Please stand by for realtime captions.

My name is Jamie Hayes I am your host for today's webinar. We also have hash Lee on tech support so if you're having any technical technical difficulties plea email Ashley. Are presented as Amy Solis and she is the librarian for social sciences data. Government documents and public administration at the University of Southern California USC. In Los Angeles California. She also serves as USC's official representative to the inter-Consortium for political and social research or IC PFR. She is responsible for overseeing data services in the USC libraries and works closely with university faculty, subject librarians and the referenced team to build and promote numeric data collections in support of data literacy, research and curriculum. In the social sciences. Before we get started I will walk you through a few housekeeping reminders. If you have any questions you would like to ask the presenter or if you have any technical issues please feel free to use the chat box. It is located in the bottom right-hand corner of your screen and I will keep track of all the questions that come in and at the end of the presentation Amy will respond to each of them. We are recording today's session and will email you a link of the recording and slides to everyone who registered for this webinar. We are also going to be sending you a certificate of participation using the email you used to register for today's webinar. If anyone needs additional certificates because multiple people are watching the webinar with you please email STL P outreach@GPO.gov and include the title of today's webinar along with the names and email addresses of those needing certificates. If you need to zoom in on the slides being shown by the presenter you can click on the full-screen button in the bottom left side of your screen. To exit the full-screen mode mouse over the blue bar at the top of the screen until it expands. Then click on the blue return button to get back to this you. At the end of the session we will be sharing a webinar satisfaction survey with you and we will let you know when the survey is available and we would appreciate it if you fill that out. Finally Amy will be screen sharing her presentation today which means once Amy starts talking you will no longer see the chat box in the lower right side of the screen. If you want to ask a question or if you just want to watch the chat traffic as Amy is presenting once screen sharing begins just mouse over the blue bar at the top and when the menu drops down download Chad to enable the chat box. With that I am done and I will be covered Amy. You're all set Amy.

Hello everyone thank you so much for attending my webinar. I'm going to start out sharing my screen. Then I need to, uh, put up my slide all right, so everyone should be looking at my slides, thank you.'s thank you so much. My presentation is [indiscernible] with working with metadata and open refine. It was a workshop presented at [indiscernible] with each of the international Association for social science information services and technology. By Leanne Trembley and Colleen short. From concordia University in Tampa. Basically is a lesson plan and I use my own data set from a documents project that I worked on and and for document librarians I think. This is an introduction to open refine you do not any need any prior experience or agenda for today we are going to talk a little bit about my background what is open refine the open refine set up separate demonstration and hands-on practice and then additional helpful resources. So our learning objectives by the end of this webinar you will be able to search, sort and filter data in a variety of ways. Restructure and manipulate a data set and then perform basic data cleanup. In terms of myself I am a fairly new government documents library and in a year ago I did my first meeting project and it was pretty difficult and I used open refine to take spreadsheets from my, uh, catalog and then try to get up something presentable to give to my state library for approval

