Data Cleaning Using OpenRefine

Slide 1: Hi and welcome to Data and Donuts, I'm Tobin Magle, the Data Management specialist at CSU's Morgan Library. Today's session is about how to clean up "messy" data.

Slide 2: To put this topic in the context of the research data lifecycle, it occurs after data collection, but before analysis. The goal of this session is to clean, and enhance data using a powerful data cleaning tool called OpenRefine.

Slide 3: Open refine has some very useful features.

- First, it doesn't alter the raw data, which is good for data integrity
- It also tracks all of the steps you took and can apply these steps to other data sets
- Finally, If you make a mistake, the changes are easily reversible

Slide 4: But before we get into using OpenRefine, let's look at the data we'll be using in this session. The rodents survey file contains data collected about animals in a field study.

- Each row is an observation of individual animal.
- Each column contains information about these animals, such as
    - the species and sex of the animal
    - the date and location of the observation
- However, these data are messy
    - the data contain misspellings, especially in the species name column
    - There are also extra spaces in the text fields,
    - and columns that contain multiple variables.

Slide 5: Let's get these data into OpenRefine! To get started, you'll need to create a project using a spreadsheet. There are a couple of ways to do this.

Demo 1:

- Then, start the OpenRefine application. The interface will open in your web browser.
- Either download the file (**https://tinyurl.com/jwtqy4w**) and select "This computer" to load the file.
- OR Select Web Addresses(URLs) and paste the link (**https://tinyurl.com/jwtqy4w**) in the box.
- Preview loads (NOT A SAVED PROJECT YET)
- Select file format under "Parse data as" (in this case .csv)
- Rename project
- Click "create project" when the data look how you want them
- Number of rows listed above the table

Slide 6: One of the most powerful features of Open Refine is what's called faceting. Faceting is a great way to check for errors in your data. Creating a text facet will generate a list of all the unique values that have been entered into a column. This allows you to easily identify inconsistencies, such as spelling errors, in your data.

Demo 2:
Let's try faceting the scientificName column

- click the blue triangle next to the column name
- Mouse over facet and Select text facet.
- The list will appear to the left of your data.
- You can edit the values here, and they will be changed in the data.
- Look at Ammospermophilis harrisi. There are 3 very similar facets, spelled slightly differently.
- You can change the spelling to the correct spelling by mousing over the misspelled facet, clicking edit, and correcting the spelling in the pop-up window

Slide 7:
Exercise 1:

1. Using faceting, find out **how many years** are represented in the census.
    - Facet year column
    - Inspect list
    - See there are 26 choices
2. Which years have the **most and least observations**?
    - Sort the facets by count (click on count)
    - Look at first and last entry

Slide 8: OpenRefine also has clustering algorithms that help you find groups of values that might represent the same thing. It's like a more efficient way of doing what we did in the facet example above. I like to think of this as "spell check" rather than editing by hand.

Demo 3:

- Faceting the scientificName column,
- Click "cluster" on the upper right of the facet window.
- A new window will appear with options for clustering methods and keying functions.
- In practice, you can play around with them to see what makes the best cluster for your case.
- For this dataset, the best is the "key collision" method and the "metaphone3" keying algorithm.
- The clusters will appear in the window. If you want to accept the clustering, click the checkbox next to the cluster and make sure the spelling on the right is accurate.
- Click merge selected and recluster until they don't find more relevant clusters

Slide 9: As I said previously, OpenRefine keeps track of everything you do. If you want to see your data cleaning history, go to the left-hand frame where we were working with facets and select the Undo/Redo tab. You can click on each step and revert to newer steps simply by clicking the step. Note that the data on the right changes when you do this.

Slide 10: We can also split columns that contain more than one variable into multiple columns. Let's split the scientific name column into one column for genus, one for species

Demo 4:

- Click the blue arrow to the left of the scientificName column
- Select Edit column
- Then select split into several columns
- Pick a separator. In this case " "
- Choose whether you want to keep the original column. I like to keep it, because I can use it to make sure the split worked as expected the split and can always delete it later.
- Once you hit "ok" the new columns will be added to the spreadsheet. See the new columns with genus and species names
    - Why are there 4 columns? (whitespace before scientific name)
    - Undo split from the undo/redo tab
- Remove whitespace
    - Click on the Blue triangle to the left of the scientificName column heading
    - Then mouse over **Edit Cells**,
    - then **Common Transforms**,
    - then select **Trim Leading and Trailing Whitespace**.
    - Redo split
- Now that we can see that the split worked, remove the speciesName column by going to "Edit column" > "Remove this column"
- You can rename the genus column by going to "Edit column" > "Rename this column" to genus

Slide 11:
Exercise 2

- Try to change the name of the second new column to "species".
    - Already a column named species
- How can you correct the problem you encounter?
    - Create more descriptive names
    - Rename species to "speciesAbr"
    - Rename scientificName2 to species

Slide 12: So far, we've been working with the entire dataset. But what if you only want to look at part of the data? This process is called Filtering. You can do this two ways:

1. If you want to select all records in a specific facet, you can click on the facet

Demo 5:

- Facet the species_abr column
- Click on the facet you're interested in
- Data changes to contain only records in that facet

Slide 13: But what if the data you want don't correspond to a facet? For example, think about unstructured text like the locality column. What would you do if you wanted to find all measurements made in Hawaii? For this you can use the "text filter" option.

Demo 6:

- Select the locality column and click "text filter". A box will pop up on the left-hand frame.
- Type in the text you want to search for, in this case 'hawaii'.
- Now when you look at the locality facets above, they all say Hawaii somewhere (not case sensitive).
- Be careful, because all of these are exact text matches, so you might lose some that have misspellings
- Can use regular expressions

Slide 14:
Exercise 3

- Goal: find all years in the 1980s where measurements were taken
    - Facet on year
    - Create a text filter to get data from 1980s
    - All years from the 1980s have entries (look at the facets)

Slide 15: In addition to filtering the data, OpenRefine also allows you to sort the data as text or a number.

