Transcript for Using SQL to Populate LibInsight

Slide 1

Thank you everyone for taking the time to view this presentation on Using SQL to Populate LibInsight. My name is Rob Behary and I am the Head of Systems and Scholarly Communications at Gumberg Library which is the main library at Duquesne University in Pittsburgh Pennsylvania.

Slide 2

What I will cover in this presentation are items ranging from the higher level “why” you may want to present your data in LibInsight to the nuts and bolts of the process, especially focusing on those areas where you may run into trouble or where you may gain the most benefit. I’ve found that things aren’t always intuitive in LibInsight, and I hope this presentation will help you negotiate some of the tricky areas I encountered.

These items include the system prerequisites, some notes about data extraction with SQL, data cleaning, data storage, and then the processes of uploading and presenting data in LibInsight.

Slide 3

What this presentation is not is a comprehensive guide to SQL. There is a great community of SQL users in the Sierra community who are always willing to help. I’m available to help as well though I acknowledge that my knowledge is largely practical and I build on what I have learned form others.

This is also not a comprehensive guide to LibInsight. I’m assuming some basics about negotiating LibInsight. Springshare support is quite responsive, and can fill in any gaps. Of course I am also happy to help where I can.

Slide 4

One quick nod to why LibInsight works for us. Data visualization is a really exciting field. We like LibInsight because it captures a lot of our processes automatically, such as LibGuides usage. It is also the place where we record our gate counts, reference transactions, ILLiad usage data, and all kinds of other things. We have one LibGuide that links to all of our LibInsight dashboards that we send to all library employees once a month to keep us focused on being a data driven organization.

The interface is really familiar and affordable. In the future, we may be moving to Microsoft’s Power BI software if the university can negotiate a site license, but Tableau has proven too expensive to deploy widely.

As you know, these are challenging times for most if not all libraries, and demonstrating our worth in a transparent and easy to access way can provide returns many times over whatever the cost of collecting and presenting data may be. At the beginning of the pandemic for instance, all of our employees were working from home, and we were able to use our metrics and our dashboards to demonstrate just how highly productive the library has remained.

Even in cases where the transactional data was not coming directly from Sierra into LibInsight, mashing up the data from the patron database with data from EZproxy transactions allowed us to show how the library continued to support each school and each university program.

Slide 5

Before you can begin, of course you will need to make sure your library subscribes to LibInsight. LIbInsight for us was an inexpensive addon to our existing subscription to the SpringShare suite of product. Check also that you can access SQL through Sierra. I believe SQL is included at all subscription levels to Sierra through III. Also either check with your IT group, or if you have more relaxed polices, see that you can install PGAdmin or a similar utility to connect with the Sierra SQL server. I like PGADMIN just because I am used to it now, but it seems like a lot of the past programs you will find on the IUG archives for SQL related sessions use PGADMIN to demonstrate. You will also need Excel. While you can natively upload the resulting .CSV files from PGADMIN to LibInsight directly, I always open the files, check for problems, clean the data, and check the date formats before uploading to LibInsight

Something within the Sierra software itself that you want to check is a software configuration setting. For privacy reasons, the default for Sierra storing circulation transactions is very short. We extended ours to 13 months of retention which works fine as we update the statistics monthly.

You will also need some expertise. Of course familiarity with uploading and downloading files, ability to edit SQL queries, the ability to use Excel, and some design work to present your dashboards. I'm still working on that myself.

Don't let these steps deter you though. There are a lot of steps that sound intimidating at first, but if you are watching this presentation you are obviously very smart and capable and this should be well within your skillset.

Slide 6

There are some great resources available beyond this presentation that you will have available to you. Postgres which is the flavor of SQL used by Sierra is really well documented. The Sierra SQL community is also great. On top of that, the tables for Sierra are very well documented on the Sierra DNA site, which you will need your help desk credentials to access. This slide shows a snippet of how the sierra DNA site is organized with arrangement by the different entities which open to the individual SQL tables.

