Attributions

This session was built using Carpentries materials and workshop materials offered by the University of Victoria. You may access their complete materials here:

- Data Carpentries link: https://datacarpentry.org/openrefine-socialsci/
- Library Carpentry link: https://librarycarpentry.org/lc-open-refine/
- University of Victoria link: https://docs.google.com/document/d/1sMO-KwmD3GSwrYuNyFGr0zV0zsxh_kkTJSAuEmmmMTA/edit

OpenRefine manual in case you want to read: https://docs.openrefine.org/

Handy GREL Cheatsheet: https://guides.library.illinois.edu/openrefine/grel

Schedule and Session Objectives

The primary goal of this session is to walk you through some key features of OpenRefine that you are most likely to use in library work. In this session you will learn how to:

- Start a project
- Use OpenRefine to clean and organize your data by:
    - Using facets and clustering to scan your data standardize terms
    - Using common transformations via GREL to shift, split, and batch edit your data
- Export your project to your preferred format

OpenRefine Tips and Tricks

Flags and Stars

Flags and stars are a great way to create a subset of your data to work with. When you add a flag or star to your data, OpenRefine treats the items as a Facet, which enables you to work with this information using the Facet functions. You can manually Flag or Star any row by clicking on the icon on the left of the screen. You may also Flag or Star multiple rows or Flag/Star more rows matching your current Flag/Star rows by using the "**All"** dropdown menu and selecting "**Edit rows → Star rows"** or "**Flag rows**."

For more information: Edit by Star or Flag

Undo and Redo

OpenRefine keeps track of every operation you perform on your data, and easily allows you to undo or redo steps as you work through your data. What is more, the way OpenRefine records the steps you have taken even allows you to take the steps you've carried out on one data set and apply it to another data set by a copy and paste operation. You can access the Undo and Redo options in the lefthand panel. Undoing or Redoing an action involves simply clicking on the step you would like to revert to. To save the history of your actions for

later, you can click the **Extract** button, where you can select the steps you wish to save. This extract will be in JSON format, which can be copied and saved in a text file. When you wish to apply these steps at a later date, you can simply paste in the JSON and click Apply. Undo/Redo data is stored with the Project and is saved automatically as you work, so next time you open the project, you can access your full history of steps you have carried out and undo/redo in exactly the same way.

For more information: History (Undo/Redo)

Removing blank rows

Sometimes your data will have blank rows scattered throughout the worksheet. Here's how to get rid of them.

1. Looking at your data in the display window, we see that there are a number of empty rows separating our data.
2. Let's search the file for rows with cells that are all blank. Click the dropdown arrow next to "**All**" (far left). Hover your cursor over **Facet**, and select **Facet by blank (null or empty string)**.
3. In the facet that appears, you will see "true" and "false" with numbers. The blank values will show up ast "true" as they meet the condition of the filter (blank rows). Hover your mouse over the true option, and select "**include**".

Trimming whitespace

Another common issue with data are spaces at the beginning and end of a cell's data. We usually want to remove these, not just because they can make the cell data look messy, but also because a machine will treat these extra spaces in your data as part of the string. You can remove leading or trailing whitespaces on import, but over the course of your transformations you may actually be creating some, especially when you split or join columns.

If you unchecked that box when importing data, or if leading or trailing whitespaces were introduced while splitting columns, or other operations, OpenRefine also provides a tool to remove blank characters from the beginning and end of any entries that have them.

1. Edit the village on the first row to introduce a space at the end, set to God .
2. Create a new text facet for the village column. You should now see two different entries for God, one of those has a trailing whitespace.
3. To remove the whitespace, choose **Edit cells > Common transforms > Trim** leading and trailing whitespace.
4. You should now see only four choices in your text facet again.