Demo 7: Sort by month (mo)

- Facet on species_abr
- Filter on the AH facet
- Click the blue arrow to the left of "mo"
- Select Sort
- Pick how you want the cell values sorted. Since the mo column contains numbers, we'll do 'numbers'. (Note, the results will be different).
- Then click ok
- Now all of the values in month are in order
- Redo the sort as text. Note that the sort changes

Slide 16: Now that a sort has been applied, OpenRefine gives you more options. For one, you can remove the sort. This function is important, because it returns the data to its original order, even if it was in no particular order to begin with. In programs like excel, your only option would be to hit undo immediately after sorting.

Demo 8: Remove sort

- Return to the sort menu, and click “remove sort”
- Now the data return to their original order.

Slide 17:
Exercise 4- sorting multiple columns

- Sort by year and month. What order are the entries in?
    - Year takes precedence (months sorted within years)
    - Unlike how sorting works in excel
- Sort by year, then month, then day
    - Year takes precendent,
    - months sorted within years
    - Day sorted within months
- What happens when you remove the sort on the second column?
    - Days are sorted within years, months out of order.

Slide 18: By default, OpenRefine imports all data as text. However, it does have special functions for numeric data. To use them, you have to tell OpenRefine that a column contains numbers.

Demo 9: Make the year column into a number.

- Select the blue arrow next to record ID
- Select Edit Cells
- Select Common Transforms
- Select “To Number”
- You can tell it works if the numbers turn green
- Make sure all filters are removed: if not, only the filtered data will be converted

Slide 19:
Exercise 5

- Convert 3 other columns to numbers (include period)
    - Year, month, date and period
- What happens when you try to convert a non-numeric column?
    - Nothing!
    - Different from R, where if it can’t coerce the value to a number, the data all turn into NAs

Slide 20: Now that we have some columns designated as numbers, we can do some really useful things with **Numeric Facets**. To create a numeric facet:

Demo 10:

- Select the blue arrow next to year
- Select Facet
- Select numeric facet
- Now, instead of a list of names as facets, you get the range of numbers with slider bars
- Slide the bar to the range you want to include

Slide 21:
Exercise 6

- In a numeric column, replace a number with text (such as abc) and one with a blank
- Create a **numeric facet** for this column
- How is this different than the numeric facet for "Year"?
    - Checkboxes for non-numeric and blank data
    - Check/uncheck boxes to filter

Slide 22: The last type of facet we're going to talk about is a **Scatterplot Facet**. This facet type allows you to subset the data based on 2 numeric facets at once.

Demo 11:

- Select the blue arrow next to any numeric column
- Select Facet
- Select Scatterplot Facet
- This pulls up the Scatterplot matrix window. This is a grid that will plot each numeric variable against all others and shows a preview. Click on one to select a facet.
- The plot you selected will appear on the left column.
- Export plot opens the plot larger in a new tab
- To subset, drag a box around the points that you want to include (acts as a filter)

Slide 23:
Exercise 7

- Click on the **Scatterplot Matrix** square for **recordID** and **period**
- Facet on species_abr
- Filter on the AB facet.
- Notice the change in the scatterplot. It might be easier to see if you click **export plot** to put it on a new browser tab.
    - Shows all points in gray, points in the filter are black/orange

Slide 24: Now we've done all this work in OpenRefine, but it is stored inside the program. So how do we extract this stuff? First, I'll show you how to save the steps you've done in the Undo/Redo tab.

Demo 12:

- In the Undo/redo tab, click extract.
- Select the steps you want to keep.
    - This generates JSON code that specifies these steps.
    - Uncheck and recheck some boxes to see the code change
- To save the code, copy and paste it into a text editor and save it as a .txt file.

Slide 25: Once you have these steps saved, you can apply them to similar files. So if you collect the same types of data over and over again with the same column headings and the same data cleaning steps needed, you can apply these scripts instead of having to point and click through the whole thing every time.

Demo 13:

- Go into the undo/redo tab
- Click apply
- Past in the contents of the text file
- Click Perform operations
- The data should change to reflect those steps and the steps will be in the undo/redo history.

Slide 26: We can also export the steps and the data together by exporting a project. You've noticed that throughout this process I haven't clicked save at all. This is because OpenRefine is autosaving everything you do as you go. But if you want to get your work off of your computer you need to export the project.

Demo 14:

- Click the Export button at the upper right hand side of the screen
- Click export project.
- The program will automatically download a compressed file that contains all your data and the cleaning steps.

Slide 27: Now that your project has been exported, anyone who has OpenRefine can view it just by importing the project.

Demo 15:

- Click open in the upper right, a new window will open
- Select Import Project on the left hand side of the window
- Open the compressed file created by the export and rename it if you would like
- The cleaned data and history and the data should be loaded.

Slide 28: Now let's talk about exporting your cleaned data. Not everyone is interested in everything you've done with your data. Sometimes, they only need the final product. Thus, you can also export your cleaned data using OpenRefine.

Demo 16:

- Click on the export menu on the top right
- Click on the type of data file you want to export. We suggest .csv or tab separated text.
- The program will download the data in the selected format automatically.

Slide 29:
Thanks for listening. If you need any help with these exercises, don't hesitate to email me at tobin.magle@colostate.edu. You can also visit our data management services website to see what else we do with regard to data management. Also, if you want to see the lessons these were based on, visit the data carpentry website and view the lessons in a bit more detail. Thanks!