Colorado State University Libraries
CSU Libraries
Training and Instruction
Transcription of Data cleaning using OpenRefine, 2/13/2018

Collection: Training and Instruction (10217/195518)
Title: Data cleaning using OpenRefine
Date: 2/13/2018
File Name: FACFLIBR_DaD-OR_TM_20180213.mp4
Date Transcribed: November 2024
Transcription Platform: Konch AI

BEGIN TRANSCRIPTION

[00:00 - 01:49] Tobin Magle: 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. The goal of this session is to clean and enhance data using a powerful cleaning tool called OpenRefine. We need tools like OpenRefine because data are messy, especially if it's entered by hand. It can include things like misspellings, extra spaces, values that don't make sense, and variables combined into one column. Luckily, we have data cleaning software that can help with identifying and correcting errors, making formats consistent, and leaving a paper trail of what you did to the data. OpenRefine has some very useful features. First, it doesn't alter the raw data, which is good for data integrity. It also tracks all the steps you took and can apply these steps to other datasets. Finally, if you do make a mistake, the changes are easily reversible. But before we get into using OpenRefine, let's look at the data we'll be using in this session. The Rodent Survey file contains data collected about animals in a field study. Each row is an observation of an individual animal. Each column contains information about these animals, such as the species and sex of the animal, and 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. 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 different ways to do this that we'll go over in the next demo. Let's start by opening the OpenRefine program by clicking the blue gem on the desktop. See the program is opening here. [pauses] An OpenRefine will automatically open a web browser to do the work in.

[01:49 - 03:36] Tobin Magle: So, you can see this is all local computer. This is a local IP address, and this is the portal that connects the web browser to OpenRefine. So to create a new project, I'm going to go here on the left and click 'Create Project'. And if you have data on your computer, you can use the, this computer tab. And then just click here and navigate to the file that you want to load. In this case it would be on my desktop. But instead for this demo, we're going to use a web address. So I'm going to click down here. And then I'm going to take the web address from the previous slide. And load the data directly from there. So, I'm going to click 'Next'. [pauses] In OpenRefine loads a preview of the data. So, you can see here that the data doesn't look that pretty. The columns actually didn't separate out correctly, and this is because OpenRefine didn't correctly identify the format of the file. So to fix this, I'm going to go down to parse data as and then click on 'CSV/TSV'. You can see now, since we've told it that it's a comma separated value file, the columns separate out nicely here. We can also rename the project and I'm going to call mine, rodents. And at this point we haven't actually created a project yet. This is just a preview. So now, when we click 'Create Project'. The data will be saved in OpenRefine. So, OpenRefine never really displays all the data at once. Currently, it's showing the first ten rows of the data. You can switch the number of rows by clicking, so if you want to see 50 rows, we can click here. Then scroll down to see that we have 50 rows. But an important thing to note is that the total number of rows in the spreadsheet will always be displayed up here above the table.

[03:38 - 05:38] Tobin Magle: One of the most powerful features of OpenRefine 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. To start, let's try faceting the Scientific Name column. To do this, we're going to go to the top of the Scientific Name column, and then click on the blue arrow to the left of the column. Then, we're going to mouse over facet, and then select 'Text facet'. Now, you can see that we have a new panel on the left hand frame. It's labeled as scientificName, the name of the column and the list here is, are all the unique values of some data in the scientificName column. Let's look specifically right now at the Ammospermophilis harrisi entry. You can see that there's three things here that look basically like they probably represent the same organism. There's 435 of these and one each of these. So, these two that are flanking the correct spelling look like misspellings. So, if you see something like this in your data, you can mouse over the misspelling, click 'edit' and then fix the spelling. In this case I'm going to remove the 'i' at the end of the word and click 'apply'. Now, you can see that the misspelling disappeared and got folded into the, this one. What this is doing is actually changing the value of that cell in the scientificName column. So when you're editing fast some of this left panel, you're actually changing the data. So, just for the sake of completeness, you can change this one as well. We'll change this 'i' to a 'u', and

click 'apply'. Now, they're all combining the same facet. Let's take a minute to test out your faceting skills by doing some exercises.

