Local Government Training Institute (LGTI) Course Data Curriculum

Data Collection Course

Module 2: After Data Collection

Disclaimer: This work is developed by School of Data with funding from The World Bank Tanzania Data programme. For more information visit https://www.schoolofdata.org.

Module 2: After Data Collection

STUDENT WORKBOOK

Introduction

You and your team have taken the time to assess the need for a data collection project and agree that this is necessary. You then spent time and resources mapping out the scope, stakeholders, budget and timeline for this exercise. You then went on to identify the necessary roles and responsibilities needed for the data collection project and have now developed a methodology and survey that will allow you to collect the data you need. You have also trained your data collectors and have practised what steps to take in case something goes wrong. You had a plan, managed to execute it and now have the data you need.

As much as you will be eager to share, analyse and present the data you have collected in its raw form, you cannot. Doing so not only means leaving in potential sources of errors and privacy issues in your data set. As responsible data experts and organisations, we want to make sure that our collected data contains what we expect and can be used without any ethical and privacy concerns.

This module focuses on what tools, techniques and best practices you need to turn your collected data into something that can be published, analysed and presented.

Lesson 1: Data Verification

Data Verification

This is the stage where you check to make sure that what you expected is what has been collected during the process. This includes ensure that the methodology proposed has been followed including the sampling methods, the targeted respondents have been reached and the number responses have been collected. This stems from reviewing the project scope and stakeholders in order to summarise the data collection project overview.

In order to verify our data, we first need to download it in a machine-readable format from KoBo Toolbox.

Downloading your data

On the main project page, you can download the data by selecting *Download data*. Your dataset can be downloaded in XLS, CSV, ZIP and/or KML formats. Once you have chosen which format, you will be arrive at the Export page where you can now download the file.

The online table analysis is very limited, thus, downloading your dataset will allow you to do your analysis offline and use your dataset in other platforms for analysis or visualization e.g. Tableau[1], Google Fusion Chart or OpenRefine.

Responses Overview

The overview should list the following information:

- Which organisation or individual organised the data collection project?
- What was the purpose of the data collection project?
- What was the timeline or data for data collection i.e what dates were the enumerators collecting data?
- How many data enumerators were involved in the project?
- How many responses were expected?
- How many target communities were targeted?
- What tools were used in the data collection?

Below is an example of the data collection overview for the Women Environmental Programme (WEP) data collection project mentioned in module 1.

> WEP, together with Open Knowledge International, trained data collectors on 28th and 29th July 2016. On 1st -9th August 2016 (7 working days), data collectors went into the communities to collect information about status of infrastructure using the Kobo Toolbox data collection kit. 20 data collectors were to collect data from 160 communities across 32 wards of the three Area Councils of AMAC, Gwagwalada and Kuje. A total of 2 responses each were expected from each community totaling 320 expected responses. At the end of the 7-working days data collection exercise, 333 responses were collected.

[1] http://www.tableau.com/

Cleaning Data with OpenRefine

OpenRefine is an open-source tool that allows a user to clean and process a large dataset using powerful functions. This section walks the reader through downloading, setting up and using OpenRefine for data cleaning.

1. Open Refine – Download it from [openrefine.org]
2. The sample Dataset – Download it from [Africa Open Data]

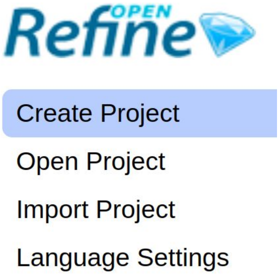
Step 1: Creating a new Project

OpenRefine is a data cleaning software that uses your web browser as an interface. This means it will look like it runs on the internet but all your data remains on your machine and you do not need internet connection to work with it. The main aim of OpenRefine is to allow you to explore and clean your data before you use it further. It is built for large datasets so as long as your spreadsheets can hold data then OpenRefine can as well.

To work with your data in OpenRefine you need to start a new project:

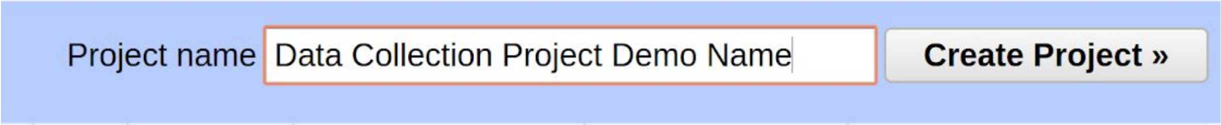
Creating a OpenRefine project

1. Start OpenRefine – this will open a browser window pointing to http://127.0.0.1:3333 if this doesn't happen open the link with your browser directly
2. Create a new project: On the left tab select the "Create Project" tab:



3. Click on "Choose Files" to choose your downloaded data collection responses file and click on "next" – you can also use the URL to the CSV directly if your data is hosted on the web.

