**LSHTM Covid-19 volunteer data cleaning**

This document is a step-by-step guide to the weekly processing required to enable making the data available to the volunteers. It uses a file supplied by the project’s data scientist of records from previous weeks plus new records from the providers.

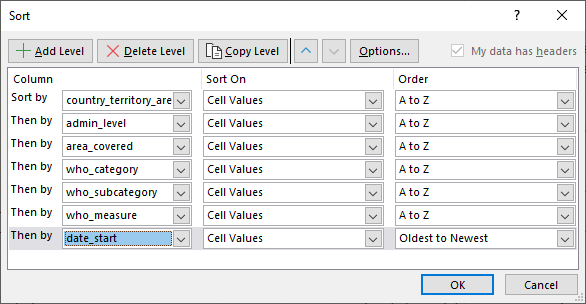
**Perform checks on the data**

1. Create a folder for this week’s data
2. Receive data file from data scientist (master\_yyyy\_mm\_dd.csv)
3. Save file in the folder
4. Rename file master\_yyyy\_mm\_dd.xlsm (it needs to be saved as a macro file for the formulae to work)
5. Rename the sheet to ‘data’
6. Do some high-level counts of the records:
   1. Open country\_list&stats.xlsx template
   2. Save in this week’s folder
   3. Update cell B1 on the country\_level\_stats sheet with this week’s master name (no extension needed)  
      This will auto-populate formulae on that sheet and the top\_level\_stats sheet
   4. Check that cells B3, B4, B5 on the top\_level\_stats sheet have populated correctly (count the rows for this week’s master):
      1. B3 = all records
      2. B4 = those which don’t require cleansing, i.e. where who\_id = 8.2, 8.3, 8.4, 8.5, 10, 11, 12, 13, 14, 8.4.1, 8.4.2 (n.b. 9 is not in this list)
      3. B5 = B3-B4 = the number of records the volunteers will clean
7. Perform data checks on master (any issues, report to Chris):

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DATA CHECKS** |  |  |
|  | **Field** | **Check** | **Reason/extra info** |
|  | who\_id | Check all populated (there are 3 records, two for San Marino, one for Bosnia which don’t have who\_id populated – these are historic so ignore them) |  |
|  | Keep | Check all records are populated with either y or n |  |
|  | keep / duplicate\_record\_id | If keep = ‘N’ and who\_code <> 10, 11, 12, 13, 14 raise this with Orlagh | This should have been updated in the post-volunteer processing. |
|  | who\_region | check values are only those permitted (no blanks) | Permitted values are detailed in the WHO\_taxonomy sheet, columns I:P (columns are hidden) |
|  | admin\_level | check values are only those permitted (no blanks) |  |
|  | date\_start | check all dates are plausible, i.e. 2020/2021 or end of 2019. Blanks are permitted (although they’re far from ideal) |  |
|  | date\_end | check all dates are plausible, i.e. 2020 or 2021. Blanks are permitted If there are text values, delete them |  |
|  | measure\_stage | check values are only those permitted (blanks permitted at this stage) |  |
|  | reason\_ended | check values are only those permitted (same as measure\_stage, plus a value of time-limited) |  |
|  | Enforcement | check values are only those permitted (blanks permitted at this stage) |  |
|  | non\_compliance\_penalty | check values are only those permitted (blanks permitted at this stage) |  |
|  | link\_live | check values are only those permitted (blanks permitted at this stage) |  |
|  | link\_eng | check values are only those permitted (blanks permitted at this stage) |  |