[05:38 - 07:33] Tobin Magle: So, for exercise one, we're going to be looking at the year column. And we want to find out using faceting, how many years are represented in the census? And secondarily, we want to know which years have the most and least observations? So take some time to figure this out on your own. Pause the video, play around in OpenRefine, and then start the video again, when you're ready to know the answers to the questions. All right. Let's look at the answer to this exercise. Using faceting, find out how many years are represented in the census. To do this, we're going to facet the year column by clicking to the arrow to the left of the name of the column, selecting 'Facet', and then selecting 'Text facet'. You can see that we have a lot of years represented here. We can scroll down and see that it doesn't look like there's any typos or duplicates. And then above this listing, we can see that we have 26 choices. So there's 26 years represented in the census. The second question is what years have the most and least observations? So currently, the data are sorted by the name of the facet. So we're going from 1977 up to 2002. But, instead let's sort by count by clicking on count here. And now you can see we're looking at count and descending order. So 1997 has 24,093 entries which is the most. And if we scroll down to the bottom here, 1997 or 1977 has 503 records, which is the least number of records. [pauses] 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 spellcheck rather than editing by hand. [pauses] Now, let's look at how to do this. Use these clustering algorithms in OpenRefine.

[07:34 - 09:24] Tobin Magle: We've already faceted the scientificName column, and that's the first step. And to use the clustering algorithms, we're going to click on the 'Cluster' button right here. This opens a new window that gives you options for the clustering and keying function methods. In practice, you can play around with them to see what makes the best clusters for your data. But for this data set, I know that the best options here are to pick the 'key collision' clustering method. And the 'metaphone3' keying algorithm. So let's change this to metaphone3. These are very complicated algorithms that have a lot of details. But for what we're doing right now, it's not really important to understand all of the details. All right. So we found two new clusters. We can see with this first one Amphispiza bilineata. We probably could have gotten a lot of these, um, by hand. But the fact that there's one where they spell it like 'E' with an 'E' in front, and the choices were sorted alphabetically, you probably wouldn't have been able to catch this by hand as easily. So, we can look at these things. Yep. They all kind of look like the same thing. Um, this one has the most number of records, so it's probably the correct spelling. And that's why this is over here. And we can choose whether or

not we want to merge these into the same thing, because theoretically there could be stuff that looks similar. Um, that is things that you don't actually want to merge into one facet. But in this case, this looks good. And same down here with, um, this one. So, we can click here. And then we can select 'Merge and Re-cluster'. And see if any more clusters appear. Because once we start merging things, then the algorithms find new things in here.

[09:25 - 11:13] Tobin Magle: Okay. So we can see that there is no other clusters found with this method. So we can, we can close. And that's how you use clustering algorithms. 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 on facets, and select the Undo/Redo tab. You can click on each step and revert to newer steps, just simply by clicking on the step that you want to go back to. Note that the data on the right changes when you do this. We can also split columns that contain one or more variable into multiple columns, using OpenRefine. Let's split the scientificName column into one column for genus and one column for species. Now let's split the scientificName column into genus and species columns. To do this, we're going to click to the left of the name like we have been. But instead of going to facet, we're going to go to edit column and then split into several columns. We're going to use a separator. But not a comma in this case. In this case, the scientificName, the genus and species are separated by a space. So we're going to delete the comma and then just put a space in there. We can also pick whether or not we want to remove the original column. I'm going to uncheck this box because I like to be able to check to see if the split worked correctly, to have the scientific name column and the other two columns right next to each other. And we can always delete it later. So when I hit 'okay', [pauses] what I would have expected here is to get two columns, one with scientificName 1 and one with scientificName 2. But instead we have four columns, all numbered scientificName 1 through 4.

