw citizen science data.

Recommended steps to cleaning CPRE's raw data download from Epicollect

1. **Download and save the raw data from Epicollect**

- Go to 'view data' on CPRE's Epicollect project page
- Click on the 'Download' button in the options bar at the top of the page
- Select 'custom' timeframe and select the start and end dates for your data download
- Choose CSV as the format to download the data in
- Click on the green 'Download' button
- Open the CSV file and save it to your computer

2. **Select and delete Epicollect's automatic metadata and other data not required for analysis contained within the following columns of the spreadsheet:**

Column A: ec5_uuid - The unique identifier of each data record entry (row of data)

Column B: updated_at - Timestamp when the entry is uploaded to the server

Column C: created_at - Timestamp when the entry was created

Column D: created_by - Email address of Epicollect user who created entry

Column E: title - Title of data entry, based on data existing in other columns

Column F: 1_Welcome_to_CPREs_w- Site safety check

Column G: 2_Full_name- Full name of citizen scientist

DO NOT DELETE COLLUM H: 3_Site_Ref

Column I: 4_Site_name – Site name as description

Column J: lat_4_Site_location - Longitude of sampling site, based on phone's GPS

Column K: long_4_Site_location - Longitude of sampling site, based on phone's GPS

Column L: accuracy_4_Site_location - Accuracy of GPS location

Column M: UTM_Northing_4_Site_location -Northing of site location, based on GPS
Column N: UTM_Easting_4_Site_location - Easting of site location based on GPS
Column O: UTM_Zone_4_Site_location - UTM Zone of site location based on GPS
Column V: 13_Continue_to_measu – blank

3. **Copy location data and Unique identifying code from CPRE’s master site log to the correct site on the spreadsheet**

- Insert three new columns to the right of column A (now 3_Site_Ref)
- Label column B, ‘Unique_site_ref’, column C, ‘Lattitde’, and column D. ‘Longitude’
- Click on cell A1 (3_Site_Ref) and select ‘Sort A to Z’ on the ‘Sort & Filter’ drop down menu – all data will be sorted by numeric value (CPRE-01 to CPRE-X)[1]
- Open CPRE’s master site log
- Copy and paste the unique site reference, latitude and longitude of the first site from the master log into cell B2, C2 and DC (also headed Unique_Site_Ref, Lattitude and Longitude) next to the correct sample site reference (in column A) of your new spreadsheet
- Highlight these three cells, then drag and copy data into all the cells in those columns with the same sample site ID, ensuring that the unique site reference stays the same[2]
- Repeat the steps above for all sites with latitude and longitude data found in the log
- Delete all data entries from your CSV file that are marked as red in the log – these are from training sites locations

! Alert the CPRE group admin if any sample site ID’s found in your data download do not exist in the log or do not have a corresponding latitude and longitude data stored in the log. CPRE will need to update the log with the correct site details. **If this is not possible, you will need to delete these entries from your CSV file.**

[1] *Tip: Where site references are out of numeric order, manually cut and paste the rows to the correct position in the spreadsheet to make the next steps easier*

[2]*Tip: So that site numbers do not autofill when dragging down multiple entries from the same site, copy and paste the three cells (*Unique_Site_Ref, Lattitude *and* Longitude*) into the first and second sample row for that site, then highlight all six cells before dragging and copying down to the remaining sample rows*

! Alert the CPRE group admin if any entries have a blank cell in column A, headed 3_Site_Ref. CPRE group admin may be able to identify them from the raw data download using the name of the sampler. **If this is not possible, you will need to delete these entries from your CSV file.**

4. **Reorder data columns**

- Select and cut column E (headed '6_Type_of_sampling_p') and insert the column after column G (headed '8_Time_of_sample')

5. **Standardise headings across all columns in line with other citizen science groups in the network, by replacing existing headings with the following titles:**