1. Copy column “who\_id” so it is after measure\_stage (column W) and rename it to   
   “who\_id (repeated from column B)”
2. Delete any blank rows below the data
3. Save file
4. Re-count records which don’t need cleansing (codes listed above) – this may have changed from above processing
   1. If the counts have changed, go the country\_list&stats spreadsheet, top\_level\_stats sheet and copy and paste values the updated values into C3:C5 (we want a count of the number of records immediately before the master is sub-divided)
5. Separate out records which don’t require cleansing into a different spreadsheet
   1. (codes listed above in 6.4.2)
   2. Copy these records to a separate file, named Extracted\_yyyymmdd\_8s10s11s12s13s and save in that week’s folder. This will be uploaded to Teams later on.
   3. Check the record count is correct against the count done earlier.
6. Create a new master spreadsheet
   1. Filter where who\_code is not in the code list (i.e. the list in 6.4.2)
   2. Copy records which we want to send to the volunteers (i.e. where who\_id is not in the list above) to a new blank file
   3. Save the master (i.e. the file we have just copied the records from) as master\_yyyymmdd\_stage1.xlsm
   4. Save the new document as master\_yyyymmdd.xlsm. Doing this means that the stats sheet will now use this document from which to pull its figures.
   5. Rename the sheet “data”
7. Format this new master file so it is suitable to send to the volunteers
   1. Hide all except the following columns:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| who\_id | Keep | duplicate\_record\_id | country\_territory\_area | admin\_level | area\_covered | prov\_category |
| prov\_subcategory | prov\_measure | who\_code | who\_category | who\_subcategory | who\_measure | Comments |
| date\_start | measure\_stage | who\_id (copied from column B) | prev\_measure\_number | following\_measure\_number | targeted | Enforcement |
| non\_compliance\_penalty | Link | link\_live | link\_eng | queries\_comments |  |  |

* + 1. Check that there are 26 columns visible
  1. Add the following sort order:
  2. Add vlookups for who\_category, who\_subcategory, who\_measure (note $Q2 was previously $P2).  
     =VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 2, FALSE)  
     =IF(VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 3, FALSE)=0, "", VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 3, FALSE))  
     =IF(VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 4, FALSE)=0, "", VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 4, FALSE))
     1. Check vlookups work – inc who\_code = 14
  3. Check keep = Y for all records
  4. Check there are no values in Duplicate\_record\_id
  5. Set date\_start to dd-mmm-yy format (this format is more likely to highlight records where the month and day have been transposed)
  6. Add the following conditional formatting is working:
     1. Strikethrough records what we’re not keeping  
        Where keep = N, strikethrough the row – amend the ‘Applies to’ box to $A:$AQ  
        =$E2 = "N"
     2. Who\_id is not permitted (list below)

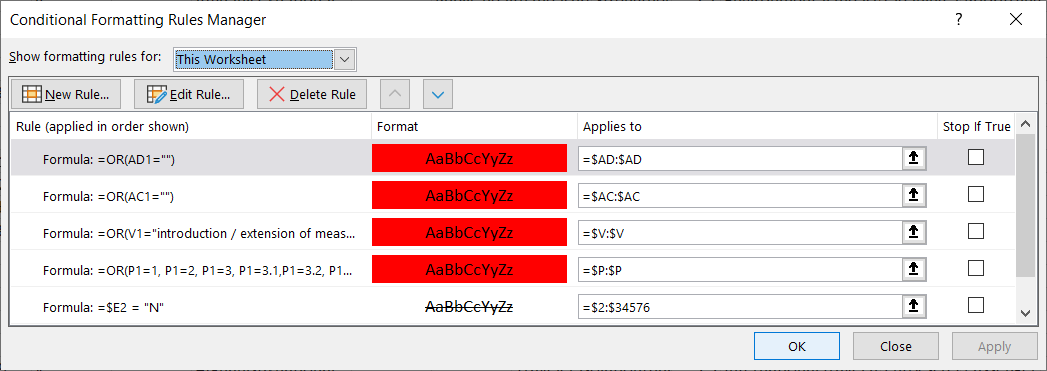
Highlight column P, select “Use a formula to decide which cells to format” and paste the following  
=OR(P1=1, P1=2, P1=3, P1=3.1,P1=3.2, P1=4, P1=4.1, P1=4.2, P1=4.3, P1=4.4, P1=4.5, P1=5, P1=6, P1=7, P1=8, P1=8.4, P1=9)

Set fill to red

* + 1. Where measure\_stage = “Introduction / extension of measures" or "Lift" or blank)  
       Measure\_stage – highlight column V, select “Use a formula to decide which cells to format” and paste the following   
       = OR(V1="introduction / extension of measures", V1 ="Lift", V1 = "")  
       set cell to red fill
    2. Where enforcement is blank

=OR(AC1="")  
set cell to red fill

* + 1. Where non\_compliance\_penalty is blank  
       =OR(AD1="")  
       set cell to red fill