4. You will get a preview on how OpenRefine will interpret your data – if you have selected a well formatted CSV or other file: this should be pretty automatic.
5. Review the preview carefully to make sure the data looks right. Double check character encoding. Much, but not all data uses UTF-8 these days, but make sure you don't see any funny characters in preview.
6. You may want to turn off "guess data types", particularly if you have data that contains leading zeros in numbers or identifiers which are significant.
7. Name your project in the box on the top right side and click on "Create Project"



Project name Data Collection Project Demo Name

8. The project will open in the project view, this is the basic interface you are going to work with: by default OpenRefine shows only 10 rows of data, you can change this on the bar above the data rows. Also you can use the navigation on the right to see the next or previous rows.

You now have successfully created your first OpenRefine project and can now start cleaning your data. **Remember:** although it runs in a web-browser, the OpenRefine server is still on your machine – all the data is there. This is especially important when handling sensitive data or personal data.

Verification Process

The assigned individual or team can now match this up with the actual responses from the data collection. The emphasis should be on the aggregate metrics that match up with what was intended.

Verification Question	OpenRefine Approach
How many responses came from each target location? Did this match up with the planned number?	Create a text facet on the location column and count the number of responses shown by each location name
How many responses came from each data collector? Did this match up with their assigned number	Create a text facet on each data collectors name and count the number of responses by their name
What target locations did each data enumerator have responses from? Did this match up with their assigned target location?	Create two text facets: one on the enumerator name column and the other on the location column. In the facet sidebar on the left, select the respective enumerator name and record the count of the respective location show in the location facet.
What dates did enumerators submit responses? Do they match up with the dates responses were to be collected?	Create a date sort on the submission date column by ascending order. This allows you to see the order of submissions. A date facet can also be created using the customized facet option.

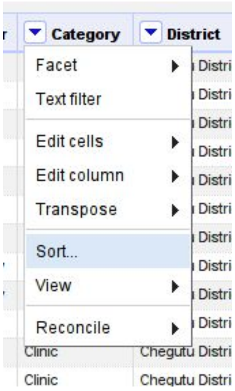
The questions above are not exhaustive but should help you begin to verify the accuracy of the data collected. Exploring these questions will either lead to additional questions which you will have to verify. Below, we provide a walk-through of the sorting and facetting features that can be powerful in your verification process.

Steps 1 & 2 Sorting and Facetting

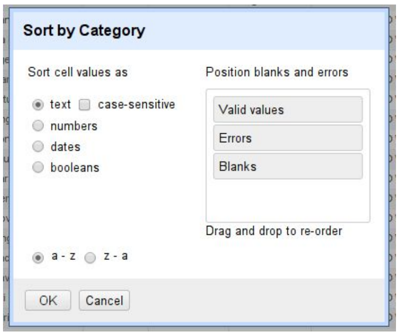
Once we created our project, let's go and explore the data and the OpenRefine interface a bit. Using OpenRefine might be intimidating at first because the interface is so different from spreadsheets but fear not, OpenRefine is a user-friendly tool that just requires a bit of getting used to. As was discussed in many of the data fundamentals course, one of the commonly used functions in spreadsheets is sorting and filtering data – to figure out minima, maxima or things about certain categories. OpenRefine allows you to do this as well and can be very helpful in the verification process.

Sorting rows

1. OpenRefine handles data similar to a spreadsheet: you have rows, columns and cells – a cell is a field defined by a row and a column.
2. To sort your rows based on a specific column click on the small downward triangle next to the column



