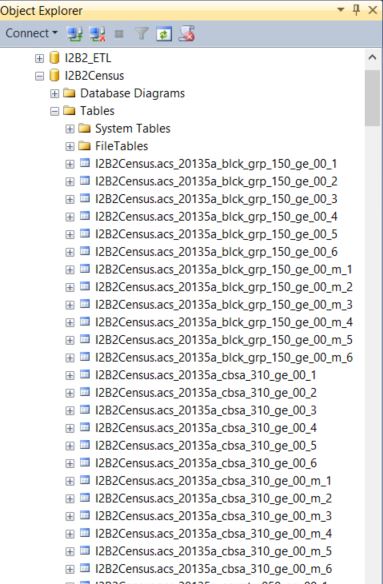
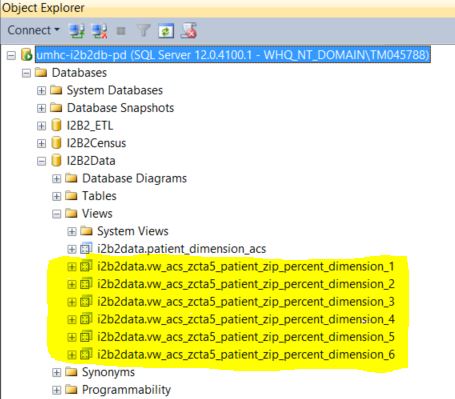
ACS and Census Ontology Data and Design

After reviewing the ACS/Census data we at MU made a couple of decisions due to the amount of space it would require to load all of the data, and to store into observation\_fact. By our calculations it would take around 3.27 TB per section to store all of this data. Since we do not have that type of space available, but also wanted to load every column for selection we came up with design alternative using views to accomplish the task.

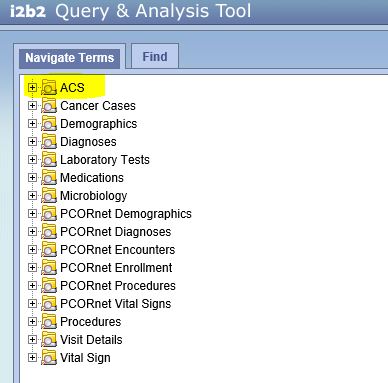
We initially used an open-source ETL tool called Pentaho Data Integrator in order to build the tables and load the ACS/Census data. We had to split out the data into multiple tables due to the number of columns in the flat file and Sql Server’s limitation of 1024 columns per table. We also only loaded data for Missouri and its surrounding states to save space.



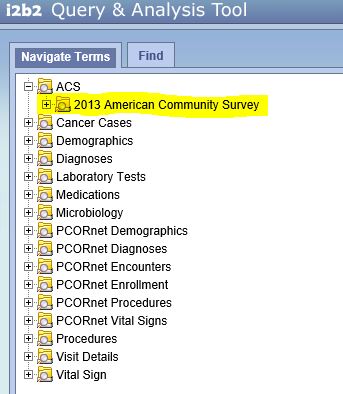
We then created a view for each table number that would link the **patient\_dimension** to the corresponding table where the column is stored. These views also calculate the percentage for the columns with a total column. We did this instead of returning the raw data in order to normalize the data across geographic boundaries. More than 85% of the data (in group a files?) has a total column associated with the data. These are the only ones we have loaded to date. We will look at using percentile for those data that do not have a total column, as well as making the raw data available for queries.



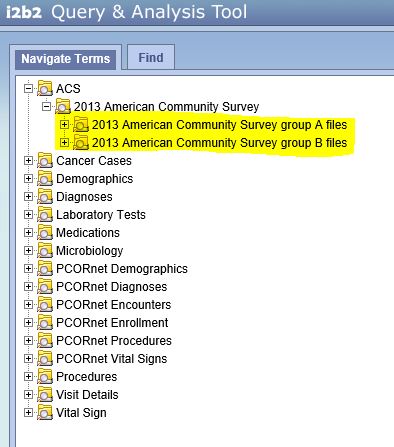
The Ontology build itself will consist of around 10 levels of data. The first level is ACS or Census. So far we only have ACS.



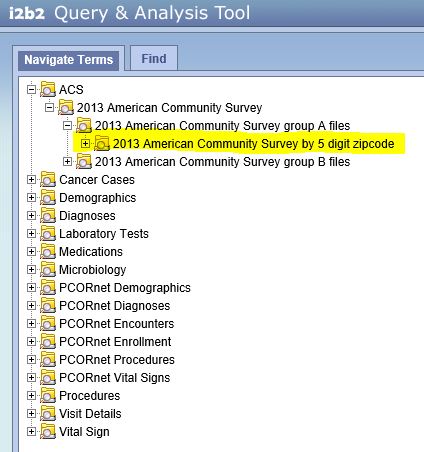
The second level will differentiate between years of data, so when a new ACS or Census comes out it can be added with the defining year.



The third level is separating out the different groups within the data. For instance ACS has an A and a B group of data.