* 1. Colour headings grey
  2. Make heading row double height, left and top aligned, wrap text
  3. Re-size columns
  4. Put cursor in cell H2 and freeze panes
  5. Set row height = 20
  6. Delete values in queries column

1. Run macros to create the country files:
   1. Copy macros SplitToFiles and CopyTaxonomyAddDropdownsColourRows from last week’s master into this week’s master
   2. Close all applications and restart computer
   3. Re-open this week’s master\_yyyymmdd.xlsm
   4. Check the correct sort order has been retained (this sounds crazy, but sometimes it isn’t retained)
   5. Run macro SplitToFiles
      1. In the dialogue "Enter the column number used for splitting", type 9 (i.e. country\_territory\_area)
      2. In the dialogue "Enter the starting row number (to skip header)", type 2 (i.e. to ignore headings)
   6. When that has successfully finished, run CopyTaxonomyAddDropdownsColourRows
   7. Check the following have all worked:
      1. does the sort order work
      2. are the vlookups for columns Q, R, S referencing the taxonomy correctly
      3. Change Keep = N, does the whole row ~~strikethrough~~
      4. Check the conditional formatting is correct:
         1. If there are any non-permitted who\_codes, check that they have a red fill
         2. Measure\_stage
         3. enforcement
         4. Non\_compliance\_penalty
      5. are all of the column headings correct
   8. When you’re happy that the data looks right, select a file and send it to Chris (via Slack) for him to double-check it. Chose a file which has approximately 5-6 new records so that if there are any issues, they’re likely to be evident.

**Create sign-up sheets and upload the data into Teams**

1. Check the stats per country
   1. Open excel template Country\_list&stats
      1. Check that there aren’t new countries this week (check the number of countries listed in the country\_level\_stats sheet B4 matches cell C4)
      2. If there are any new countries, add them to the list
      3. Check that the number of records per country, totalled matches the overall number of records we’re expecting (cell C5 in the Top Level Stats sheet)
2. Separate the country files which don’t need cleansing into a new folder
   1. Copy the list of countries with no new records to the .csv tab. Column B will populate the list of countries with .xlsx file extension. Copy this column to a new spreadsheet and save it as countries.csv in the \split folder. Important: close the spreadsheet!
   2. Create a sub-folder in the \split folder called no\_cleansing\_needed
   3. Open a command prompt (from the Windows button, type Command prompt)
   4. Navigate to the current \split folder. To do this, type the following:   
      cd c:\Kate\lshtm-covid19\data\_cleanse\_20210421\split  
      where c:\etc is the path of the current split folder  
      press Enter to execute the command
   5. Copy the following text (changing the file path to your file path) and press enter  
      For /f "delims=" %N in (countries.csv) do move "%N" C:\Kate\LSHTM-Covid19\Data\_cleanse\_20210421\Split\no\_cleansing\_needed
   6. This should move the files in the countries.csv file into the no\_cleansing\_needed subfolder. However, for an unknown reason it doesn’t move the first file, nor those with a comma in the country name, so these need to be moved manually   
      (currently Bolivia, Plurinational State Of; Bonaire, Sint Eustatius And Saba; Iran, Islamic Republic Of; Micronesia, Federated States Of; Northern Mariana Is lands, Commonwealth Of The; Venezuela, Bolivarian Republic Of).   
      Obviously, some of them will not be in the ‘no new records needing cleansing’ list in any one week.
   7. Check the correct files have been moved
      1. If you want to check which files are in any of the folders, go to a command prompt and type dir >list.csv  
         This will create a csv file in that folder, listing all of the files in that folder. (it can be called anything, and if you want a text document, type dir >list.txt).
3. Create the sign-up sheet
   1. Go into Teams>Coding Team>Files
   2. Download Countries\_requested.xls
   3. Save it as Sign-up\_sheet\_yyyymmdd
   4. Copy the total # records and # new records from the Country\_list&stats spreadsheet and paste values into Sign-up\_sheet\_yyyymmdd
      1. Check which countries have no new records and set the Name field = n/a and Status field = finished
      2. Whichever country was checked by Chris, assign that to him in the sheet and set Status = finished
      3. Check the totals at the bottom of the sheet are correct
      4. Hide columns F:I
4. Create the Admin sign-up sheet
   1. Save Sign-up\_sheet\_yyyymmdd as ‘Admin\_sign-up\_yyyymmdd’
   2. Un-hide columns F:I
   3. Filter on column F, Coding Team to remove entries in columns D and E where the person isn’t a member of the Coding Team.
   4. Re-hide columns F:I
5. Upload the country files which need cleaning (i.e. those in the Split folder) to Teams\Data to clean
   1. Upload the Sign-up\_sheet\_yyyymmdd to the same folder
6. Upload the country files which don’t need cleaning (i.e. those in the no\_cleansing\_needed subfolder) to ADMIN ONLY – cleaned and checked.
   1. Upload the Admin\_sign-up\_sheet\_yyyymmdd to the same folder
7. Check that all files have uploaded (Teams can be slow and temperamental)
8. Make files available for Orlagh for next week:
   1. In Teams>Project Team, create a new subfolder called Task xx PHSM weekly update xx master file etc (increment the numbers in the folder name from last week)
   2. Copy the following files into this folder  
      master\_yyyymmdd (the excel sheet which you run the macros from)  
      Extracted\_yyyymmdd excel spreadsheet  
      country\_list&stats
9. Let Chris know that the data is ready so he can put a message up on Teams.