Downloading OpenRefine

1. Download OpenRefine from the following website: http://openrefine.org/download.html

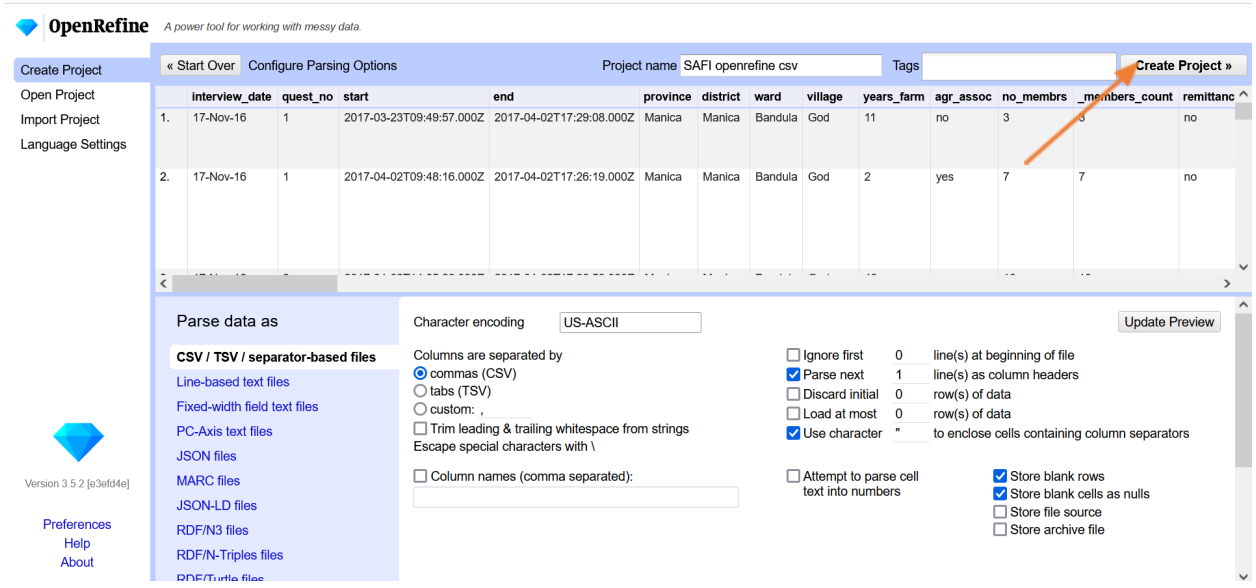
Please use 3.5.2, the latest stable release of OpenRefine 3.5, released on January 26, 2022. Please backup your workspace directory before installing and report any problems that you encounter. A change log is provided on the release page.

- **Windows kit**, This requires Java to be installed on your computer. Download, unzip, and double-click on *openrefine.exe* or *refine.bat* if the former does not work.
- **Windows kit with embedded Java**, includes OpenJDK Java, available under the GPLv2+CE license. Download, unzip, and double-click on *openrefine.exe* or *refine.bat* if the former does not work.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. You do not need to install Java separately.
- **Linux kit**, Download, extract, then type *./refine* to start. This requires Java to be installed on your computer.

2. Download the dataset here: https://ndownloader.figshare.com/files/11502815

Starting a project in OpenRefine

1. Run the OpenRefine app (or double click on the openrefine.exe file). N.B *Although the OpenRefine application will open in your default browser window, it is being run locally and data is not being uploaded or shared online.*
2. The first page you will see is the data import page. We will be creating a new project by importing a data file. Make sure that "**Create Project**" is highlighted in the left-most menu. OpenRefine allows you to connect to either local or online sources. We will be uploading a local *.txt file.
3. Under the **"Get data from"** heading, select "**This Computer**" and click "**Choose Files**…"
4. In the window that opens, navigate to the "**SAFI-openrefine-csv**" file, select it and click "Open".
5. Back in the OpenRefine window, click "**Next** »"
6. Before importing your data, OpenRefine provides a preview of how it is being read.
7. Our file organizes the data contained by tab-separated values (TSV). Under the "Parse data as" menu, select "**CSV / TSV / separator-based files**". Note all the different file types you can choose from!
8. See the different options you have for parsing data. You can, for example, ignore lines at the beginning of the file or parse some lines as headers.
9. Now that your import settings are set up, you can click "**Create Project** »" in the top right corner.