[11:14 - 13:11] Tobin Magle: Why do you think this is? Well, it's probably because there is some whitespace before the species or the scientific name in these two columns. So to fix this error, I'm actually going to go back into the Undo/Redo tab. And go back one step. Now, we can see we are back to just one column for scientificName. And before we move on, we're going to talk about how to remove whitespace. OpenRefine contains built in functions that are commonly used in data cleaning, such as removing whitespace to do this, we're going to click on the arrow to the left of scientificName like we have been. And we're going to go to 'Edit cells'. And then we're going to go to 'Common transforms'. And then we're going to click on 'Trim leading and trailing whitespace'. We don't really see a lot of changes in the data, but that's because we can't see whitespace to begin with. But to see if this actually worked, we're going to do the Text facet or the split again, sorry. So click on the arrow. Go to edit column, split into several columns. We're going to do it by a separator, which is a

space, and we're not going to remove the original column, right? So, when we hit okay, as expected, we get two columns, one called scientificName 1 and scientificName 2. And these look like they match up pretty well with the original scientificName column. So, I'm going to go here and click the arrow. Go to edit column, [pauses] and then scroll down here to remove this column. Now the scientificName column is gone to get rid of repeated data. Now we're going to rename this column to genus. So I'm going to click on the arrow. Go to edit column. Go down here to rename this column and I'm just going to type in genus.

[13:17 - 15:11] Tobin Magle: Let's apply what we've learned with another exercise. So for Exercise 2, you should try to change the name of the second new column to the species. And I'm going to warn you that you're going to encounter a problem. Figure out the cause of the problem, and then figure out how to correct it. So, pause the video to play around with OpenRefine now, and then restart the video again when you're ready for the answer. Yes. Once again I'm going to click on the arrow, highlight edit column, then go down to rename this column. And I'm just going to type in species here. You can click okay, but it's throwing me an error and it's saying that another column is already named species. So I'm going to click okay. And I'm going to look over here to the left of genus. And there is in fact a column called species. And it just has a two letter abbreviation of the species, the genus and species name. I would solve this problem actually by renaming both columns. Since this is an abbreviation, I'm going to go to edit column. Rename this column. Do underscore a-b-r for abbreviation. And then I'm going to go back to scientificName 2, edit column. Rename this column, species. So now they all have unique names that are a little bit better describe what's actually in the column. So far, we've been looking at the entire data set. But what if we only want to look at part of the data? This process is called filtering. You can do this in two ways. First, if you want to select all records in a specific facet, you can click on the facet. For this demo, let's look at the species abbreviation column. So, let's click on the arrow to the left.

[15:11 - 17:12] Tobin Magle: Highlight 'Facet' and then click 'Text facet'. You go down here and see that the species abbreviation column gets split into 48 facets. But, let's say we only wanted to look at the ones that are tagged as Amphisis- Amphispiza bilineata. In this case, we're going to click on AB. Note that, before we click, note that it's 35,000 columns about up here, but after we click it, it goes down into 303 matching rows. What if the subset of data that you want doesn't correspond to a facet? For example, think about unstructured task like the locality column. What would you do if you wanted to find all the measurements made in Hawaii? For this, you can use the text filter option. Let's demonstrate by applying a text filter to the locality column. So we click on the blue arrow. And select 'Text filter'. This brings up a search box in the left hand panel where we've been dealing with text facets. And so let's search for entries that contain the word 'hawaii'. You can see that we have

15 rows of the data that have 'hawaii' somewhere in the locality column. You can see here that I spelled Hawaii with a lowercase 'h' on the left, but it's still picking up Hawaii with an uppercase 'H' on the right. So we've learned that by default this search is not case sensitive, but we can make its case sensitive by clicking this box. And now nothing matches because there's no, no text that says hawaii with a lowercase 'h'. You can also make this search more flexible using regular expressions. This is a very complicated topic, so we're not going to get into it here. But just know that, if you are familiar with regular expressions, you can use them in OpenRefine. All right. Let's test out your filtering skills with Exercise 3.

[17:12 - 19:06] Tobin Magle: The goal of this exercise is to find all years in the 1980s where measurements were taken. So to do this, you're going to pass it on here and then create a text filter to get all the facets that include dates in the 1980s. When you were, so pause the video, when you're ready to continue and hear the answers. They'll continue after, right after this. All right. So to complete Exercise 3, we're going to be working with the year column. So we're going to click here. We're going to facet it and we're going to do a text facet. So you can see we have the 26 facets that we did, when we did this earlier. And now we're going to apply a text filter. So, I'm going to click the arrow, go to 'Text filter'. And I'm going to type in '198'. And now we can see after it's updated, we've got 1980, '81, '82, '83, '84, '85, '86, '87, '88 and '89. So we know now that for every year during the 1980s, measurements were taken. In addition to filtering the data, OpenRefine also allows you to sort the data as text or number. Let's look at sorting the month column. So it is set up for this demo. You can see that I have faceted on species abbreviation. And then I filtered to only include the Ammospermophilus harrisi data by clicking on 'AH'. Okay, so let's sort the month column. I'm going to click to the left of the month column and then go down to sort. And since we know that months here are being represented by numbers, I'm going to switch to a numeric sort and hit 'OK'. So we scroll down here, we can see that January is up first and then we switch into February eventually down here. Okay. Let's go back into the Sort option.