of my reading project. I am hoping that open refine also helps you in your reading project. So everything that I am going to show you, the slides, the handouts, the data sets, everything is located in [indiscernible] URL.com Dosh [indiscernible] open refine. Open refine is locally on your computer so even though we are going to use a web browser a copy of all your data files are saved locally to your computer and this is great, this is actually data management 101. You always want to keep your original data for this and just make a copy. [indiscernible] to open that freeze that is great. A little bit about messy and clean data. A lot of people heard the term messy data so what is it, what is clean data collection so clean data sets require very few transformations are changes. And messy data by contrast requires a lot of changes to the data to reorganize it and make it more consistent. An example of this is is somebody, uh, a visitor walks into an office every day, every day checks in to a guest blog for example he might put his name John Smith, J Smith, JL H Smith, JOM Smith. So different variations of their name. This creates a lot of messy data. So column a on the screen shows an example of messy data. By contrast clean data is uniform. That's probably the main difference, column B you can see that is an example of clean data so the last name first, the first name second and both names are separated by a comma. So basically it is a lot more consistent. So, messy data has a lot of typos and you know a lot of data entry is done by humans and as humans we are not perfect and we always make mistakes or put things, enter in things differently so by definition whenever there is home human entry it usually results in messy date of. Right? What is open refine. We are going to use the open source tool we are working with messy dated to clean it up. Was formerly called Google refine. Is a powerful tool for working with messy data, to clean up, transform it, since 2012 it is no longer supported by Google. It is not entirely run by volunteers. So why open refine. You have Excel and then there is open [indiscernible] so it's a great program but sometimes it takes a lot of time to clean your data especially large data spreads in Excel. Sore great place to start and for beginners working with data's I use and [indiscernible] together. On the other hand if the program language is is supported by the our foundation and it's also another [indiscernible] to clean data. The learning curve for this, somebody like myself who does not have programming skills, I have not been able to start using [indiscernible]. So open refine is a nice compromise because open source but is also known as Excel it's a great alternative to clean data for non-programmers. A typical refine and also XML files as well. So URL.com/[indiscernible] 2019 open refine so, everything again is going to be in that link so we are going to start with our demonstration and hands-on practice. I will go through it and rehearse it a few times so we'll have actually time at the end for a bunch of questions. So do not worry so much about asking questions throughout my demo. So this activity has no prior knowledge to open refine. We will be importing a spreadsheet, in Excel spreadsheet into open refine and exploring it. The goal of this activity is a simple data sets to introduce you to open refine. And the user interface and it has some basic tasks you can accomplish. In this in a met we will first review our Excel data set, loaded onto open refine and then perform a really basic data [indiscernible] so you can get familiar with interface and we will do some clustering which is, uh, another way of saying combining similar values together. We are going to convert a column from text to numeric use. We are going to use open refine just sort and [indiscernible] are data and also restructure the data set by removing columns and rolls. The really cool redo feature and at the end we will export the data and refining them back into XL. So hopefully everyone has loaded all the documents from the URL.com Dashe [indiscernible] 2019. Folder onto your computer. I am going to show you how that looks. Right now. Okay so this is how the folder looks, there is an open refine handout here. The slides I just showed you are here. An indoor golf dock Pro

project Excel project this is the file we are going to use today. Then we have a Leedy project activities instruction. So clearly every single thing I am going to say I wrote down the steps in this activity instruction so if you want to read I will recommend opening the instructions so I would recommend downloading this in town folder off your desktop or anywhere accessible on your computer. Hopefully it is not too crowded in here. I have only the folder on my desktop so I am going to open the golf dock cleaning project. So it looks like this. So this is a data set from our government document from leading project from the University of Southern California library. It contains metadata information such as [indiscernible] number OCLC number, title, publication date, etc. for items considered for discard for a leading project I did last year. Notice the following, the data file has columns or roles here and it has formatting applied on the top here you can see that we have descriptive information that is not part of art data [indiscernible]. The permanent call number here has some so we will take a look at, a closer look at this. In open refine. So this is what we are working with. So I just wanted to show you that, I am just going to close this Excel file right now. So to open refine depending on if you have a Mac or a PC it's a bit different. So, I gave instructions on how to open, open refine I have a map for example so I have you can have a diamond or you can go to your finder's folder and applications. And then where you have all your other programs like Microsoft Word it should be a diamond where it says open refine. Do a right click to open. Is going to actually open in the browser. And right now I am using chrome. I would recommend using chrome I think it works pretty well. If you have trouble open up Safari or Firefox Firefox, try it, something with Max there is this with the issue so in the instructions there is a special URL code that you can copy and it looks like this onto your browser. And to open it when I tell my students the hardest part of open refine is actually just downloading it and opening it on your computer once you have that actually using [indiscernible] it is easier. So your biggest barrier is trying to download it and open it.'s all right so now let's get started with open refine. If you are following along in the instruction sheet we are in step number four. Okay so now wait, let's see. Actually we are in step number three. So make sure on the left side that the correct project is selected. Then we are going to click on choose file and then you want to make sure that on your desktop you have your [indiscernible] of the library folder. Then you are going to select God.leaving leading project. Then click open. And then we are going to click on next. And it's uploading a copy of your Excel sheet. Your original Excel sheet will be the same. Is just a different copy. All right we are in step number four. So you are now viewing the data set and preview mode. You can see what the data will look like when it is loaded. Make changes to what data open refine will load. Notice on the top of the descriptive text it is showing in the preview and is messing up [indiscernible] to identifying the setting. You can see on the topper column headings are just numbered if you actually want to [indiscernible]. We are going to click on the bottom where it says ignore first. And we want to ignore the first [indiscernible] lines of our spreadsheet. It does not automatically update, you will click on update preview on the right-hand side. All right so now you see where we have our correct headings now on her spreadsheet. Then on the top you can see where it says project name, go back to project where you can change this and it will leave it as it is. Then I'm going to click create project. It's uploading it, all right, so this is how it looks like the data when it is loaded. Again this is just a copy in your original file is still on your computer. So all right so now we are going to perform some basic data cleanup to get familiar with the open refine interface. So on the top you can see that is is 313 rose that is how many rows are in the spreadsheet. You can see we can only see 10 so you can actually lick on 50 at the top. Now we are going to be able to see 50 rose and if you want to see the next 50 you can click