Note: OpenRefine has issues with Firefox and will not work with Internet Explorer.

Scanning and standardizing your data with Facets and Clustering

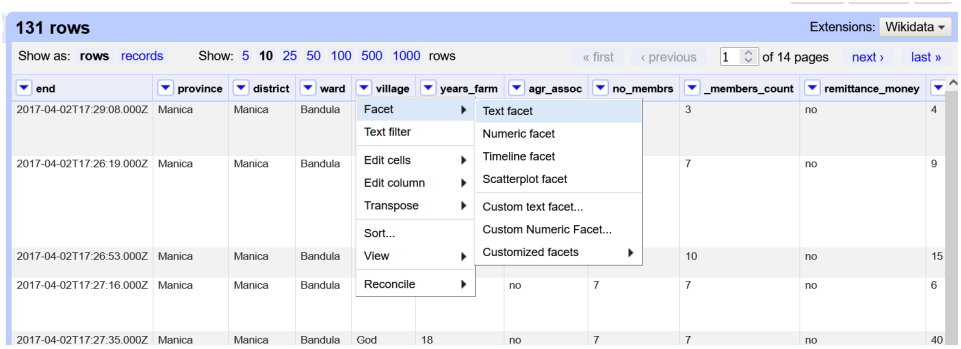
Facets

Using the **Facet** function groups all the like values that appear in a column and allows you to filter the data by these values and edit values across many records at the same time. Facets enable you to scan and cluster your data and track errors and inconsistencies by showing you a bigger picture of your data, while also giving you the option to filter down to subsets that you would like to change in bulk. There are several facet types: text, numeric, timeline, scatterplot, and custom. We're going to focus on text facets today.

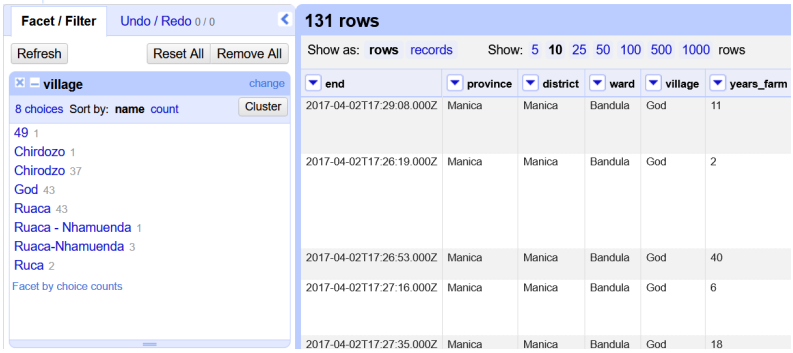
Learn more: Exploring Facets

Here we will use faceting to look for potential errors in data entry in the **"village"** column.

1. Scroll over to the **"village"** column.
2. Click the down arrow and choose **Facet > Text** facet.



3. In the left panel, you'll now see a box containing every unique value in the **"village"** column along with a number representing how many times that value occurs in the column.
4. Try sorting this facet by **name** and by **count**. Do you notice any problems with the data? What are they?
5. Hover the mouse over one of the names in the **Facet** list. You should see that you have an edit function available. N.B. At this point we could clean up the errors that we see, but we're going to show you how to clean these up with clustering in the next section.
6. You also have an include option that allows you to include as many facets as you wish.