**Macros**

1. **SplitToFiles**
   1. This macro splits a single excel file into multiple files, based on the specified column (i.e. column 8, country\_territory\_area). The data must be sorted by that column. N.B. it is case specific so Afghanistan and afghanistan are treated as distinct values.
   2. If the file has a header, it is copied to the start of each of the new files.
   3. The macro takes the location of the master file, and creates a sub-folder called Split. The new files are saved into this sub-folder.
2. **CopyTaxonomyAddDropdownsColourRows**
   * This macro processes each file in the Split folder in turn, and
   * Copies the WHO\_taxonomy sheet to the file
   * Creates data validation on specified columns
   * Creates conditional formatting for the new rows.
   1. **Data validation** (i.e. a drop-down list to ensure that the volunteers can only select approved values) is added for the following columns:
      1. Keep
      2. Admin\_level
      3. WHO\_code
      4. Measure\_stage
      5. Enforcement
      6. Non\_compliance\_penalty
      7. Link\_live
      8. Link\_eng
   2. These values are taken from the WHO\_taxonomy sheet, columns I:P.
   3. **Conditional formatting**
      1. Where processed = “not cleansed” or “cleansed”, the row fill is coloured yellow – this highlights the new records for the volunteers  
         (N.B. “cleansed” was originally used to for John Hopkins data – originally coloured blue - and “not cleansed” was for the other providers but we no longer need to differentiate between them).

Kate Nolan

21st December 2020

**Appendix**

1. Create a working folder for this week’s data, ideally named ‘CURRENT\_WEEK’.
2. Check that the following files are already in the folder in MS Teams Project Team/ Task xx PHSM weekly update yy. If not, then ask the Data Engineer to supply them:

Volunteer\_Rows\_ yyyy\_mm\_dd.xlsx

No\_New\_Records\_ yyyy\_mm\_dd.xlsx

Extracted\_code8\_ yyyy\_mm\_dd.xlsx

Extracted\_other\_ yyyy\_mm\_dd.xlsx

master\_yyyy\_mm\_dd.xlsm (note that this is *macro-enabled*)

country\_list&stats\_ yyyy\_mm\_dd.xlsm (note that this is *macro-enabled*)

update\_merge\_ yyyy\_mm\_dd.xlsx (only necessary if for whatever reason you intend to examine historical prop\_IDs)

1. Download the following files from the folder in MS Teams Project Team/ Task xx PHSM weekly update yy, and save them in your working folder:

No\_New\_Records\_ yyyy\_mm\_dd.xlsx

master\_yyyy\_mm\_dd.xlsm (note that this is *macro-enabled*)

country\_list&stats\_ yyyy\_mm\_dd.xlsm (note that this is *macro-enabled*)

