## Project Details

This project is connected to the Data Wrangling course. You have the choice between two databases for this project: SQL and MongoDB. For an explanation of the differences between these two databases, see **this node**. There are separate instructions where relevant below for each database choice.
Here's what you should do:

### Step One - Complete Programming Exercises

Make sure all programming exercises are solved correctly in the "Case Study: OpenStreetMap Data" Lesson in the course you have chosen (MongoDB or SQL). This is the last lesson in that section.

### Step Two - Review the Rubric and Sample Project

The **Project Rubric**. will be used to evaluate your project. It will need to Meet Specifications for all the criteria listed. Here are examples of what your final report could look like:
**SQL Sample Project**
**MongoDB Sample Project**

### Step Three - Choose Your Map Area

Choose any area of the world from **https://www.openstreetmap.org**, and download a XML OSM dataset. The dataset should be at least 50MB in size (uncompressed). We recommend using one of following methods of downloading a dataset:

- Download a preselected metro area from **MapZen**. Note that XML files are compressed in a `.bz2` format, so you may need a tool like 7-zip to obtain the full XML data.
- Use the **Overpass API** to download a custom square area. Explanation of the syntax can found in the **wiki**. In general you will want to use the following query:(node(minimum_latitude, minimum_longitude, maximum_latitude, maximum_longitude);<;);out meta; e.g. (node(51.249,7.148,51.251,7.152);<;);out meta; the meta option is included so the elements contain timestamp and user information.
- You can use the Open Street Map **Export Tool** to find the coordinates of your bounding box. *Note: You may not be able to use the Export Tool to actually download the data, the area required for this project is too large. After selecting a region, you can try clicking the Overpass_API link down the tool bar to perform the export and download.*

## Step Four - Process your Dataset

It is recommended that you start with the problem sets in your chosen course and modify them to suit your chosen data set. As you unravel the data, take note of problems encountered along the way as well as issues with the dataset. You are going to need these when you write your project report.

### SQL

Thoroughly audit and clean your dataset, converting it from XML to CSV format. Then import the cleaned .csv files into a SQL database using **this schema** or a custom schema of your choice.

### MongoDB

Thoroughly audit and clean your dataset, converting it from XML to JSON format. Then import the cleaned .json file into a MongoDB database.

### Hints and Tips

- Feel free to adapt the code from the Case Study lesson to help you approach the auditing of your data. It will help your organization by creating a new script for each aspect of your dataset that you audit. Each field that you audit should also include a function that will help you update your dataset.
- You may want to start out by looking at a smaller sample of your region first when auditing it to make it easier to iterate on your investigation. See code in the notes below for how to do this. You can use a small (1-10MB) sample to make sure that your code works, and then an intermediate sample to check for the most common problems to clean.
- Remember to perform data cleaning when you convert the XML into CSV or JSON format. You won't change the original data file, only the data that you plan on inserting into your database. This is where your earlier organization will pay off, since you can just import the update functions from your auditing scripts into the cleaning and conversion script.

## Step Five - Explore your Database

After building your local database you'll explore your data by running queries. Make sure to document these queries and their results in the submission document described below. See the **Project Rubric** for more information about query expectations.

## Step Six - Document your Work

Create a document (pdf, html) that directly addresses the following sections from the **Project Rubric**.

- Problems encountered in your map
- Overview of the Data
- Other ideas about the datasets

Try to include snippets of code and problematic tags (see **MongoDB Sample Project** or **SQL Sample Project**) and visualizations in your report if they are applicable. Use the following code to take a systematic sample of elements from your original OSM region. Try changing the value of `k` so that your resulting `SAMPLE_FILE` ends up at different sizes. When starting out, try using a larger `k`, then move on to an intermediate `k` before processing your whole dataset.