Exercise

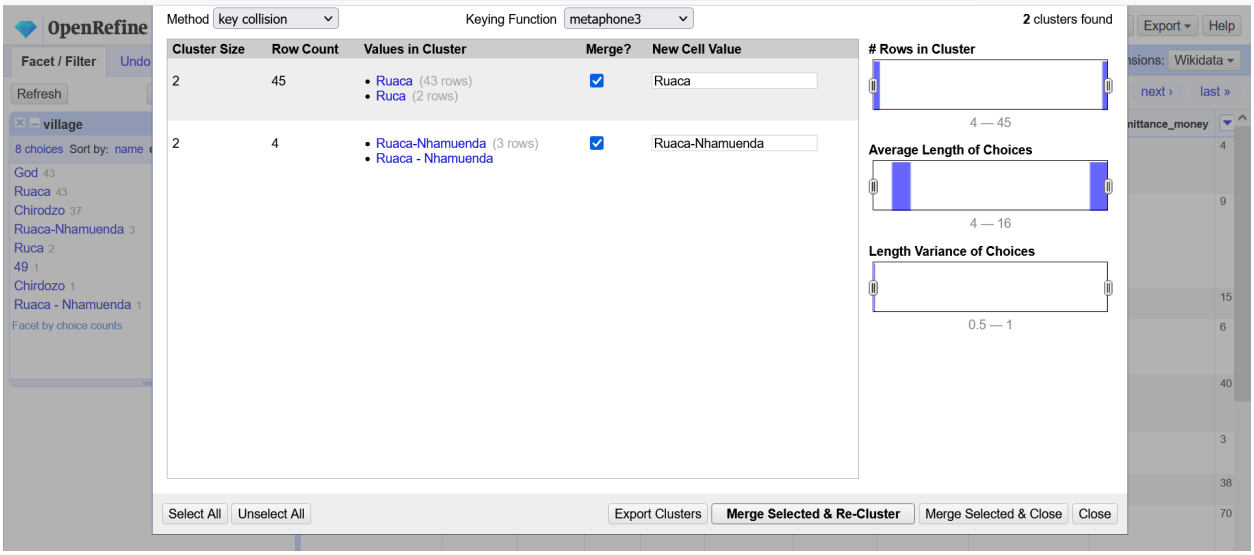
1. Using faceting, find out how many different "**interview_date"** values there are in the survey results.
2. Is the column formatted as Text or Date?
3. Use faceting to produce a timeline display for "**interview_date**". You will need to use **Edit cells > Common transforms > To date** to convert this column to dates.
4. During what period were most of the interviews collected?

Clustering

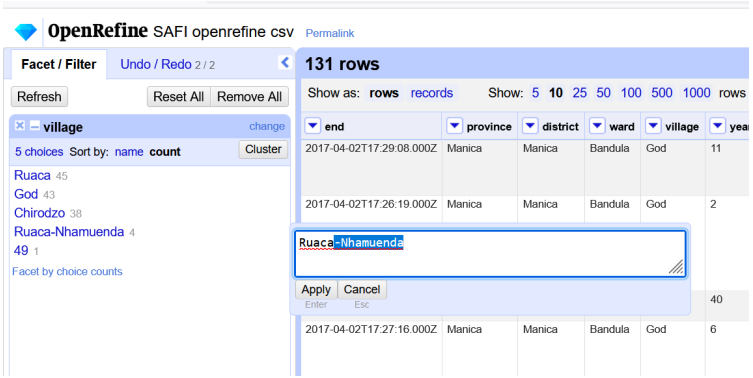
Clustering is a great way to find typos, abbreviations, case differences, and variant spellings. Clustering searches through data in a column and gathers the various facets together that might be variations of the same thing. This is crucial when dealing with text facets, since OpenRefine treats a text facet as a string, meaning any variation (including letter case) will read as a distinct facet. For example, through clustering you can find street, Street, and St. and standardize them to the term of your choice. In the next example we will walk you through some clustering. For now, it's not crucial to understand the exact algorithm for each clustering method. What is important to remember is to keep track of what you are using, experiment with what works, and be careful not to merge terms that aren't the same!

