Wrangling OpenStreetmapData with MongoDB

Selected map area: Budapest, Hungary

https://www.openstreetmap.org/relation/22719
https://mapzen.com/data/metro-extracts/metro/budapest_hungary/

Table of contents

1. Problems Encountered in the Map

My first problem was that Hungary doesn't have so many big cities. :) My birth town and the place where I live currently was accessible in the database, but its size was well below the necessary 50MB.

Next, I chose the region where I live (Southern Great Plain in Hungary). It was large enough (around 2GB uncompressed), but this area lays along the border, so there were numerous street names from Serbia and Romania. Finally, I chose Budapest, the capital of Hungary.

To assess clarity of data I listed and collected user defined fields from a subsample of the whole dataset (~10% of the original).

First, I checked the country codes. I found where any country code was given it was correct and belongs to Hungary.
Second, I checked if there is any typo, abbreviation of the name of any city, but I found everything correct here.

After the mentioned audit I found the following potential errors:

- Inconsistent house numbers (17, 21B, 21/B, 21b ...).
- Inconsistent postal code (1023 vs. H-1012)
- Varying capitalization of street types (utca vs. Utca)
- Inconsistent phone number format

Inconsistent postal code

The postal codes in Hungary consists of 4 digits. Sometimes (especially on international mails) they write as H-1213 (where H denotes the country). In almost every case, only the number were given, so I removed the leading "H-".

Varying capitalization of street types

The street types were surprisingly correct in my sample, only a couple of times was some capitalization issue.

Inconsistent house numbers

The basic house number is a single decimal number. If a land was divided, they put a / after the number and then a single letter like: 17/A, 17/B. In my dataset I found various forms: lower case letter and/or without slash. I transformed all of them to the standard format.

Inconsistent phone number format

Phone numbers in Hungary has the following pattern:

- begins with country number (+36)
- followed by a 2 digit number (area code)
- then followed by 7 digits (for mobile phones) or by 6 digits (landlines)

One can write a phone number in various formats e.g.:

- +36201234567
- +36 20 123 4567

- +36 (20) 123-4567
- +36/20/1234567

For simplicity, I transformed each of them to the standard format (without space, dashes, brackets...)

2. Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

3. Additional Ideas