- A1: '3_Site_Ref' to 'Sample_site_ID'
- E1: '7_Date_of_sample' to 'Sample_date'
- F1: '8_Time_of_sample' to 'Sample_time'
- G1: '6_Type_of_sampling_p' to 'Sample_position'
- H1: '10_Phosphate_test_st' to 'Phosphate_strips_lamotte_ppb'
- I1: '11_Nitrate_test_stri' to 'Nitrate_strips_hach_ppm'
- J1: '12_Are_you_measuring' to 'Other_parameters_query'
- K1: '15_Hanna_Digital_Pho' to 'Phosphate_checker_hanna_ppm'
- L1: '16_Electrical_conduc' to 'EC_probe_HM3_uS'
- M1: '17_Temperature_c' to 'Water_temp_probe_HM3_C'
- N1: '18_Turbidity_NTU__En' to 'Turbidity_secchi_NTU'
- O1: '19_Waterlevel_measur' to 'Water_level_fixed_method'
- P1: '20_Waterlevel_m' to 'Water_level_fixed_meters'
- Q1: '21_Photo_of_Gaugeboa' to 'Gaugeboard_photo_url'
- R1: '23_Rainfall_in_last_' to 'Rainfall_24hr_description'
- S1: '24_Flow_conditions' to 'Flow_ description'
- T1: '25_Relative_water_le' to 'Water_level_relative_description'
- U1: '27_Can_you_see_an_al' to 'Algal_bloom_query'
- V1: '28_If_YES_how_big_is' to 'Algal_bloom_size'
- W1: '29_Can_you_see_any_o' to 'Pollution_signs_query'
- X1: '30_Describe_any_sign' to 'Pollution_signs_description'
- Y: '31_Please_take_a_photo' to 'Pollution_photo_url'
- Z: '32_Any_comments_abou' to 'Sample_comments'
- AA: '33_Take_a_photo_of_w' to 'Fixed_point_photo_url'

Wye Salmon Association (WSA)

- This information sheet is based on raw data downloaded from WSA's 2021 Epicollect form set up and project status of 'public' (December 2021). If you are reviewing this guidance on more recent data downloads, please beware that the information provided here may be out of date.
- Data can be viewed directly from https://five.epicollect.net/project/wye-water-quality-monitoringl. However, you may need to register as a user on Epicollect and be granted access by WSA to download the raw data set. You can contact Stuart Smith (stuartsmith@wyesalmon.com) for more information on downloading WSA's raw citizen science data.

<u>Recommended steps to cleaning WSA's raw data download from Epicollect</u>

1. Download and save the raw data from Epicollect

- Go to 'view data' on WSA's Epicollect project page
- Click on the 'Download' button in the options bar at the top of the page
- Select 'custom' timeframe and select the start and end dates for your data download
- Choose CSV as the format to download the data in
- Click on the green 'Download' button
- Open the CSV file and save it to your computer

2. Select and delete Epicollect's automatic metadata and other data not required for analysis contained within the following columns of the spreadsheet:

Column A: ec5_uuid - The unique identifier of each data record entry (row of data)
Column B: updated_at - Timestamp when the entry is uploaded to the server
Column C: created_at - Timestamp when the entry was created
Column D: title - Title of data entry, based on data existing in other columns

Do not delete columns E and F ('1_Date_of_Sample', '2_Time_Sample_was_ta')

Column G: 3_Data_Recorder_init - Citizen scientist's initials

Do not delete columns H, I and J ('4_County', '5_River', and '6_Location')

Column K: lat_4_Site_location - Longitude of sampling site, based on phone's GPS
Column L: long_4_Site_location - Longitude of sampling site, based on phone's GPS
Column M: accuracy_4_Site_location - Accuracy of GPS location

Column N: UTM_Northing_4_Site_location -Northing of site location, based on GPS
Column O: UTM_Easting_4_Site_location - Easting of site location based on GPS
Column P: UTM_Zone_4_Site_location - UTM Zone of site location based on GPS