on next for the last 50. You can click on that right-hand side. So the first thing we are going to do is remove a column. So take a look at this column, actually a labeled column. It does not have any information so we want to remove it. So your if you're following along in the instruction seat everything that is in bold it is her action to what we are actually going to do. If you are reading along. So we are on step number seven. If you click here on the downward menu, it is going to look like up, I downward pointing arrow. This is basically our bet bread-and-butter of open refine. Everything you do is going to require clicking this pulldown menu from the column. Then you're going to click on edit column and then remove this column. Yea, so this is the first thing we have done. I want you to notice that it tells you what we did, remove the column. It gives you an opportunity to undo it as well. This only last for a few seconds so at the end I will show you how to go back if you make a big mistake and you really want to go back in time. All right so let's do the same thing with column number two. It does not seem like we have anything there. So click on the pulldown menu for column one or two. Then click on edit menu and then remove this column. Great, so now we have removed to column so that is the first thing we have done. So now let's take a look at the permanent column number which is right here. And you can see that we still have these invisible spaces here and this is pretty common when you have really messy spreadsheets where you have all these weird extra spaces that you do not want so that is where you want to get rid of them. If you ever use statistical software they do not like these like weird spaces either. So let's take a look and see if they are really there if you then hover over it now and click edit and actually there are these weird extras bases right. Then you can hit click counsel. So now is get rid of them. So from the pulldown menu from the permanent [indiscernible] you can click the pulldown menu and then click edit style. Then common transformation and then trends leading and trailing whitespace. On the top it actually says contrast form [indiscernible] which is pretty great. This is awesome, this will save you a lot of time especially if you compare your experience using Excel where you literally have to delete a bunch of things and go into save you a lot of time. So now we are going to do some clustering which is when you clean your data by combining similar values. So we are going to try the clustering function on the column publication. So you notice for this column is a bunch of years. That have all these extra characters on them. I want to get rid of all that. So I will click on the pulldown menu for publication date and then I am going to click on facet and then text facet in your facet window should appear here on the left-hand side. And notice that I did not choose it's because open refine has not recognized the numbers in this column it. It's because they have all these extra characters here like?'s and brackets. So we need to edit and clean all of this before we tell open refine that these are actually numbers. You notice that some of these look very similar [indiscernible]. It's [indiscernible] this is 1981 without a?. We want to combine all the similar values together. And make them more consistent. All right, so we were click on cluster here. And now basically there are and they roughly go from strict unforgiving. It says this feature helps you find groups of different cells that might have alternative representations of the same thing. In the lowercase so it helps you find [indiscernible]. So it says that we have 31 rose that all have 1987 but is different characters in them. So let's clean this up. So first one you want to verify that they all meet the same thing in the values and cluster so it's all 1987. You're going to click on merge, the merge button. In your new cell value column you are going to have the year with but without any of the extra extra brackets. This is going to be done and that all the values mean the same thing. Then you're going to click on emerge buttoned and you will have all the extra characters. As we go along and do this together you can see that for every value is telling you how many in this case there are 24 rose that have maybe 80 on brackets. So we are going to clean this up