Slide 7

You definitely want to think about the kind of data that you want to extract from Sierra. You also want to think about the way that the data is formatted and how you can translate that data into something that is useful for humans. As you know, there are labels associated with codes, and being able to export both the codes themselves which is very helpful for error checking, and the code names which will be used as the labels in LibInsight is critical to getting the most out of your data. The SQL that I will share, and that I use for our process, uses a series of Case statements to provide a human-readable name for each of the Sierra codes used. When extracting the data though, I make sure to include both the code and the human readable name in the output.

Clean data is also an issue. Depending on which extraction tool you use, you have to make sure that there are no errors in your data and that your date field is in a consistent format before importing into

LibInsight. I stick with the simple DD/MM/YYY format to be consistent across datasets. You want to make sure to look for any anomalies in your data. Usually I do a couple of quick sorts by columns that I know should contain no data just to see if there is any data present. Usually if there is data, it indicates a problem either with PGADMIN or with the export process itself. One time after an upgrade to PGADMIN, I noticed some extraneous data in the last column where the export to CSV was not working correctly. Since only a few lines were affected, I was able to correct the file without too much trouble. It's really things like that that can be a headache when moving data between two systems which is why opening the file in Excel can be super useful.

Slide 8

Think also about data storage. We are using our institutions box repository, but you can use a shared drive or some other place to park the actual data that you extract. You have to be careful to make sure you are handling your data and storing your data according to your institution's guidelines for such practices.

There are several reasons why we place our data in external storage as well as uploading to LibInsight or relying on Sierra alone. The first is that having data stored as CSV files or Excel files allows us to be ready to move to another system is LibInsight no longer serves our purposes. It's possible to download data from LibInsight as well, but having the Excel files saves us a step of having to re-download. Also, if there is a mistake or if we ever need to reload data, we can always wipe out a dataset and reload with our existing data. We're also able to share our data with other university systems, such as with the one person in institutional research who has access to Tableau. Finally, the data that we download and store from the SQL report is anonymized and cannot be recombined with transactional data in Sierra after the 13 months has expired. This protects our patron privacy.

As I note on the bottom of this slide, when you are working with data, be sure to be in compliance with your institution's data governance policies, and of course anything related to personally identifiable information, be careful about that.

Slide 9

An overview of the process for importing is to first prepare your SQL query, then connect to PGADMIN. Next use the Query Tool in PGADMIN to run your query, and executing the resulting CSV file to wherever you have decided to store your data. Check your data for errors, and then upload the CSV file. I'll go through each of these steps on the following slides.

Slide 10

I like to write my SQL in notepad2 or a similar text editor and store the query as a text file in the same folder where I store the source data. Never start the query from scratch. The Sierra community has been doing this for several years now, and you can always find someone who is sharing code that will match your purposes. I'll either share the code I use for this in our institutional repository or on the conference website with this presentation. You can always email me as well. Keeping the query in the same place as the data makes it easy to find every month when I do the uploads. Once I have the query in good shape, all that I have to change each month is the date range.

I want to point out something about the query in this slide that gets a little into the weeds o SQL but I think is an important consideration for writing queries that work well with LibInsight. Notice in Figure 3 that in the case statement, you will see both the code used by Sierra and the human-readable translation of that code. For example the code n is used to indicate that the transaction that was recorded is a hold. Similarly further down in the SQL, you will see that the code that translates into Gumberg 1st Floor is "g1" Because these codes are 5 characters long, and because all of them have similar starting strings, I've included as many spaces as needed to make the codes distinct. If I would have only used g and the wildcard directly after, which is the percent sign that you see on the lines, then the query wouldn't have been able to distinguish between codes that begin with g1 say and g1a. The human readable codes are really important as we will see in a few minutes when we go into LibInsight.

Slide 11

Once I have the SQL in good shape in a text editor, I'm ready to move it to PGADMIN. This slide shows PGADMIN. Notice how I have pasted my query into the query tool and after running the code, the table I will export appears at the bottom. In PGADMIN, the place where you can copy and paste your query is called Query Tool. There is a little lightning bolt icon that executes the query.

Slide 12

Here is the export tool which I use to download the resulting CSV file to our box folder for storage and uploading purposes.