3. Select "Sort..." to open the sorting dialog



Sort by Category

Sort cell values as

● text □ case-sensitive
○ numbers
○ dates
○ booleans

● a - z ○ z - a

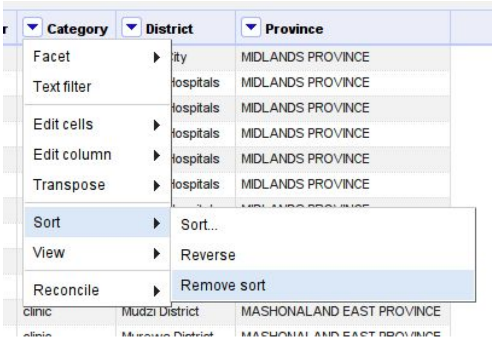
Position blanks and errors

Valid values
Errors
Blanks

Drag and drop to re-order

OK Cancel

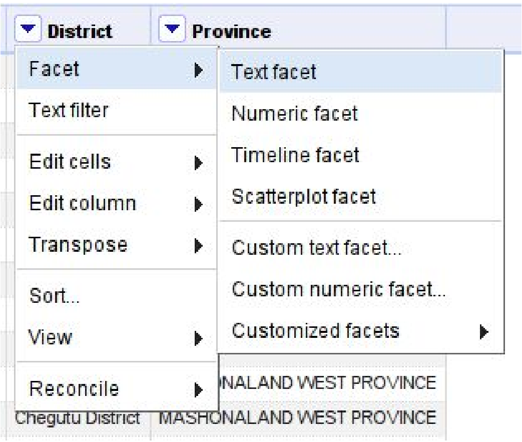
4. You can select what to sort the values as and then what order to sort in. (We'll sort in text, since for now we only have text columns)
5. Click "OK" and your rows will be sorted based on the column
6. To undo the sort, click on the column options again, select "sort" then "remove sort"



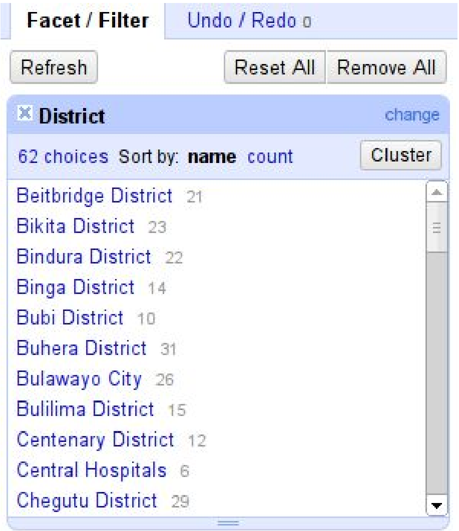
Facetting rows based on a column

The other frequently used function in Spreadsheets is filtering – in OpenRefine this is called facetting. Facetting in OpenRefine is incredibly powerful.

1. Select the column options for the column you want to facet with
2. Select "Facet"



3. You can facet differently for text, numbers or dates – let’s facet as text – click on “Text facet”
4. This will open a facet in the left bar



Beitbridge District	21
Bikita District	23
Bindura District	22
Binga District	14
Bubi District	10
Buhera District	31
Bulawayo City	26
Bulilima District	15
Centenary District	12
Central Hospitals	6
Chegutu District	29

5. Now select one or more of the choices and you’ll see how your data rows are limited to just those selected.
6. Of course you can add more than one facet and thus filter more than once.

With knowledge of the sorting and facetting features of OpenRefine, we can now go ahead and verify and clean our data.

Step 3: Dealing with Blank Cells

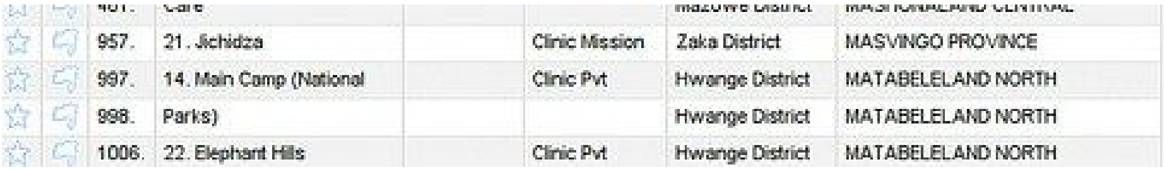
The next stage of data verification is identifying and explaining any blank cells that exist. Blank cells in our collected data could represent responses that were not applicable or failure to complete an entry for a specific questions. In an ideal world, including automatic validation in our mobile data collection form should reduce the presence of blank cells. However they cannot always be avoided.

To ensure the quality of our data, there are two options we can take in the case of blank cells. To fill them in with the right values or to delete them. In the case of filling it the right values, this is where collecting the name of the enumerator who collected the response is vital. The individual performing data verification should reach out to the respective enumerator with the help of the project manager. Once clarity is obtained, the team can decide whether to fill in the blank cell or delete the entire response.

If you look closely at your facets, you'll notice that on the bottom you have a selector saying "(blank)" for this – we'll need to deal with it. This is a common problem in a data collection project but it is something that can be dealt with easily in OpenRefine.

