Chase Africa Data – Data Repository Setup Instructions

Maja

29.1.2019

Intro:

This document contains the step-by-step instructions for how the master data repository was set up to combine existing and future clinic level data collected by Chase Africa's partners.

These instructions have been tested by Maja but will be implemented by Catherine on the actual data in order to develop and keep the knowledge in-house.

This setup will only be completed once, but the information contained here should be archived as part of the Data Managment Plan.

Outline

These instructions cover the following:

1. Description of the data
2. Preparation of the data
3. Merger of the data

Describing the data

To begin with we should have three tables:

1. `responses`: the responses from the Google form.
2. `migration`: the raw data that was input from the Excel files.
3. `manual`: the raw data that will be input from the partners in the future if they don't use Google forms.

1. `responses` table

This table is the direct output of the Google form. This means the order of the columns and the names of the variables are defined automatically, so we'll stick with them as the baseline.

For easier reference in this document I will refer to these n variables as `var-1`:`var-n`: these are core variables that will continue to be collected from the partners.

The number of rows in this table keeps growing automatically with every new filled out form.

2. `migration` table

This table was created (as a one off) manual migration of the data contained in 6 Excel spreadsheets.

The table contains all the variables `var-1`:`var-n` as well as some additional *legacy* variables which are no longer used, these are referred here as `old-1`:`old:n`. The order of the columns and variable names were ad hoc, but they will now need to be adapted to conform with the `responses` table.

This table has 893 rows and that is it.

3. `manual` table

This table will start out as an empty copy of the `responses` table and will be used to manually enter data from partners who continue to report their data using Excel instead of using Google forms.

It therefore contains variables `var-1`:`var-n` in the same order as the `responses` table.

The number of rows increases manually with every new entry by Catharine.

Preparing the data

All three tables should be placed into a single Google Sheet workbook, each on it's own sheet, named appropriately.

1. Preparing the `responses` table

- This table will only get some *protection*. The data here should never be manipulated manually. To avoid accidentally corrupting the data you need to protec the range. There are two ways you can restrict permissions:
- show a warning when soemone trys to edit the range
- restrict who can edit by entering their email.

I would only suggest using the second option if you can use an email that is not normally used to access the sheet. In other words this *administrator* should be a different email, not your customary one, because otherwise it offers no protection at all, as it would essentially be saying "only the only person who has access can change the content". In that case it makes more sense to use the warning instead.

Highlight the whole sheet, rightclick on it, select `Protect range`, click `Set permissions` and pick your choice.

2. Preparing the `migration` table

Apply the same protection as above.

Now re-order the columns so they have the following order: `var-1`:`var-n` and then `old-1`:`old:n`. So the first columns must be exactly in the same order as the ones in the `responses` table.

- This means moving the other ones to the back i.e. right of the table.
- Don't forget to add the ones we discussed 21.1.: the flag for estimated data and the flag for multiple day clinics.
- This also means rethinking the `fund_round` or `fund_quarter` variable, because they don't currently exist in the `responses` table. It would probably be best to include it there, possibly with a dropdown: Q1, Q2..

You can use the Copy/Paste transposed function to easily compare the headers of both sheets side by side.

3. Preparing the `manual` table

To initialise this table simply copy the header from the `responses` table into a new sheet and make sure there is no protection.

Merging the data

Now open a new `master` spreadsheet that will contain all the merged data.

You will use two functions to get the data:

- importrange() - which allows you to connect to a different spreadsheet