[19:08 - 20:44] Tobin Magle: And change it to text instead of numbers, just to illustrate how OpenRefine handles these two things differently. So when I click 'OK', it's going to resort the data. We're back up at the top now. When we scroll down, instead of going from 1 to 2, we go from 1 to 10. And this is because when OpenRefine is sorting numbers, it's going to put them in numerical order. But when it's sorting text, it's basically trying to alphabetize the numbers. Now that a sort has been applied, OpenRefine gives you some 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 to get back to the original order. So to undo the sort on the month column, we're going to

click to the left of the month column. Now we can see instead of just clicking on sort, we have some options to pick from. We did this before when we went back into the sort to change it from numeric to text, but instead we're just going to go down here to remove sort. And you can see the data returned to its original order. Now let's practice some of your sorting skills with Exercise 4. This exercise is about sorting sorting multiple columns. So, we're going to start by sorting by year then month. And then look at the data and how it's sorted and see what order it goes. In the next one, you're going to sort by month, day and year in any order and then remove the sort on the second one to see what happens. So [unintelligible] now to play around with the sorting. And then the, the answers will be after the pause.

[20:46 - 22:41] Tobin Magle: All right. We're going to start sorting by year and then sort by month. So I'm going to go into year, sort as a number and click 'OK'. Now, we have the oldest years at the top. Now, I'm going to go by month and click month -> sort -> numbers, click 'OK'. And now we can see that the years didn't change order at all. Were still going from the oldest, the newest. And then within each year we have the month sorted there. So if we look specifically at 1985, we have January, February, April, et cetera. So, this is unlike how sorting works in programs like Excel, where the second thing that you saw takes precedence. In OpenRefine, the first thing that you sort takes precedence. All right. So now let's get the year involved here. So I'm going to go to- sorry the day -> sort -> numbers, and click 'OK'. And then our assignment was to remove the sort in the second thing that we did. So first we did year, then we did month, then we did day. So when I remove the sort, on month, [pauses] we still have year in order, primarily because it was the first thing that we sorted, and then the months go back to the order that corresponds to the day being sorted by day. So within 1985, no matter what month it's in, we're going in order. So 7, 8, 15, 16 and then the months are out of order. So, this is just to illustrate and to give you some experience playing around with how things are sorted, because it's not intuitive if you've used other programs like Excel. 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.

[22:42 - 25:19] Tobin Magle: To demonstrate this, let's turn the recordID column into a number. So we're going to click to the left of recordID. We're going to go to Edit Cells -> Common transforms, where we went for the leading and trailing whitespace. We're going to come down here and click 'To number'. So, you can see that OpenRefine turned them into numbers by turning these green. But actually, let's remove this filter. For species abbreviation, I'm just going to click on 'AH' again. And now we can see that we actually only turned the record IDs that were included in this 'AH' set, green. So, I'm going to actually redo this. So recordID -> Edit cells -> Common transforms -> To number. And now all of them are green. Now let's practice this with Exercise 5. So first convert three other

columns into numbers. But make sure to include the period column. Now try to convert non-numeric column into numeric and see what happens. Pause the video now and we'll have the answers up next. All right. Let's convert some numeric columns into numbers. Let's see. So. Edit cells, come in, transforms, To number. I'll call him, oops. Keep wanting to edit the column. To number. Well, dear, for good measure. And then finally period. Okay, so that all worked well. The all turned green. So what happens if we try to do this with a column that doesn't actually have any numbers. We'll do a species abbreviation. So I'm going to click here edit column. Oops. Edit cells. Common transforms. To number. And you can see what happened here is that nothing changed. So, this is actually different from other programming languages. Like if I did wrote the code to convert this column in R to a number, it would obliterate the data. It would turn all of the things that it couldn't coerce into being a number into an A, which is basically the, the signal for no data in R.