together and then go down and this is probably one of the best features of open refine because I have the find and replace feature in Excel and that always takes so long so this is a good alternative. Just really cleaning up your spreadsheet especially if you're going to hand it over to somebody else so when you go down and clean all the years went to verify they actually doing the same thing and delete any of the extra characters and make brackets in such. We are going to go down the line all right, so we are almost done. It looks like we cleaned everything up. Once you are done you are going to click on in the bottom merge selected and re-cluster. So you notice that merge and at the top in yellow it says we actually changed 286 rows of cells right now. We did a pretty fast in a few minutes. So, uh, this is a great way to clean it up. So actually our spreadsheet is not that messy because we have not found any more clusters. Normally if you work with other spreadsheets is going to give you more options and you can play around with the other methods. You're going to find more similar values for you. This is a great thing to open and to use an open refine. Now we are going to close this window. Click on close. If you're following along in the activities instruction sheet we are in step number 14. So in the facet window on the left-hand side you notice that not only [indiscernible] were corrected for this column after we clustered. We can still [indiscernible] additional characters so you can hover your cursor over here in the facet window and you can click edit. Then delete your bracket and then click apply thank you. Three cells, it tells you on the top there. To clean this up we are going to click on edit and then delete the bracket and then click [indiscernible]. [indiscernible] 1978 also so click edit and delete that period. Let's make sure that there is no other characters here. There is a bracket in 2000, hover over it and then delete that bracket. Then click apply. This is another way you can clean up your, uh, your values as well. You notice again over here in gray it tells you how many rows are associated with the value. So at the end of this here you notice that there is a value labeled [indiscernible]. This means there is no publication year value for the cells. So they are available you can hover over, your cursor over the blank spots and then click on edit. You can give it a better description, right? Say something like not available. Then click apply. Or you can say something like oh you need [indiscernible]. Something like that. So that somebody else clicks on the publication year. So you can add whatever descriptive text. You can see in our publication date column it says not available. All right, so now we want to ask out of this window, you can click here on X. All right. All right so now we are going to convert a column from text to numeric value. So what do you do when open refine incorrectly values a text when it is actually a number. So in this case for publication column we need to sort by numbers but open refine things that are just text because we have all these extra values there, characters there before, right? So we need to tell refine that these actually numbers. So if you click on the publication date the pulldown menu. Then you are going to click on edit cells. Then common transformations and then to numbers. Let me do it again. XL, common transformation, then to numbers. And now open refine has identified these values as numbers and we know that because they are green now. That is awesome. So now we can do a lot more things now that open refine has identified them as [indiscernible]. So now we are going to sort filter and [indiscernible] the data. So the rose of the data are initially loaded in the order they appear in the original data file. You can change how they appear in sort by Texter number. So from the publication column we can sort by numbers, right? So if we click on the pulldown menu in the publication date column and then click on sort a new window should appear and then click on number and we are going to click on largest first. And click okay. So now we know that the most recent publication date in the spreadsheet is from 2002 which is great. Now we know okay to discard something we have at least five years in their collection. And I've had this way longer than five years. So that cancels

that requirement, right? So this is pretty cool to do if you have a huge data set, right? A lot of things, so now we are in step number 18. So filtering allows us to search for certain information that is in our data set. So let's say we want to display rose from the B2C library which is one of the libraries that here at USC. So we are going to click on the library name column and scroll over to the right, see right here. So library names, click on the pulldown menu and then click on text filter. Your text filter should appear on the left-hand side and in here we are going to type in [indiscernible]. So you notice that we went from 313 [indiscernible] to only ¬25. So these are the rose that are from the QC library and you can see we were only looking at the be QC library. This is a really cool if you have a spreadsheet that is 1000 rows and you only want and [indiscernible] of them. So let's see only some of the information. One thing I always like to point out at this point is that if you export, click here on export you're only going to export these 25 rose. Which is something you might want to do if you want to parse out a huge data set, right? We can have text filters on more than one column at a time. So from the publication date right here we can click the pulldown menu and click on text filter and then we have another text filter for publication date, let's say I want to see the, uh, items that are from 1990. I could type in 1990 and then you can see that with the two filters combined now it's showing us for government documents that were published in 1990 that are located in [indiscernible] library. This is really good to really filter out a huge data set and create other subsets, right? So we are done with filtering, we want to remove the filter so we are going to click on the X here on the left-hand side and we are back to our 313 rose. We are in step number 21. Now Lex Explorer and [indiscernible] way in which the data sorted. It lets you select which data to view as well is provide ways to edit the data. So from the library name column go back here to the library name column you're going to click pulldown menu, click on facet and then click on text facet and we have text facet here on the left-hand side so, it shows you how many total values there are in this column and how many rows you need. It allows you to sort the [indiscernible] by name or account. If I click on account it saying our repository library is our storage facility. Has the most it has 266 rose which is not surprising, we have most of our docs in there so. If we click on the [indiscernible] library we are selecting only the items from the B2C library and we are back to our 25 rose just like we did the filtering. But this time we can do more, we can select a different value so I can hover my cursor over grand repository and also click include. Now we have the [indiscernible] library and the repository. So this is a great way to really your huge data set into different categories. So you notice that the included values are on this red orange color now. That is how we know they are selected. You can exclude one of them or exclude both of them so let's say I want to exclude and [indiscernible] library, I will hover over the be QC library and then click on exclude. All right so now only the grand depository rows are showing. Which is 200 and [indiscernible]. Matching rose. So just like filter you can combine multiple facets at the same time as well. So let's say we also want to do, facet four resource type in the graphic detail. So our resource type in the graphic detail column is right here and I'm going to click on the pulldown menu and then click on facet and then text facet and so now you can see what reverse types are in the repository here on the left-hand side. You have books, migrants [indiscernible] other materials. Let's say I'm only interested in books. You can hover your cursor over books and include this orange color and now we are only seen the 257 matching rose of physical books that are located in the depository library. This is again a great way to say like okay I'm only going to work with books from this library and like if your director gives you a huge spreadsheet, and you can figure out what you need to do, so we are going to X out of this and we are back to our 313 rose. So now we are going to restructure the data set by removing columns and rows. And work