Let's take a look at an example.

1. In the **"village"** Text Facet we created in the step above, click the **Cluster** button.
2. In the resulting pop-up window, you can change the **Method** and the **Keying Function**. Try different combinations to see what different mergers of values are suggested.
3. Select the **key collision** method and **metaphone3** keying function. It should identify two clusters.



1. Click the **Merge?** box beside each cluster, then click **Merge Selected and Recluster** to apply the corrections to the dataset.
2. Try selecting different Methods and Keying Functions again, to see what new merges are suggested.
3. You should find that using the default settings, no more clusters are found, for example to merge Ruaca-Nhamuenda with Ruaca or Chirdozo with Chirodzo.
4. To merge these values, we will hover over them in the village text facet, select edit, and manually change the names. Change Chirdozo to Chirodzo and Ruaca-Nhamuenda to Ruaca. You should now have four clusters: Chirodzo, God, Ruaca and 49.



Important: If you **Merge** using a different method or keying function, or more times than described in the instructions above, your solutions for later exercises will not be the same as shown in those exercise solutions.

Want to learn more? Check out the documentation: [More on clustering]

Introducing Transformations

Transformations are ways of manipulating data in columns beyond clustering/filtering. In OpenRefine, these transformations are done by using a language called GREL (General Refine Expression Language) which you can think of as similar to the formulas in Excel. With transformations you can:

- Split data that is in a single column into multiple columns (e.g. splitting an address into multiple parts). You can also join columns!
- Standardize the format of data in a column without changing the values (e.g. removing punctuation or standardising a date format)
- Extract a particular type of data from a longer text string (e.g. finding ISBNs in a bibliographic citation)

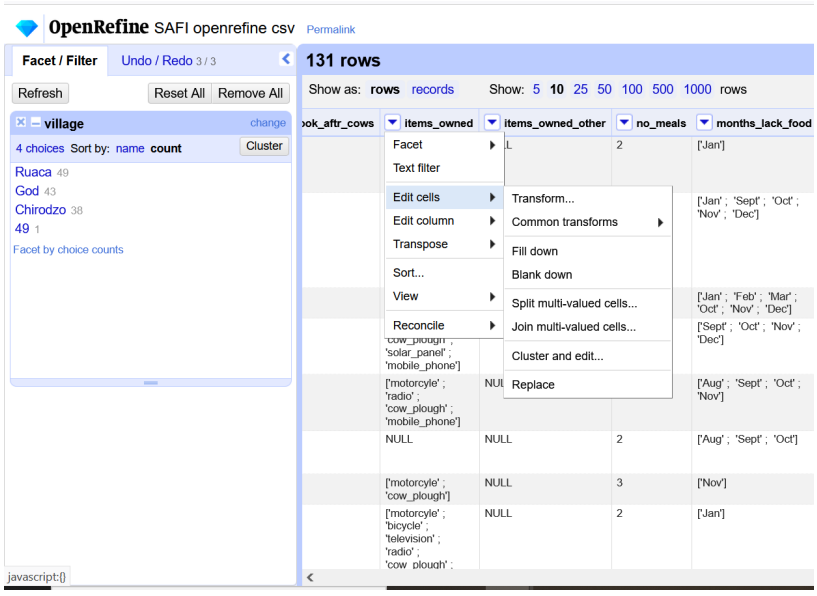
We do not have time to cover everything you can do with GREL today, but we're going to cover some popular transformations so you can get a sense of the power of the program, and how you can use it alongside Excel to clean your data.

Learn more about what GREL can do: https://docs.openrefine.org/manual/grelfunctions.