[25:21 - 27:19] Tobin Magle: Now that we have some columns designated as numbers, we can do some really useful things with numeric facet. To create a numeric facet, we're going to click the, the blue arrow next to year. Go to Facet. And then instead of selecting 'Text facet' we're going to select 'Numeric facet'. Now we can see we have a new window here, on the left underneath the, the text facet for species abbreviation. And instead of having a list of values, it actually has a range of values. So, if you wanted to only take the newest data, I could take the slider bar and slide it over. And that just subset the data from 1995 to the most recent datasets. Let's explore some more advanced faceting with numeric facets in Exercise 6. So first pick your numeric column and replace any of the numbers with texts such as abc, for example, and then another number with just a blank space. Create a numeric facet for this column that you've edited. How is this facet different than the numeric facet for year? Now, we've done all this work in OpenRefine, but it's stored inside the program. So how do we extract this stuff? Well, first I'll show you how to save the steps that you've done in the Undo/Redo tab. So to save the steps that we've done, we can go into the Undo/Redo tab and you can see all the steps we took here in order. And click 'Extract'. Now, you can see that we have all the steps in JSON format here on the right. Now you can pick which steps you want to include and exclude. And to illustrate this, let's just unselect them all and we have no script. And then click 'Select All' and they come back. We can also uncheck individual boxes. So I'm going to click a few here. And you can see the more we click, the shorter the script gets.

[27:19 - 29:26] Tobin Magle: So, I'm going to select all again because I want to keep them all. I'm going to put my cursor in here. Select all of the text, and copy, and then go over into a text editor. Paste, and then save. 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 headers and the same sort of data and cleaning steps needed, you can apply these scripts instead of having to point and

click through the entire thing every time. To illustrate this, I just created a new project with the unedited Portal rodents file. So to apply the steps that we have already had, I can go to the Undo/Redo tab, and then click 'Apply', and it gives you a text box where you can apply the JSON history. So, I'm going to go back to my text document. Select all the text. Copy. And then paste it in here. And then click 'Perform Operations'. And now you can see, we no longer have the the speciesName column. We have it split into genus and species. We've changed some other column names and converted these to numeric, all with just pasting one chunk of text into OpenRefine. We can also export the steps and the data together by exporting a project. You've noticed that throughout this process, I haven't actually hit save at all. And that's because OpenRefine is auto saving everything you do as you go. So, but if you want to get your work off of your computer, you need to export the project. Now that your project has been exported, now anyone with OpenRefine can view it, just by importing the project. To import a project into OpenRefine, we're going to go to the open button on the upper right, and it brings us back to this page where we started.

[29:26 - 31:20] Tobin Magle: These are projects that I've already been working on. So to import one I'm going to click 'Import Project'. And then I'm going to navigate to where the the compressed file that we downloaded earlier was on the desktop. I'm going to select this tar file and click choose. I can rename the project. I'm going to call it 'project2' just to not get it confused and then click import. Now, we can see all the data is here. Cleaned up. And then if we go to the Undo/Redo tab, we have all the steps that we did before. Now, let's talk about exporting your clean data. Not everybody is interested in everything you've done with your data, sometimes they only need the final product. Thus, you can export your clean data using OpenRefine. To export our final data set, we can go up here on the upper right and click the export button. You can see we can export it in a variety of formats like tab separated, comma separated, Excel, etc. For fun, let's export it as an Excel file. I'm going to click Excel. And then an auto starts a download after it makes the file. So, I'm going to open up my downloads folder. We see we have an Excel file with the project name attached to it. So, I'm going to double click it to open it in Excel. You can see here, we have a clean data set in Excel format. Thanks for listening. If you need any help with these exercises, please don't hesitate to email me at tobin.magle@colostate.edu. You can also visit our Data Management Services website, to see what other things we do with regard to data management. Also, if you want to see the lessons that these were based on, please visit the Data Carpentry site and you can follow along the lessons in a little bit more detail. Thanks.

END TRANSCRIPTION