with our undo and rollback changes under reduced feature all right so let's say we are unhappy with that [indiscernible]. So data we cannot do much with. So we can remove whole columns which is exactly what we did on the beginning of her activities, right so let's do that first just to review then I will let say I want to get rid of this [indiscernible] ID column. So I will click on the pulldown menu and then click on the [indiscernible] column. Cancel, that's not what I want to do, pulldown menu then edit column and then [indiscernible] column. There you go. So we are for the first time we are going to use this really cool all feature which is the first column of your spreadsheet here. We are going to click on the pulldown menu for all and then click on edit columns. And then reorder and [indiscernible] columns. Now I new window should open in here you can drag columns that you want to delete from the right to the left. Let's say we want to we want to delete material types physical item and then we want to delete network number so moving it to the right is [indiscernible] basically. Material types [indiscernible] physical items that is the only one to delete. Also we can reorder [indiscernible]. Let say we want the title to be on the top before permanent call number. So moving at the top is the same thing is moving it to the left in your spreadsheet. If you're going to move to the title with all the way the top. Then I want my permanent dates to be after my permanent call man number. My permanent call member. Again moving it means moving it to the left. All right once you have your columns delete you can click here on okay so now you see so instead of working with columns we are going to work with Rose, so another feature of open refine is the star rose in its [indiscernible] by this way. So this is very similar if you use Outlook you can plug your email, so an easy way to flag rose to just click on the flagpole next year roll. So let's just flag a few of these rose here. And then everyone flag euros now we are in step 27. So we can also facet are data sets to show rose and then automatically flag was rose. For example to see how many rows have a blank value from a complete column let's do that for the resource type of the graphic detail. Then you are going to click on the pulldown menu and then click on facet. And customize facet. Then facet by blanks. So now you have your facet on the left-hand side. You are going to click on true, we are going to actually see that there is a column and not only this column the resource place there is no information here so we want to actually delete this. This is a pretty cool feature actually because if you have a really big data set you also can't really see if you are missing information. This is the thing to do for every column I think. You have a Biggs spreadsheet. All right so now we want to delete this row. So we are going to click on the all column, click on the pulldown menu, click on edit row and remove all matching rose. So we actually just remove that row that had no information. So that was great. On the left-hand side you can click on the exit, on the extra exit out of here. So now we only have 312 rose. All right so we are in step 28. So now let's find the rose that are in the Journal. Is go to our materials type here. Click on materials type and then click on the pulldown button and then click on facet and then text facet and you can see that it is giving our material site for journals, we want to look at and now we have a subset of we elicit we are not really interested in working this. So we want to flag them. So from are all column here we are going to click on edit row and flag these rose. So now we flagged all our rose that were identified as Journal inner type. So let's reset so we can click on the reset all button here. It resets to the 312 row. You can also X out of this. Window as well. So we are in step 30. Later let's say you want to decide, you want to remove anything, that you are not sure of. Okay I actually do want to delete everything that I flagged. So farmer pulldown menu of all click on the pulldown menu and then click on facet and then [indiscernible] by flag. And on the left-hand side you have your window and you are going to select true. So now all the romance that you have flagged are showing. Let say okay, I am ready to delete all these row's and then you are