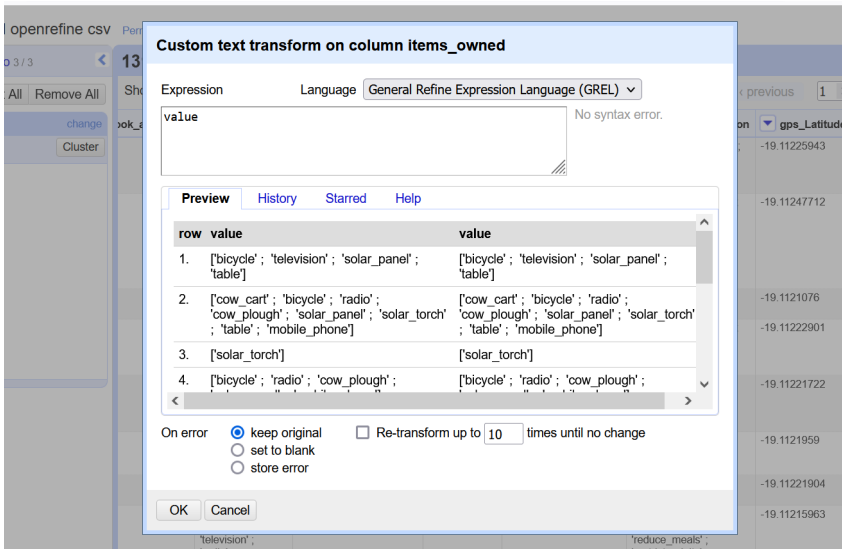
Cleaning Up Cell Data with Transformations

The data in the **"items_owned"** column is a set of items in a list. The list is in square brackets and each item is in single quotes. Before we split the list into individual items in the next section, we first want to remove the brackets and the quotes.

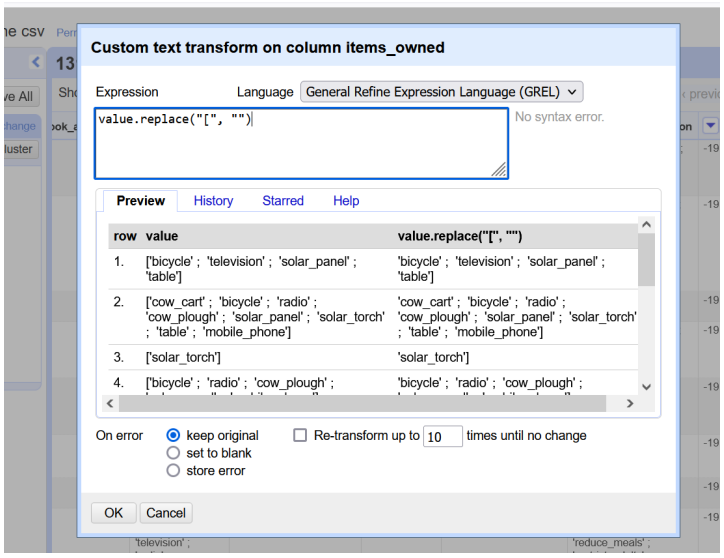
1. Click the down arrow at the top of the "**items_owned**" column. Choose **Edit Cells > Transform...**



2. This will open a window into which you can type a GREL expression.



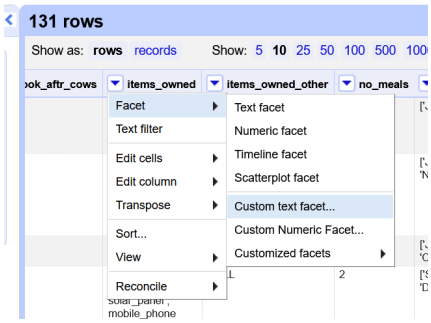
3. First, we will remove all of the left square brackets ([). In the Expression box type: value.replace("[", "") and click OK.