Volunteer\_Rows\_ yyyy\_mm\_dd.xlsx

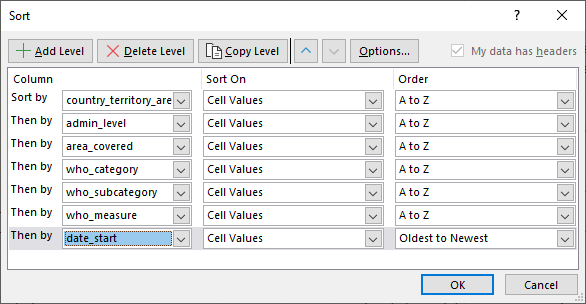
1. Do some high-level counts of the records:
   1. Open both country\_list&stats.xlsm *and* master\_yyyy\_mm\_dd.xlsm – for the latter, check that the tab is renamed **data**.
   2. Check that cell B1 on the country\_level\_stats sheet has the date corresponding to this week’s master  
      This will auto-populate formulae on that sheet and the top\_level\_stats sheet
   3. Check that cells B3, B4, B5 on the top\_level\_stats sheet have populated correctly (count the rows for this week’s master).

B3 = all records that are *not* in the two Extracted files. They include both the number of records in the Volunteer\_Rows\_ yyyy\_mm\_dd.xlsx file and also the No\_New\_Records\_ yyyy\_mm\_dd.xlsx. Note that B4 is no longer used and will show zero, and that consequently B5 = B3.

1. Open No\_New\_Records\_ yyyy\_mm\_dd.xlsx, and create a new (i.e. second) tab. Copy the list of Countries from column H of the first tab, and paste them into column A of the second tab. Remove duplicates, so that a unique list remains.
2. Now go back to country\_list&stats.xlsm and go to the first tab (country\_level\_stats). Click the filter in cell D9, and ensure that only 0 is selected. Select the corresponding list of Countries in column A, and copy it over into column G of the second tab in No\_New\_Records\_ yyyy\_mm\_dd.xlsx. Check that all the countries in both columns match, save, and upload No\_New\_Records\_ yyyy\_mm\_dd.xlsx to the folder in MS Teams Project Team/ Task xx PHSM weekly update yy, overwriting the version of the file that may already be present. Going back to the country\_level\_stats tab, release the applied filter in cell D9.
3. Download the pre-formatted *macro-enabled* template Master\_Volunteer\_Formatted\_Template.xlsm from the top of the ‘Project team’ folder on MS Teams, and add it to your working folder. Once it is in the working folder, rename it to the truncated form: Master\_Volunteer.xlsm (without the date). Open the file.
4. Now open the file Volunteer\_Rows\_ yyyy\_mm\_dd.xlsx, and in the date\_start column (U) check the number of future dates, and make a mental note of this value. Make sure that you release the filter when you have done this check. Copy across all the data (as **values**, i.e. without the formatting) from Volunteer\_Rows\_ yyyy\_mm\_dd.xlsx to Master\_Volunteer.xlsm.
5. In Master\_Volunteer.xlsm. set date\_start to dd-mmm-yy format (this format is more likely to highlight records where the month and day have been transposed), and again check the number of future dates in column U. Make sure that this matches the number in step (8), and remove the filter.
6. Hide all except the following columns:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| who\_id | Keep | duplicate\_record\_id | country\_territory\_area | admin\_level | area\_covered | prov\_category |
| prov\_subcategory | prov\_measure | who\_code | who\_category | who\_subcategory | who\_measure | Comments |
| date\_start | measure\_stage | who\_id (copied from column B) | prev\_measure\_number | following\_measure\_number | targeted | Enforcement |
| non\_compliance\_penalty | Link | link\_live | link\_eng | queries\_comments | trigger |  |

1. Check that there are 27 columns visible.
2. Add vlookups for who\_category, who\_subcategory, who\_measure  
   =VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 2, FALSE)  
   =IF(VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 3, FALSE)=0, "", VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 3, FALSE))  
   =IF(VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 4, FALSE)=0, "", VLOOKUP($Q2,WHO\_Taxonomy!$A$2:$D$71, 4, FALSE))
3. Check vlookups work – inc who\_code = 14. In practise, it is sufficient to check the first and last row.
4. Add the following sort order, by initially clicking on a single cell (and *not* the whole spreadsheet) before adding it:



1. Save the spreadsheet, and close it. **Very Important**: Close *all* other instances of Excel that are currently open, as other instances can seriously affect the performance of the Macros.
2. Reopen Master\_Volunteer.xlsm, and check that the Sort Order is still present. If not, then add it again and repeat both step (15) as well as this step.
3. Check the (active) end cell by putting the cursor initially in cell A2 and then press (Ctrl) and (End) simultaneously. Check that the cursor ends on the last cell of the bottom row in the table, rather than an empty row. If this condition is not satisfied, then run the Macro ‘ResetLastCell’.