going to click, then click edit row Max and then remove all matching rose. Depending on how many rows you flagged you might have different numbers here. Then you want to X out of your window pack. Now for example I have 286 rose. Do not worry about so we are in step number 32 so what happens if you do a few things and then you wish you could take some of it back. So now we are going to work with open refine undo and [indiscernible] feature. So if you click on the undo tab which is on the left-hand side here next to the facet filter here on the left, click on undo. You will see the number of steps that are like anything we did in this data set. But the great way to keep track of what you have done, this is pretty cool if you are working with multiple people and you say I wonder what he or she did you can see exactly every step. You can also rollback your changes from a previous version by clicking on the last step. Everything after that then rolls back and is deleted. So you can actually go back and forth. You can take it look at a data set in a particular point. So is pretty great. So for example look on the item that says read order columns. Number 12. Reorder columns. You will see that the steps that are grayed out. Greens that has not happened yet. So for this example those flagged rose have now not been deleted. You should see them here in the data set. Oh my gosh I didn't mean to delete that. To the previous step. Start making new changes and transformations. It's basically all the subsequent steps will be deleted permanently. So let's do that. Let's go ahead and start starring some row's. You can click on start here and you notice actually okay everything you done was deleted and now new steps of been added. So you should see that the steps now have been deleted and now you have your new starring row's here. So if we have had a similarly structured data set for example, let's say you take a survey every month, right? You have the exact same questions, the same data and you have to clean it up the same way. You can actually perform perform with her data set with a whole new data set. By clicking on the extract button here. So it is showing you every, you will see the code in the window that describes every single step. So for example you have a similar data set, every day, every month for example. You can copy the code and then click on close and click here, upload your new, uh, data set and click apply. Then paste your code onto there. Then it will do the exact same steps for that new data set. Which is pretty cool. So now we can just click counsel. All right and then the last thing we are doing is exporting our, our files. So you can click here on export, you can export to Excel, to CSV and then open by new Excel worksheet. This is again a copy. This is just a copy of the Excel spreadsheet the original what is intact. That is exactly what you want to do. Let say you are finished with this data set but you want to do another one. You just need to click here on the open refine icon in the upper left-hand corner. It takes you back to where we started from and then on the left-hand side you can click on open projects and see all the refined [indiscernible]. So all your projects, everything always stays. Okay. So we are finished with our activity. I am going to go back to my spreadsheet, to my slides I mean. So everything I have done I also have screenshots of all of the main things we have done. So this is for your, if you are going to do this again or if you missed something I gave all the steps to do it. Then I forgot to show you how to close open refine you have to hit command queue at the same time and it closes, uh, the software. Then you wait until there is a message that shuts down. So this is a very introductory, uh, webinar. You can do a lot more with this. So here are some great resources to keep on going with open refine. I suggest doing the software carpentry, the open refine workshop they have all their lessons online. They have [indiscernible] now it's linked in learning. They also have really great, uh, all night online webinars for and I did not show you this but you can actually do some coding and if you have coding experience you can use codes and directly put them into [indiscernible] that is what this last one is referring to. You can keep on going with this this is just an intro. Now we have plenty

of time for questions. Again all the slides, handouts, editor that are at tinyurl.com/FDLC2019OPENREFINE. I'm going to shop sharing my, my, uh, my computer. No. Okay

---

Okay. So I got a couple questions Amy while you were presenting. I will go ahead and read those out. Now, if anybody has any questions please feel free to chat them in the chat box at the bottom right-hand but, make sure that you are submitting it to all presenters so that everyone can read the questions. So, uh, first it's kind of a common question. So, the question is I can't seem to find open refine program in the tiny URL folder. So is that there?

That is a good question, so in the folder there is a sheet that says downloading and opening open refine handout. In that hound out our instructions on how to download open refine and it takes you to the open refine website to follow the instructions and download open refine. So the folder did not include the actual software.

