Reconciliation of Point-of-Interest Locations on OpenStreetMap

Students

Marc Havrilla

Damian Dasser

Advisors

Prof. Stefan F. Keller, Nicola Jordan

Subject Area

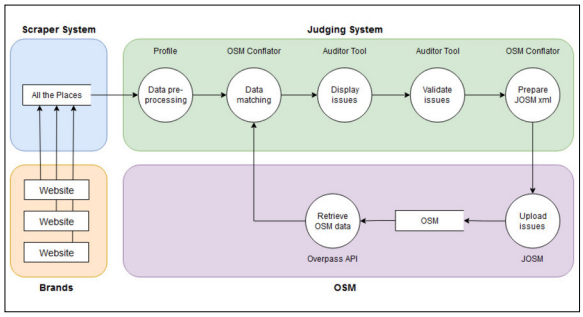
Software

Introduction: OpenStreetMap (OSM) is a crowdsourcing project, which creates and distributes open geographical data of and for the world. OSM is largely built and maintained by volunteers. The continuous change of the real world circumstances can hardly be managed and reflected on maps in general. Consequently, only selective and punctual updates are made and this causes the data on OSM to be neither perfectly accurate nor complete.

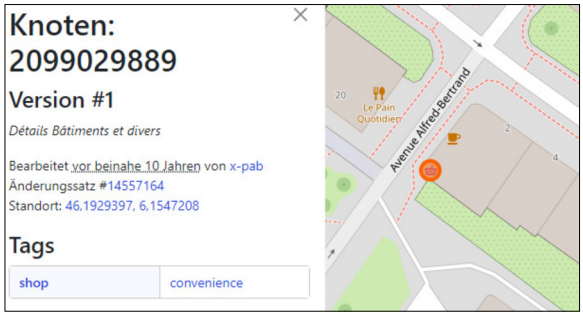
Problem: To overcome the before-mentioned shortcomings of OSM, web scrapers can be used to collect publicly accessible brand data. The collected data will be stored in a scraper system and compared with OSM by a judging system to identify discrepancies. The results will be manually validated by OSM members and then be uploaded to OSM. Various open-source tools have been assessed to find an ideal scraper and judging system to perform the matching between publicly available data and the OSM database. The combination of "AllThePlaces" (ATP, www.alltheplaces.xyz) as the scraper system and "OSM Conflator" as the judging system has turned out to be the most suited. The OSM Conflator obtains the external data from ATP. Profiles are used to pre-process the data before the OSM Conflator performs the matching. Currently, for each brand a separate profile has to be created.
During this task, five profiles were created in Python and the data output by the OSM Conflator was analysed. The brands for the profiles were selected by their availability on ATP and their localisation in Switzerland. Thus, the selection consists of Fust, Aldi, Jumbo, Coop Vitality and the public toilets of Zurich city. On average 12.6% of the entries from the external sources do not exist in OSM and were flagged as new entries. One problem is the questionable quality of the tag elements, which prevented a match between external data and OSM clues for almost 2% of the entries. On top of that, the data quality of the external sources has room for improvement. In 1.6% of all analysed entries, a matching was not possible as the coordinates from the external sources were too unprecise.
Additional findings concern the creation of queries within the OSM Conflator. As a result of schema mismatches between the different data sources, it is challenging to include suitable tags in the profile to create a precise and performant query for a certain brand. Besides the non-standardised schemas, also the matching of the data is challenging due to spelling aspects of brand names, regional differences in brand naming and real-world circumstances, like e.g. company takeovers.

Conclusion: The output of the OSM Conflator shows that data reconciliation of OSM data is important. A more generic profile, which pre-processes the data for the OSM Conflator, would be ideal. Allowing a profile to dynamically extract different tags from the external data source and directly generate multiple individual queries for OSM, might constitute an interesting improvement for the OSM Conflator. Furthermore, standards for the data sources would be required to establish a more generic profile. The input data needs to be accurate, consistent and trustworthy. These improvements would also enable the reconciliation of regions rather than brands alone.

Dataflow from brand websites over ATP to the matching and validation in OSM Conflator and being stored in OSM
Own presentment



Example of an OSM element using a minimal schema which therefore is impossible to unquestionably match it to a brand
OpenStreetMap



A matched store location with suggested changes (in green) in OSM Conflator during the manual validation process
OSM Conflator, OpenStreetMap, admin.ch