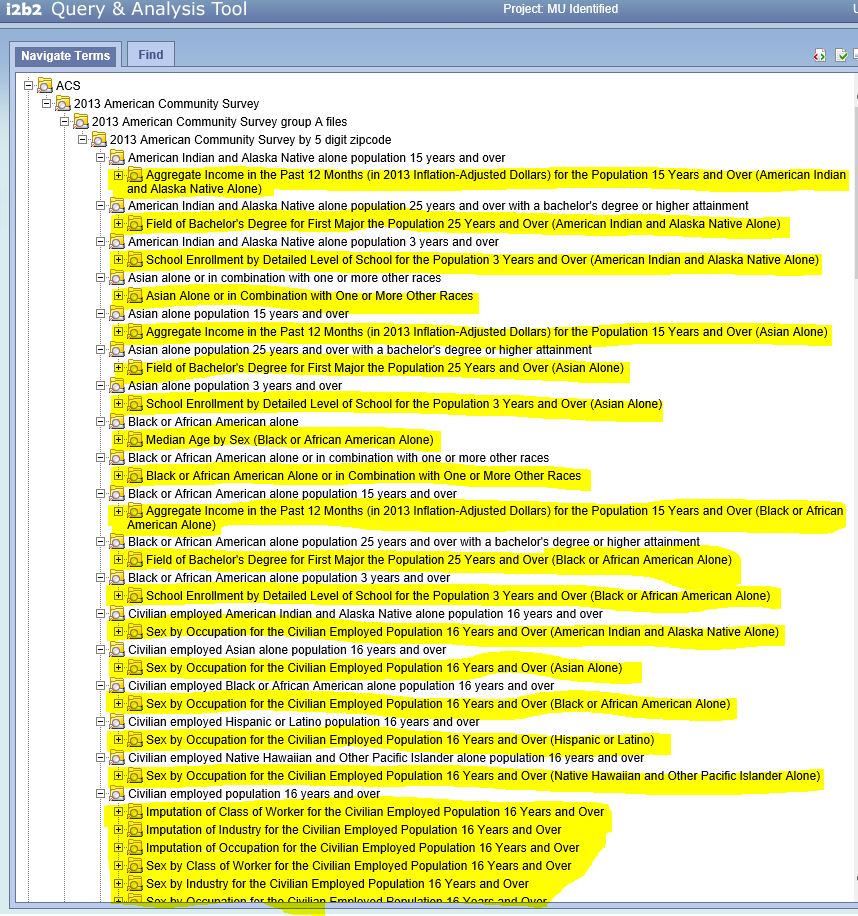
The fourth level is separating out each layer of data. ACS will have zipcode, blck\_grp, cbsa, county, cty\_sub, metdiv, place, state, tract, and urb\_area.



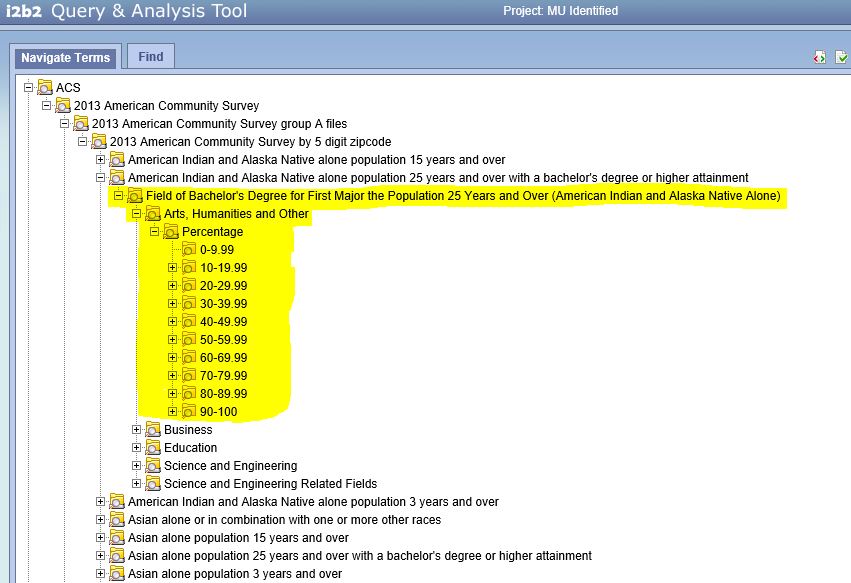
The fifth level is the top tier categories that start breaking down the data to Families, Race, Gender, Population, etc…



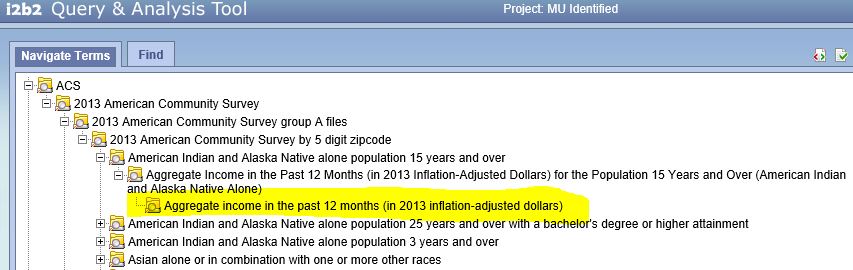
The sixth level of the ontology, is for the sub-categories.



The seventh through 10th levels are going to be the breakdown and calculations. So either returning the raw data, or percentile, or percentage data for the corresponding areas.



Percentage example. There will be a final level broken down by 1% increments.



Percentile/Raw example. (not populated yet)