Okay, another question. Why work certain [indiscernible] missed on the first go around quick

---

That's a great question. So open refine has a method, an algorithm that it uses. Like the nearest neighbor and other methods. So, I have no idea why some of them are missed. I was a little surprised that they were missed. It has something to do with their algorithm and how they cluster. But, yeah, I will look more into that because [indiscernible].

Okay another question, some of the date you merge look like series. How do you merge some of the dates and not others?

That is a good question. You know what I would do is I would use some of the filters, some of the like filters for like materials types and then first filter out who [indiscernible] and just have your series and then do, then change the publication date. Yeah, so, if you have a giant spreadsheet just pick out your series 1st and then work with that.

Okay. Is there a way to unmerged or separate columns for which cells have been merged in some but not all row's ?

Yes there is a way. I've seen it done using the codes. As I said before open refine has special coding that you can learn how to use. Then that helps with the, uh, with the merging and on merging. Actually the original lesson done by Leanne and Kelly Shores does have a section on merging and on merging and I can definitely share that out with everyone as well. The original. It takes you more into that type of, uh, process and open refine.

Another question, is a control Q for Windows?

Yes it should be the same for both of them.

Okay and what is the different between flag and star in open refine do they have different functions?

No it does not seem like they do. I have not seen any different functions which is interesting why they have both of them. I think they just give you this is an option. They do the exact same thing.

How many records can open refine handle?

That is a good question that I have got before and from other participants in my workshop they have put in a lot of, a lot of, a lot of data and they still have been able to work with open refine. So right now it does not have a limit. So that is a good question and it seems like you can put in as much data as you want.

There is a comment, that would be great if you can share the original lesson plan. How do you want them to get that? Should the email you?

Yes, that would be great. Email me, I can send you everything.

Okay. Yes the email is at the bottom of her slide right there. Do you know if open refine will soon support access tables?

I do not know what access tables are. What are access tables?

Can't you export your table from Microsoft access?

I have not used Microsoft access. So I am not sure. I have not used Microsoft access.

Yeah I wonder what type of data files they are.

We have got a couple of more minutes for questions so if you have any questions please chat them in the bottom right-hand for Amy. But while we are waiting we will cover a couple of the upcoming webinar so that we have., It once, might computer Hayes. There we go. So on February 5 we have got the United States government manual, leave to digital. That February 10 is adoption foster care in children's well-being resources from the children's Bureau. Then on February 20 with that webinar introduction to federal research, a resource unreal education in the U.S. So feel free to sign up for any of those webinars with STP L Academy just on as TPL.gov. So, uh, it just scrolled up on me., From who? Just read it I can't find it.

What other projects have you used open refine for quick

What have I used other than for my reading project. I think it's done in the sciences or in the social sciences I mean like for surveys. People when they gather their own surveys and they want to clean up the data that is also another great way to use open refine. You know basically any type of Excel huge data set or file you have you can use open refine with to do simple cleanup. That are so time-consuming and Excel.

A sorry, thought you were done.

I was just looking at the chat and it looks like access is similar to Excel. So if handles Excel spreadsheet you should be able to use open refine.

Okay another question, can you treat a number as a date even if it has an extra character in it. So for example 1989 Dashe.

That was the issue that I encountered since it had the-are the extra character it did not let you sort by number. So first, you have to, uh, clean that up. Take those out before it let you do any sorting. Or filtering by number.

All right we are just about out of time so, uh, we are going to push that survey one more time. One last question though for you Amy before we go. In your opinion what are the top three advantages to using open refine over Excel for data cleanup?

Okay. First of all I would recommend using both. But, if you want to use open refine it is faster, it let you clean huge amounts of cells all at the same time. Use on one example we did like 200 cells all at once. Whereas in Excel it will take you hours to do the exact same thing. Also to verify that you are not missing the second, it let you verify that you are not missing information so instead of scrolling down in Excel spreadsheet and looking at everything you can just do some of the commands and, uh, open refine does that for you. Checks for blanks, or leading spaces, extra spaces. It is faster and it let you, uh, verify missing data.

Okay so we are at three clock, 3 PM Eastern time on the dot. So I just want to thank you Amy for presenting this wonderful webinar. There are plenty of really great comments in the chat box and I want to thank all of our, uh, viewers for tuning in today and we will see you next time.

Thank you. Thank you also. Thanks for coming.

[Event Concluded]