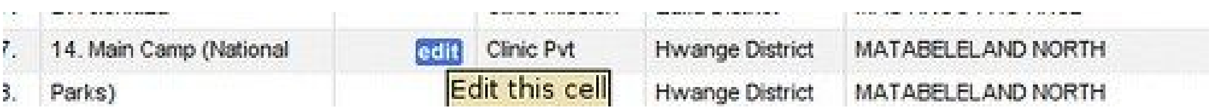
Filling in the (blank)s

1. Choose the "(blank)" facet in your "Owner" column
2. If you look at some of the rows, you'll see that there was a mis-split of the columns and the owner actually ended up in the "Category" Column

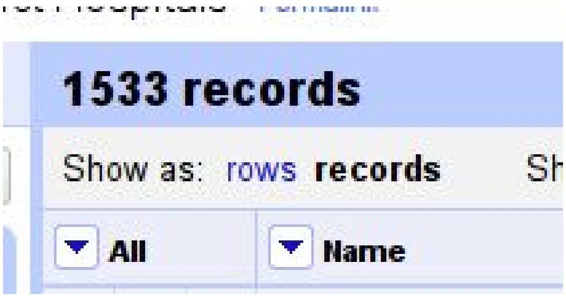


☆		[illegible]	Care			Mazowe District	MASHONALAND CENTRAL
☆		957.	21. Jichidza		Clinic Mission	Zaka District	MASVINGO PROVINCE
☆		997.	14. Main Camp (National		Clinic Pvt	Hwange District	MATABELELAND NORTH
☆		998.	Parks)			Hwange District	MATABELELAND NORTH
☆		1006.	22. Elephant Hills		Clinic Pvt	Hwange District	MATABELELAND NORTH

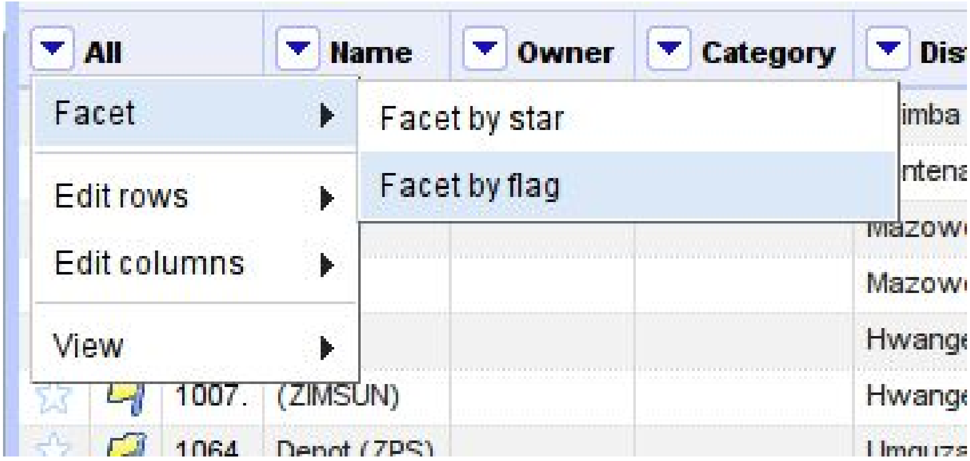
3. To fill this into the "Owner" Field hover over the cell you want to fill in and click the "edit" button.



4. If you click the “Edit” button you can add the Owner there – don’t forget to also correct the “Category” cell.
5. You’ll notice some rows seem to be erratic – they don’t have a name that makes sense and no further information – you can flag these for deletion by clicking on the little flag.
6. Do the same with the “Category” Column – the Category is sometimes joined with the “Name” column
7. Now let’s delete the flagged rows – make sure you are in row mode for this: for this click on “row” in the top left corner above the data.

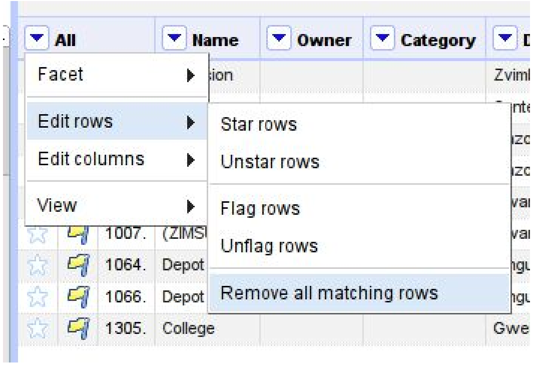


8. Open the column options for “All” and select “Facet” – “Facet by Flag”



9. Now you can select “true” in your flag facet on the left.

10. Now let’s delete the flagged rows: in the Column options for “All” select “Edit rows” – “Remove all matching rows”



Lesson 2: Data Cleaning

Cleaning typically happens after verification but the reality is that both may need to be happening hand-in-hand. Verifying a response may lead to correcting some aspect that was entered wrongly. The takeaway for this stage is to get the data into a format that can be easily read, analysed and easily visualised without any setbacks to the user.

