Open Street Map Project

Damian Brunold

Problems in the data

I focused on the street names. Since the map region is located in switzerland, the street names are generally in german, although some swiss-german names do occur. Since the street names in german use no separating space between the name and the word “Strasse”, the audit process had to be adapted from the one used in lession 6.

First I defined a set of common street-name-endings and matched the street names against these. Everything not matched was printed. I did this several times, each time expanding the set of endings until no obvious street ending was left. The resulting list of non-standard streets were manually scanned. This resulted in some obvious errors (e.g. “strassse” instead of “strasse”).

The final set of endings was:

```
expected = ["strasse", "weg", "gasse", "feld", "halde",
            "gässli", "gass", "-Strasse", "-Weg", "platz",
            "rain", "hof", "garten", "park", "matt", "weid",
            "berg", "grund", "acker", "höhe", "blick", "bach"]
```

But then I decided to list all unique street names in the dataset and check them manually for problems. This was a little bit of work, but it yielded quite some more problems.

I wrote a “corrector” function, that automatically corrected these problems. It was based on the following bad-to-good mapping:

```
mapping = { "Vorstdt": "Vorstadt",
            "Steinhauserstrassse": "Steinhauserstrasse",
            "St.Oswalds-Gasse": "St. Oswalds-Gasse",
            "Campus GrüentalUnset": "Campus Grüental",
            "Eintrachtsrasse": "Eintrachtstrasse",
            "Zur Weid Rossau": "Zur Weid",
            "Altschlossstrasse;Reidholzstrasse": "Altschlossstrasse",
            "Bahnhofsplatz": "Bahnhofplatz",
            "Bärenacher-Strasse": "Bärenacherstrasse",
            "Chrisimatt": "Chriesimatt",
            "General-Guisan Strasse": "General-Guisan-Strasse",
            "Kapperlerhöhe": "Kappelerhöhe",
            "Saentisrain": "Säntisrain",
            "Schellenmatstrasse": "Schellenmattstrasse",
            "Schnellenmattstrasse": "Schellenmattstrasse",
            "Tellen-Strasse": "Tellenstrasse",
            "Tuergass": "Türgass"
            }
```

As is evident the following classes of problems were found:

- Typo (e.g. Vorstdt instead of Vorstadt)
- Inconsistent/Incorrect name (e.g. St.Oswalds-Gasse instead of St. Oswalds-Gasse)
- Data entry errors (e.g. Campus GrüentalUnset instead of Campus Grüental)

I used a combination of personal subject matter knowledge (I live in this region) and an official map on the internet for figuring out the correct versions of names.

Doing this cleaning work, I realized how varied the street names in switzerland are. Since we have both standard german as well as our dialect swiss-german, many street names exist in some variants. It is not that easy to decide whether there is an error or whether these are really correct variants.

Another complication are umlauts (e.g. ä, ö, ü). These are really common, but sometimes a variant is used, i.e. Bühlenstrasse vs. Buelenstrasse. Is the latter an error or a correct variant? Only official street name data or personal subject matter experience can help decide this question.

Exactly this problem was evident in the name of the street I live in: 'Saentisrain' was the name in the xml file. But as I live there, I happen to know for sure that it is spelled 'Säntisrain'.

Apart from these problems, I had the impression that the data was of a rather high quality.

Overview of the data

File sizes:

map.osm.xml 108M

map.osm.json 115M

Number of documents

```
> db.osm.find().count()
```

```
563137
```

Number of nodes

```
> db.osm.find({'type': 'node'}).count()
```

```
488530
```

Number of ways

```
> db.osm.find({'type': 'way'}).count()
```

```
74607
```

Number of disticts streets

```
> db.osm.distinct('address.street').length
```

```
1049
```

Number of unique users

```
> db.osm.distinct('created.user').length
```

```
758
```

Other ideas about the dataset

As noted in the section on problems in the data, the street names in the german speaking part of switzerland are a mix of german and swiss-german. As swiss-german is not homogeneous but instead is a collection of quite different dialects that are regionally distributed, one idea for using this data springs to mind: Why not use dialect-specific street names to map the street locations to dialects. In this manner, it would probably be possible to make a crude map of the distribution of swiss-german dialects.

For example: A street named ‘Main Street’ in an english speaking country might be named ‘Hauptstrasse’ in a german speaking country and ‘Rue Principale’ in a french speaking country. Switzerland does have four official languages: german, french, italian and rumantsch grischun. So it would be rather easy to map the whole country according to the language of the words of the street names.

More subtle is the situation in the german speaking part of switzerland. Here there are about a dozen dialects (variants) of swiss-german. And although most street names are written in non-dialect german, there are quite a few that are in dialect swiss-german.

In order to accomplish this, we would need to use a much larger area of the map data, in order to get enough different dialects in the data. We would then scan the street names and compile a map of street name parts and the associated dialect. But running this map over the whole data set, it would be possible to assign each street that does match one of the map name parts a dialect. By using a geospatial mapping tool, we would be able to produce the crude dialect map.

Note: this has nothing at all to do with the users createing the data or the language they speak. It is about the street names themselves!