Slide 13

Now that I have my data, I can go into LibInsight. Let me give you a brief overview. I go into LibInsight under manage datasets where I can create a new dataset. You only need to create the dataset once afterwards you will only visit to upload data

Slide 14

The first screen in LibInsight asks you to select a dataset type. From the dropdown select Custom. Even though there are some other database types that are transaction based, I've found that Custom gives me the most flexibility to build the dataset in any way that I need.

Slide 15

Pass through the second screen and accept the default permissions, and move to the third screen to start selecting the options for the dataset. Here are some tricks that I think work best when creating a dataset. I only select the display one date/timestamp option for data that I am importing. The other options don't seem to work as well in this context of Sierra.

Slide 16

When you are adding or editing your fields, make sure to use the Multi Select option. If you simply load your data, you will have all the transactions present in LibInsight, but you will be very limited in how you can analyze the data. Multi Select is definitely the trick to being able to do something other than the most basic counts without any breakdowns by category.

Slide 17

Here’s an example of setting up the Multi Select for Item Type Code. I put all of our itype codes into the options field. I did the same for all of our other headings in the other multi select fields. The one that took the longest to populate were our program type codes, which are codes that correspond to all of the individual majors at Duquesne. Once I had them populated though, I was able to do some in depth analysis. Notice all of our codes appear in the Options area.

Slide 18

With the dataset created, I’m ready to add my data. Here I use the plus sign associated with the dataset and upload my file.

Slide 19

One quick caution, be sure that your date fields in the spreadsheet you are uploading match the date format of your dataset in LibInsight. You will get some wacky results if the month and day fields are inverted. I try to standardize the date entry for all of my datasets as MM/DD/YYYY so that I am less likely to make a mistake during the upload process.

Slide 20

Create a dashboard by going to the Dashboard Manager field and selecting Add New Dashboard

Slide 21

Give your dashboard a title, a friendly url, and decide whether you want your dashboard to be public or private. I make all of our dashboards public because it helps me direct all of the library staff to the dashboards without having to worry about password access issues. You could if you want however make the dashboards private to force users to login to see the dashboards.

Slide 22

Once you have created the dashboard, you can begin to add elements to the dashboard. These will be a variety of graphs with texts that can be customized. I usually just take the default format of the graphs produced, but there are more colorful options available. To start adding elements, select the gear on the bar associated with the dashboard. I made a quick dashboard just for this presentation called IUG Presentation, and I made a short URL called IUG.

Slide 23

From within the dashboard, it is now possible to add visualizations. At the top of the screen, you can select the type of element that you would like to add. For this example, I am just selecting one wide chart from the dropdown list and clicking add new row.

Slide 24

A default chart will appear. Click the tool in the upper right portion of the screen to begin editing.

Slide 25

You will be taken to a settings screen where you can populate the Title, the dates you want to cover, and most importantly the data set. Notice that there is also a place to select background colors, but I will leave that those of you with better design skills than me.

Slide 26

From the Chart group, you will have the ability to add a title to the chart, select a chart type, and select what your Y axis will include. Here is where having all of the multi-select fields really pays off. If you did not have several multi-select fields, you would have the same data for every chart. Pie charts would be pretty much useless for instance because all data would be lumped into a single category. Note that if I had select a row type with 2 or more charts, there would be more than one chart group.

Slide 27

Finally we come to the data point. This is simply an aggregate count of all of the points associate with a field. You can do a sum of the field for numeric data, but because we want a count of all transactions, I have selected count of the field. I labeled the field Total Tranasactions.

Slide 28

The resulting chart shows the totals from each of the transaction types along with the total transactions for the data selected. You can continue to add rows to your dashboard until you have provide a single place for those who need the data to go each month, and once the dashboards have been created, they will update automatically when you add new data to LibInsight.

Slide 29

Once you have completed our upload, you will have a dataset that has many categories, many options for display because of the multi select fields, and you will have a relatively inexpensive pathway to creating basic visualizations.

Slide 30

Thanks so much for viewing this presentation. I have my contact information should you have any questions or need any help. I invite you to look at some of my other publications where I am trying to contribute to the library literature of building a culture of decision making grounded in data.