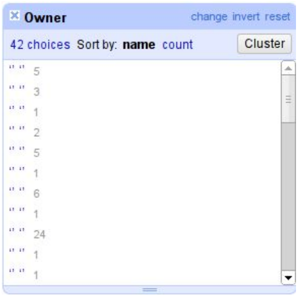
Cleaning can involved several issues including removing whitespaces (spaces and newlines), spelling mistakes, mismatched capitalistions, and mislabelled responses. These usually result from entry of responses into text fields. Fortunately OpenRefine’s ability to create facets on specific columns allows us to isolate categories of responses and directly edit them correctly.

Step 4: Fighting the Invisible Man

As illustrated in The Data Fundamentals Course, having spaces or newlines in your data fields is a problem. Since this is a very common problem, OpenRefine has specific functions to remove whitespaces that shouldn’t be there.

Removing hidden whitespaces

1. Let’s start cleaning our Dataset with the Owner Column
2. Create a Text Facet for the Owner Column as described above
3. You will notice that there are several things odd in the column: It starts with a long list of similar looking entries – we’ll deal with it later.

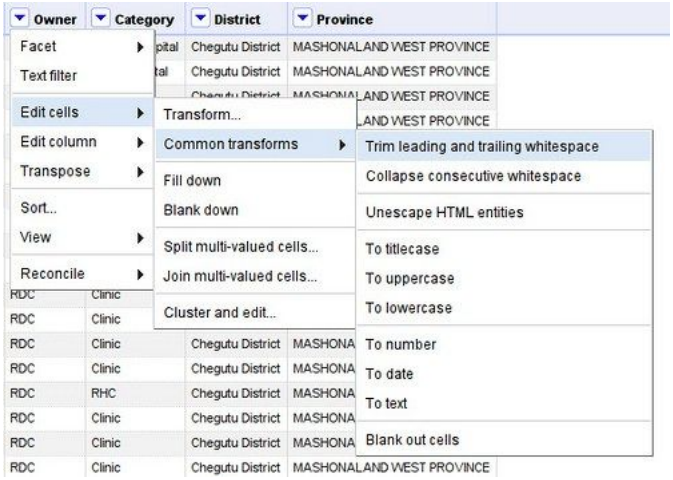


Owner

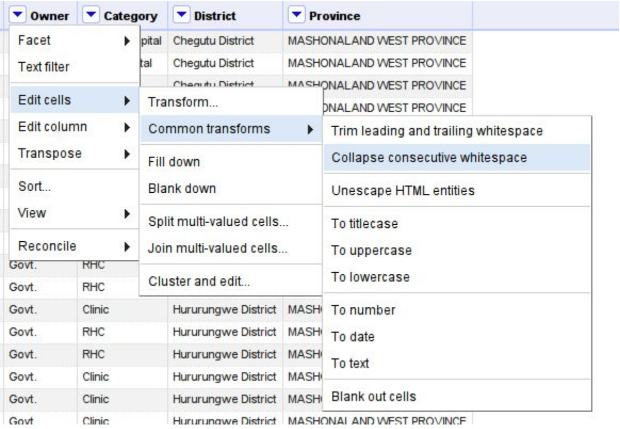
42 choices Sort by: name count

" "	5
" "	3
" "	1
" "	2
" "	5
" "	1
" "	6
" "	1
" "	24
" "	1
" "	1

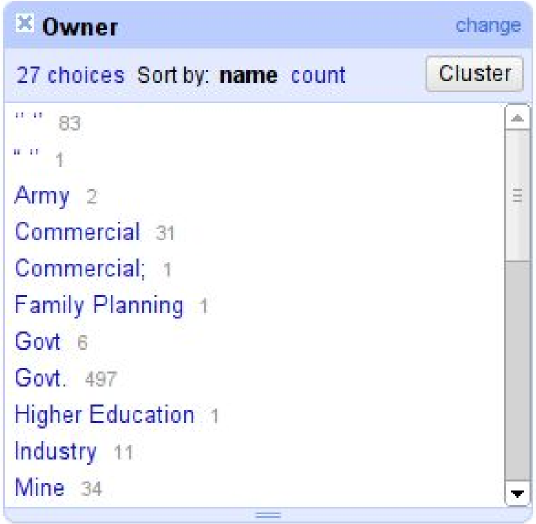
4. Although they look similar to you, they are different for the computer – there is a different number of spaces between the quotes.
5. Scroll down and you’ll notice that some entries will be there twice – although they look similar. There are two entries for Municipality that look exactly the same. This is because they have whitespaces at the end.
6. Refine can help you clean this up in an instant – open the column options for the “Owner” column
7. Select “Edit Cells” – “Common Transforms” – “Trim leading and trailing whitespaces”



