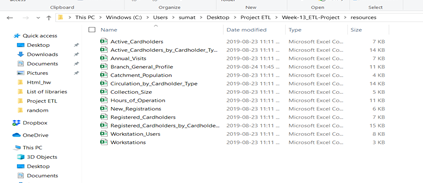
**Week 13-ETL Project report**

The ETL project comprised of Toronto Public Library data, which combined multiple data sources to perform data extraction, normalisation and finally transformed them to depict how different branches performed along with their in-house programs offered in the past years.

The primary web source we used was “<https://opendata.tpl.ca/>.” This link lead us to some equally interesting data to work with. The data was in different formats like JSON, CSV, HTML

**Extract: Data sets**

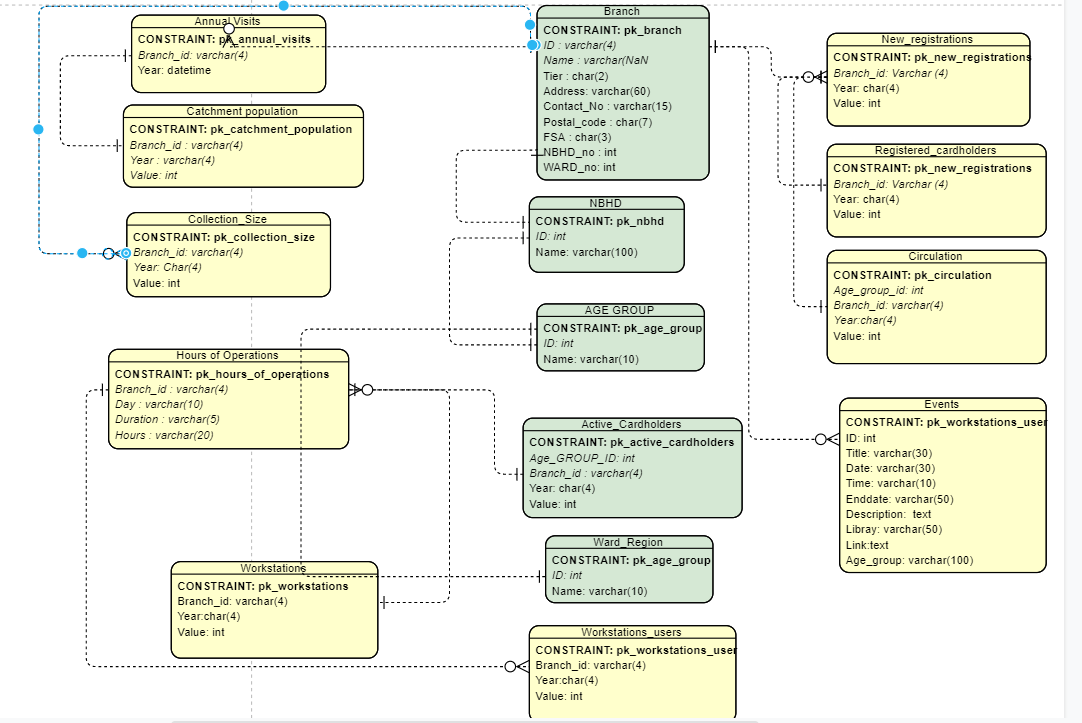
First step for the project was to extract the data. The link shared above lead us to further  to different CSV, JSON and HTML files. We also used the URL for the list of library branches. Some of the datasets we considered for our project are about active-cardholders, branch information, hours of operation,workstations and annual visits. Following is the list of the csvs used to build the database.



**Transform: Data cleaning or transformation**

The data sets had multiple values so the first step was to normalise the data. Followed by data cleaning with python scripts. There were different tables that need to be connected through primary key. It took us a lot of time to decide the schema. We used ERD to finalise the schema.

**ERD Diagram**

****

**Load: the final database**

The final step was to load the data. For this we used SQL alchemy to save the data. All the data is in relational database mode. This data can be further used to run queries and find out more information about Toronto Public library.

**Why we chose this ?**

Majority of the data was in CSV format so it seemed like a perfect choice to use relational database to store the final database.

**Additional learning:**

Web scraping with beautiful soup. We tried to use some of the web scraping skills we learned this to get a list of books from the library web page.

<https://www.torontopubliclibrary.ca/books-video-music/new-items.jsp?category=Adult+Fiction>