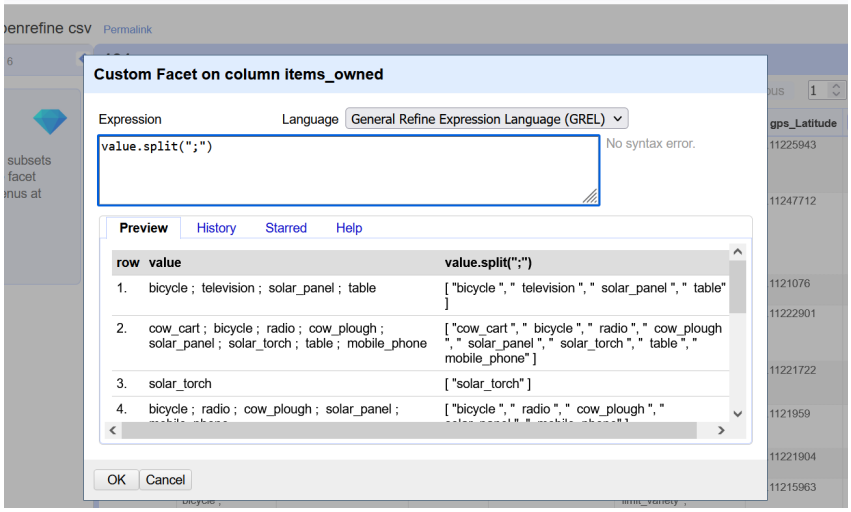
4. What the expression means is this: Take the value in each cell in the selected column and replace all of the “[” with “” (i.e. nothing - delete).
5. Now try removing the closing bracket and quotation marks using the same method.
6. Click OK. You should see in the “**items_owned”** column that all of the brackets and quotation marks have been removed.

Now that we have cleaned out extraneous characters from our “**items_owned”** column, we can use a text facet to see which items were commonly owned or rarely owned by the interview respondents.

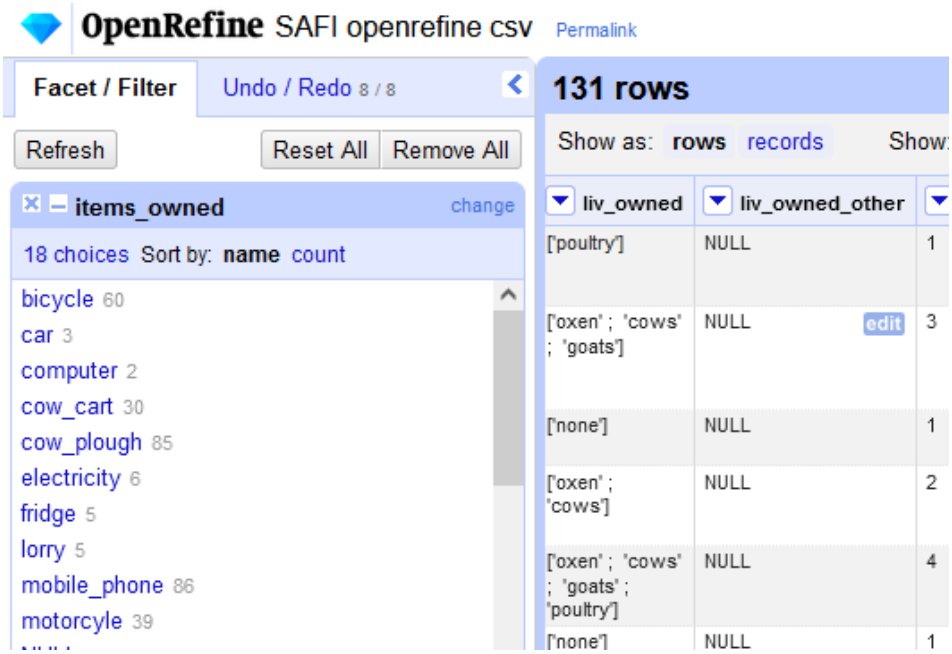
1. Click the down arrow at the top of the “**items_owned**” column. Choose **Facet > Custom text facet...**



2. In the Expression box, type value.split(";").



3. Click OK
4. You should now see a new text facet box in the left-hand pane with every unique value in the cell data.



Exercise

Perform the same clean up steps and customized text faceting for the "**months_lack_food"** column. Which month(s) were farmers more likely to lack food?