8. This will remove whitespaces in the beginning and at the end of your column
9. Check Municipality and you'll note that there's only one choice now – perfect. Now let's deal with the list at the beginning.
10. Select "Edit Cells" – "Common Transforms" – "Collapse consecutive Whitespaces"



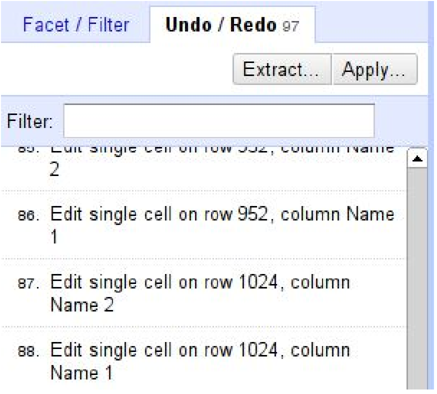
11. You'll see the multiple choices have been reduced to two choices in an instant



" "	83
" "	1
Army	2
Commercial	31
Commercial;	1
Family Planning	1
Govt	6
Govt.	497
Higher Education	1
Industry	11
Mine	34

12. Now our list already looks a lot cleaner!
13. Go ahead and apply the two transforms to all your columns.

Once you made your transforms you might wonder: What if I made a mistake? Also if you work with data you generally want to keep track of what you did to the data. Since OpenRefine was build with data processing in mind it keeps track of what you're doing with your data and allows you to go back and forth in time. To see your history of changes click on the "Undo/Redo" tab on the left.



85. Edit single cell on row 952, column Name 2
86. Edit single cell on row 952, column Name 1
87. Edit single cell on row 1024, column Name 2
88. Edit single cell on row 1024, column Name 1

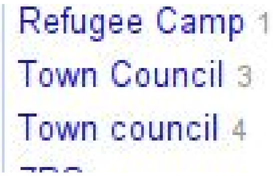
You see all the changes you made – by simply clicking on one of the steps you'll be undoing all the changes after the step (don't worry you can redo pretty much the same way). Play with this system until you are comfortable.

Step 5. Reconciling categories

A quick look at our categories and you'll notice that not everything is well in Owner land – still some categories that should be the same are not. The same for the "Category" column – let's reconcile them.

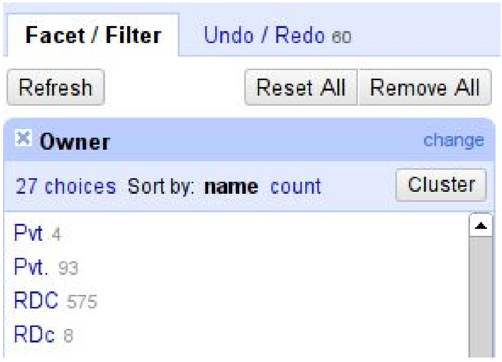
Reconciling Categories

1. Create a Facet for the column you want to reconcile (in our case this is "Owner")
2. The first step is to bring the categories to the same case – see for example "Town Council" and "Town council" – the difference is just one letter.



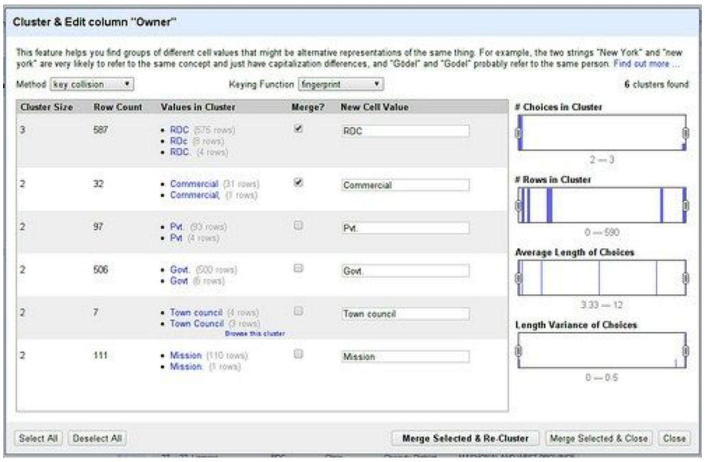
Refugee Camp 1
Town Council 3
Town council 4

3. Refine can help you to automatically find the categories that belong together – a feature it calls "Clustering". To activate clustering click on the "Cluster" button in your facet.