3. Delete any unwanted entries from the wrong timeframe

- Select 'Filter' on the 'Sort & Filter' drop down menu
- Click on the dropdown arrow on cell A1 (now '1_Date_of_Sample') and select all dates that fall outside the range you are interested in
- Highlight and delete all rows that appear
- Unselect 'Filter' on the 'Sort & Filter' drop down menu and you will be left with all samples taken within the date range of interest

4. Rearrange the order of columns in the spreadsheet

- Cut the column labelled '6_Location' and insert cut cells into column A

5. Copy location data and Unique identifying code from FOL's master site log to the correct site on the spreadsheet

- Insert three new columns to the right of column A (now 6_Location)
- Label column B, 'Unique_site_ref', column C, 'Lattitde', and column D. 'Longitude'
- Click on cell A1 (6_Location) and select 'Sort A to Z' on the 'Sort & Filter' drop down menu
- Rename the first few sets of site names which appear with quotation marks around them, by removing the quotation marks and dragging and dropping the new site name
- Click on cell 1A (6_Location) again and select 'Sort A to Z' on the 'Sort & Filter' drop down menu a second time – all sites should now appear in alphabetical order and match those found in WSA's master site log
- Open WSA's master site log
- Copy and paste the unique site ID, latitude and longitude of the first site from the master log into cell B2, C2 and DC (headed Unique_Site_Ref, Lattitude and Longitude) next to the correct sample site name (in column A, headed "6_Location") of your new spreadsheet
- Highlight these three cells, then drag down and copy the data into all the cells in those columns with the same sample site name (in column A), ensuring that the unique site reference stays the same each time it is copied

- Repeats the steps above for all sites with latitude and longitude data found in the log
- Delete any entries where multiple site names appear in the same cell in column A of your CSV file

! Alert the WSA group admin if any sample site ID’s found in your data download do not exist in the log or do not have a corresponding latitude and longitude data stored in the log. WSA will need to update the log with the correct site details. **If this is not possible, you will need to delete these entries from your CSV file.**

! Alert the WSA group admin if any entries have a blank cell in column A, headed 6_Location. WSA group admin may be able to identify them from the raw data download using the name of the sampler. **If this is not possible, you will need to delete these entries from your CSV file.**

6. **Standardise headings across all columns in line with other citizen science groups in the network, by replacing existing headings with the following titles:**

- A1: ‘6_Location’ to ‘Sample_site_name’
- E1‘1_Date_of_Sample’ to ‘Sample_date’
- F1: ‘2_Time_Sample_was_ta’ to ‘Sample_time’
- G1: ‘4_County’ to ‘County’
- H1: ‘5_River’ to ‘River’
- I1: ‘8_Phosphate_PO4_mgl_’ to ‘Phosphate_checker_hanna_ppm’
- J1: ‘9_Nitrate_mgl_or_ppm’ to ‘Nitrate_strips_hach_ppm’
- K1: ‘10_TDS_mgl_or_ppm’ to ‘TDS_probe_HM3_ppm’
- L1: ‘11_pH_00’ to ‘pH_probe_HM80’
- M1: ‘12_Ammonical_Nitroge’ to ‘Ammonia_strips_hach_ppm’
- N1: ‘13_Water_Temp_C’ to ‘Water_temp_probe_HM3_C’
- O1: ‘14_Water_Colour’ to ‘Water_colour_description’
- P1: ‘15_NRWEA_Water_Gauge’ to ‘Gauge_board_location’
- Q1: ‘16_Water_Height_m_if’ to ‘Water_level_fixed_meters’
- R1: ‘17_Water_Level_if_no’ to ‘Water_level_relative_description_alt’
- S1: ‘18_ Weather’ to ‘Weather_condition’
- T1: ‘19_Livestock_Near_wa’ to ‘Livestock_query’
- U1: ‘20_Comments’ to ‘Comments_misc’
- V1: ‘21_Photo’ to ‘Photo_misc’