Splitting Data in Columns

Cells

Sometimes you need to split information that is contained in the same cell into separate rows. In the previous section we worked in the "**items_owned"** column, cleaning out some extra characters and learned how to facet multivalued cells. But if you are planning to do some analysis in another program like Excel you might want to split up this data.

1. Click the drop down arrow next to the "**items_owned**" column heading. Hover over "**Edit cells**" and select "**Split multi-valued cells**…"
2. In the menu that appears, specify that column will be split by separator, and input a semi-colon "**;**" into the text box.

Take a look - you'll see that each value has been added into a new row. Isn't it great the way that it displays? This is a great option for when you want to analyze multiple answers to the same question in something like a pivot table.

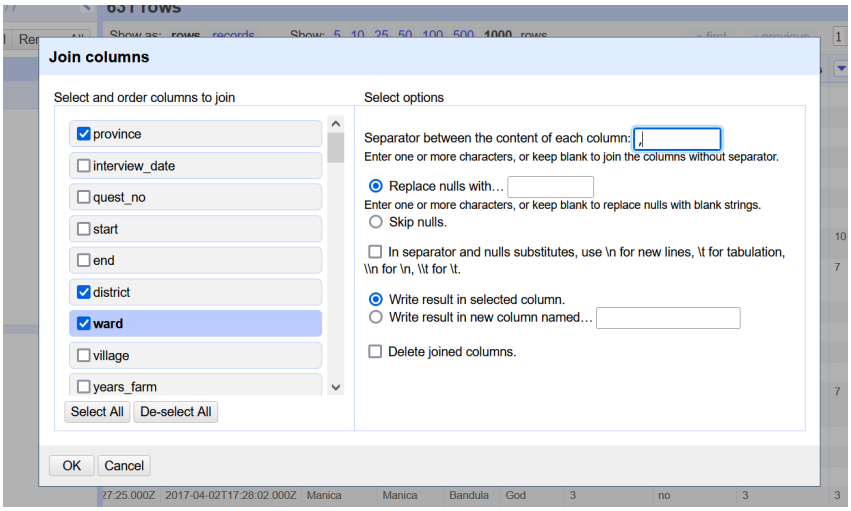
Exercise

Can you split **"liv_owned"** in the same way? When would you do this? How would you do this?

Joining Columns

You can also join the information from two or more columns. When you perform this function all of the strings will be added into the topmost cell in the record in the order that they appear.

1. To get started, select the first column that you want to join, for us it will be **"province".** In the dropdown menu select **"Edit Column",** then **"Join Columns".**
2. A popup window will open with a list on the left where you can pick the columns that you wish to join, so in this case we would like to join province and district to make a phrase "[province], [district], [ward]".



3. Add a separator. In our example we will add the following **“, “** (comma space). You should generally pick a separator that isn't commonly used in your data so you know where the change is, and so that it is easily removed programmatically. A “|” (pipe) or “;” (semicolon) are two good choices. Since we’re joining this to make a phrase for people, we’re adding characters to make it read normally.

Example
Try joining **”ward”** and **“village”.** Are there any other instances in this spreadsheet that you would like to join? What kind of separator would you use?

Final clean up and export

Time to save and export. What file type will you choose? OpenRefine has lots of options. OpenRefine saves your progress automatically.

1. If you want to export your work in OpenRefine so that you can resume it on another computer, click the **Export** button in the top-right corner, then select “**Export project**”. You will have the option to save the file to a local directory, or to Google Drive.
2. If you are finished cleaning your data, you can choose to export a file with the current settings enabled. the **Export** button in the top-right corner, then select your preferred format (e.g. Excel (.xls), Tab-separated value, etc.).