Pvt 4
Pvt. 93
RDC 575
RDc 8

4. You will end up in the clustering menu – as you can see Refine is pretty smart about which things should belong together



Cluster & Edit column "Owner"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

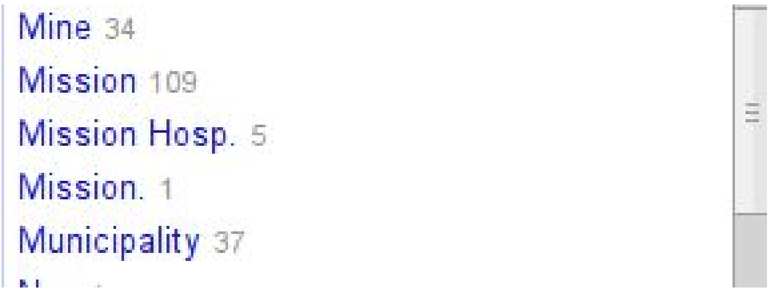
Method: key collision

Keying Function: fingerprint

6 clusters found

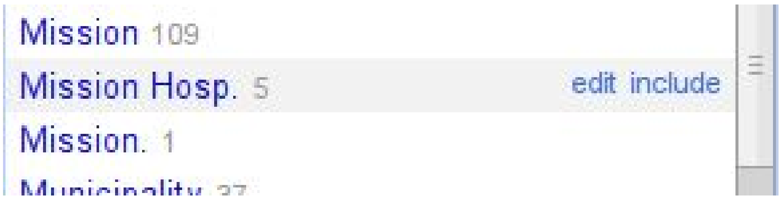
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	587	• RDC (575 rows) • RDc (8 rows) • RDC. (4 rows)	☑	RDC
2	32	• Commercial (31 rows) • Commercial, (1 rows)	☑	Commercial
2	97	• Pvt. (93 rows) • Pvt (4 rows)	□	Pvt.
2	506	• Govt. (500 rows) • Govt (6 rows)	□	Govt.
2	7	• Town council (4 rows) • Town Council (3 rows) Browse this cluster	□	Town council
2	111	• Mission (110 rows) • Mission. (1 rows)	□	Mission

5. Check the "merge" checkbox if you want the two categories to be the same. Once you marked all the categories you want to merge click on "Merge selected & Re-Cluster"
6. If Refine doesn't find more values to be similar change the "Keying Function" and see whether you can find more similar categories – if not: simply click close to continue.
7. This reconciled some of your values – let's go on.
8. Look at "Mission" for example we have three different categories for what should be one – Refine did not automatically find them.



Mine	34
Mission	109
Mission Hosp.	5
Mission.	1
Municipality	37

9. Let's change them all to mission
10. Hover over "Mission Hosp." notice the "edit" button at the end?



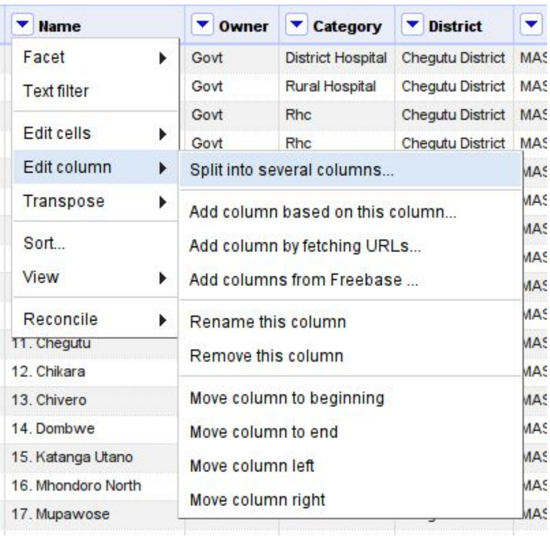
11. Click on Edit – this will open the field for editing. Change the name to “Mission” – this will change “Mission Hosp.” to “Mission” in all cells where it appears – continue on to change all the fields you can find.
12. Repeat reconciling for “Category”

Step 6: Splitting Columns

If you look at the “Name” column in our dataset you’ll notice that the names commonly start with a number (this is an enumeration of hospitals in a district – and is an artifact from extracting the data). Let’s clean this up and split the number and the name.