1. Run macros to create the country files. After each Macro is completed, check the timestamps on each Country file to check that the Macro has been applied to it:
   1. Run macro SplitToFiles
      1. In the dialogue "Enter the column number used for splitting", type 8 (i.e. country\_territory\_area)
      2. In the dialogue "Enter the starting row number (to skip header)", type 2 (i.e. to ignore headings)
   2. When the macro SplitToFiles has completed, open the file one of those Countries that has been generated by this macro, and check that the Sort Order is still present. If yes, then immediately go to the next step. If not, then run the macro SortOrder, and when this is complete, again check that the SortOrder is present.
   3. When both steps (a) and (b) are successfully completed, run CopyTaxonomyAddDropdownsColourRows, and again check one file from a particular Country (preferably the last file).

**Create sign-up sheets and upload the data into Teams**

1. Check the stats per country
   1. Open excel template Country\_list&stats
      1. Check that there aren’t new countries this week (check the number of countries listed in the country\_level\_stats sheet B4 matches cell C4)
      2. If there are any new countries, add them to the list
      3. Check that the number of records per country, totalled matches the overall number of records we’re expecting (cell C5 in the Top Level Stats sheet)
2. Create the sign-up sheet
   1. Go into Teams>Coding Team>Files
   2. Download Countries\_requested.xls
   3. Save it as Sign-up\_sheet\_yyyymmdd
   4. Copy the total # records and # new records from the Country\_list&stats spreadsheet and paste values into Sign-up\_sheet\_yyyymmdd
      1. Check which countries have no new records and set the Name field = n/a and Status field = finished
      2. Whichever country was checked by Chris, assign that to him in the sheet and set Status = finished
      3. Check the totals at the bottom of the sheet are correct
      4. Hide columns F:I
3. Create the Admin sign-up sheet
   1. Save Sign-up\_sheet\_yyyymmdd as ‘Admin\_sign-up\_yyyymmdd’
   2. Un-hide columns F:I
   3. Filter on column F, Coding Team to remove entries in columns D and E where the person isn’t a member of the Coding Team.
   4. Re-hide columns F:I
   5. Going back to Sign-up\_sheet\_yyyymmdd, check what names are already allocated to certain countries. For a Country that does *not* currently have a Volunteer name associated with it, choose a file which has approximately 5-6 new records so that if there are any issues, they’re likely to be evident. Check the following have all worked:
      1. does the sort order work
      2. are the vlookups for columns Q, R, S referencing the taxonomy correctly
      3. Change Keep = N, does the whole row ~~strikethrough~~
      4. Check the conditional formatting is correct:
         1. If there are any non-permitted who\_codes, check that they have a red fill. Also check that 4.1.2 and 4.2.2 have an orange fill.
         2. Measure\_stage
         3. enforcement
         4. Non\_compliance\_penalty
         5. trigger
      5. are all of the column headings correct
      6. When you’re happy that the above looks right, perform a test-clean on your chosen file of 5-6 new records, and when completed it is to be uploaded into the Cleaned Data folder on MS Teams, and the ‘status’ column in the Sign-up sheet is to be updated to ‘Finished’.
4. Upload the remaining country files which need cleaning (i.e. those in the Split folder, but not the test-cleaned file) to Teams\Data to clean. Upload Sign-up\_sheet\_yyyymmdd to the same folder, and pin to the top.
5. Upload the Admin\_sign-up\_yyyymmdd to ADMIN ONLY folder, and pin to the top.
6. Check that all files have uploaded (Teams can be slow and temperamental)
7. Make files available for Orlagh for next week:
8. Once the run is finally complete, rename the current working folder as dd\_mmm\_yyyy, and let Chris or Orlagh know that the data is ready so they can put a message up on Teams.

updated by Prakash Sarnobat

17 November 2021