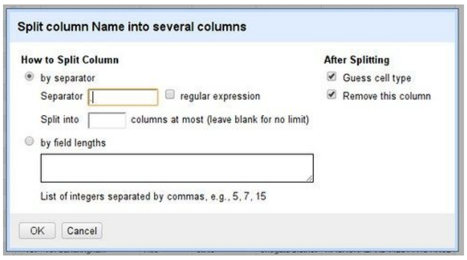
Splitting Columns

1. To Split a column select “Edit Column” “Split into several columns”



2. We want to split at a “.” since the number generally ends with a “.”

3. Enter "." into the Separator for in the split menu – since we only want to have two new columns enter 2 into the field below so the sentence reads "Split into 2 columns at most"



Split column Name into several columns

How to Split Column

● by separator

Separator [ ] □ regular expression

Split into [ ] columns at most (leave blank for no limit)

○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☑ Guess cell type

☑ Remove this column

4. Click on "OK" and you'll end up with two columns.
5. On some of the rows the split will fail – to fix those, create a facet on the second column and select "(blank)"
6. You can now manually fix the cells.

Taking Care of Data Privacy Issues

Due to potential privacy concerns, any columns with information that can be used in the re-identification of data collectors or the people they surveyed should often be removed. This step should usually be performed after verification and cleaning is complete since personal data may be needed by the team for these stages.

The specific columns to delete will really depend on the type of questions that were included and it is important to decide on this with the survey designer. KoBo Toolbox also collects some individually-identifiable data such as ***simserial, subscriberid, deviceid, phone number*** which should be deleted.

Once you have decided on which columns to delete before publication, obtain another copy of your raw data (i.e one that has not be verified and cleaned) from KoBo Toolbox and delete the selected columns that contain personal detail from it. This will be the

anonymised raw data file that will be published together with the ***anonymised cleaned data file***.

To get a more in-depth overview of dealing with privacy issues and making your data anonymous, see this hands-on guide from the Open Data Institute. In a later chapter, we will go over how to think about the privacy and security of your data collectors, respondents and data users using the concept of "responsible data". For more information now, you can review the responsible data handbook.

Publishing Data

At the stage, your data is ready for publishing. Depending on your project scope and stakeholders, you will have to decide on the best platform to publish your data. However, data publication is not just about publishing the cleaned data but also the data collection methodology and steps taken to turn the raw data into the cleaned version. Fortunately, you have all these resources already available and just have to organise them in a coherent structure for publication. The following resources should be included in your publication:

- Data folder
    - Anonymised raw data file
    - Anonymised cleaned data file
- OpenRefine Extracted History JSON file
- Survey Methodology
- Column Name Description

Once you have these resources ready, you can choose publish your data on Google Drive[2], GitHub[3], DataHub[4]

[2] https://gsuite.google.com/learning-center/products/drive/
[3] https://guides.github.com/activities/hello-world/
[4] http://datahub.io/

Data Management & Best Practice

Common Mistakes in Data Collection:

Obviously, you can only collect data about thing you have access to so data collection is constrained by the available access to the subject being studied. The first part of defining a data collection strategy is understanding when or where will the act of data collection will take place and taking care to think about the following common challenges and mistakes:

- Enumerators accidentally delete photos from gallery.
- Enumerators upload trial data in an actual dataset.
- Enumerators take blurred images and some pictures are taken in low light.
- Enumerators type data with spelling mistakes. Example: 0 (zero) is spelled as O (alphabet O).
- Enumerators change the default storage location to internal memory in the devices provided to them. This causes confusion as half the data is stored in removable memory and half the data is stored in device memory, causing panic among enumerators who believe the data is lost.
- Enumerators use an additional device but forget to configure it properly.
- Typos and mistakes appear, due to enumerators hurrying.
- Enumerators try to record the GPS location from inside a building, which results in poor accuracy and delays.
- Enumerators try to submit high volume data (with photos) from a place with low bandwidth internet service, which can result in frequent upload unsuccessful errors. This can raise frustration levels and give the impression that there is a problem with the data upload.
- Enumerators accidentally delete files, using the file manager, which are essential for survey software to function.
- Whilst playing with and exploring the app, enumerators change default settings. This may result in the app behaving differently from the way in which the enumerators were taught to expect during the training and result in confusion.
- A device is damaged. The most common type of damage in the field is breakage of the device screen.

- Enumerators do not complete many necessary fields which may not be compulsory in the survey form.
- Enumerators do not fill out some data points, as they think might have already captured it previous questions. For example, in one case enumerators did not capture a photo of the main building of a school, as it had already been captured in a previous photo of the school sign.
- Device auto-corrects, causing errors in the entries.
- Enumerators try to keep a backup of the data but do not copy the whole folder which results in loss of data.
- Enumerators accidentally delete data while transferring data from SD card to laptop.