Qualitätsprüfung der Verknüpfungen von OpenStreetMap nach Wikidata

Diplomanden

Timon Erhart

Jari Elmer

Ausgangslage: OpenStreetMap (OSM) ist eine freie Landkarte der gesamten Welt mit rund 1 Milliarde geografischen Objekten. Wikidata (WD) ist eine freie Wissensdatenbank mit über 97 Millionen Einträgen, die strukturierte Daten zur Verfügung stellt. Beides sind crowdsourced Open-Data-Projekte und haben grosse und aktive Communities, welche die Daten laufend aktualisieren und ergänzen. Seit 2014 ist OSM in der Lage, über sogenannte Tags (Key-Value-Pairs) eine Verknüpfung zu WD herzustellen. Zurzeit existieren rund 5.5 Millionen solcher Wikidata-Tags mit einer stetig wachsenden Popularität. Mittels dieser Verknüpfung lassen sich interessante Produkte bauen, beispielsweise eine Karte mit Burgen und Schlösser, die mit Sachdaten aus WD angereichert wird. Die Qualität dieser manuell erfassten Verknüpfungen in OSM ist bislang jedoch unbekannt und ungeprüft.

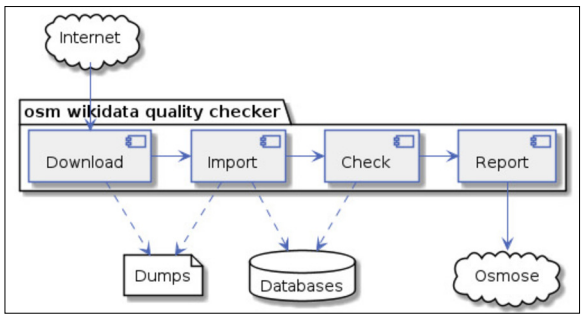
Ziel der Arbeit: Es soll eine Applikation entwickelt werden, welche die bestehenden Verknüpfungen von OSM nach WD prüft und aufbereitet. Die gefundenen Fehler - beispielsweise ungültige WD-Einträge in OSM - werden mit einem Korrekturvorschlag an die externe Fehlerdatenbank Osmose gesendet. Ziel ist es, dass die Applikation ein dauerhafter Teil der Infrastruktur zur Qualitätssicherung von OSM wird. Es muss mit den grossen Datenmengen der beiden Datenbanken (je ca. 1.5 TB) zurechtkommen und die wöchentlich erscheinenden Datenbank-Dumps innerhalb dieser Frist verarbeiten können. Ausserdem ist auf gutes Software-Engineering und Code-Qualität zu achten, sodass das Tool gewartet und weiterentwickelt werden kann.

Ergebnis: Die gesetzten Ziele wurden alle erreicht und die entwickelte Applikation "osm wikidata quality checker" liefert einen wertvollen Beitrag zur Qualitätssicherung von OSM-Daten. Das Tool läuft auf einem Server in einem Docker Container, besorgt sich selbstständig die Datensätze, verarbeitet diese und sendet die gefundenen Fehler am Schluss an das Osmose-Frontend. Es ist in der Lage, diverse Typen von fehlerhaften Verknüpfungen mit einer hohen Treffsicherheit von > 95% zu finden. Durch den Einsatz von Multiprocessing und des entwickelten Datenbankmodells, bei dem nur die relevanten Daten extrahiert werden, ist es in der Lage, mit den grossen Datenmengen umzugehen und die gesamte Welt innerhalb der geforderten Frist zu prüfen. Auch Schwierigkeiten im Umgang mit crowdsourced Daten, bei denen mit unvorhergesehene Datenfehlern gerechnet werden muss, wurden erfolgreich gemeistert, sodass eine hohe Fehlertoleranz erreicht wurde. Eine ausführliche Dokumentation sowie die leicht verständliche Architektur ermöglichen es, das Tool auszubauen und weitere Checks zu implementieren. Die optionale Konfiguration bietet die nötige Flexibilität im Betrieb und hilft bei der Weiterentwicklung. Aktuell werden insgesamt über 30'000 Fehler in neun Kategorien gefunden:

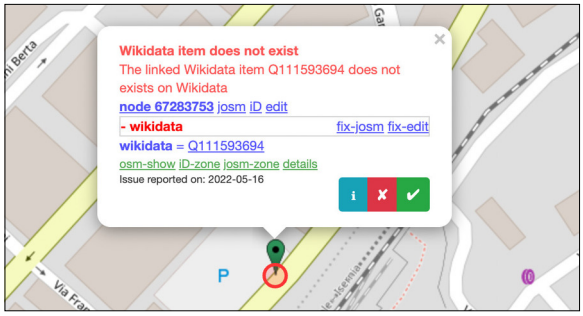
- Der verlinkte Wert hat ein ungültiges Format
- Der WD-Eintrag existiert nicht
- Der WD-Eintrag ist eine Weiterleitung
- Die WD-Kategorie ist nicht erlaubt (WD Internals, Listen)
- Der WD-Eintrag hat eine unerlaubte Taxonomie für den Link-Typ
- Eine Örtlichkeit (Dorf, Stadt, Land usw.) ist falsch verlinkt
- Die Distanz zwischen OSM- und WD-Eintrag ist ungewöhnlich gross
- OSM- und WD-Kategorien stimmen nicht überein
- WD-Kategorie passt nicht zum Link-Typ

OpenStreetMap-Objekt "Rapperswil" mit Wikidata-Tag, dessen Wert Q688539 auf www.wikidata.org/wiki/Q688539 verweist
Eigene Darstellung

Ablauf und Datenfluss der Applikation "osm wikidata quality checker" (Osmose ist eine externe Fehlerdatenbank)
Eigene Darstellung



Kartendarstellung im Osmose-Frontend eines gefundenen Fehlers "WD-Eintrag existiert nicht"
Eigene Darstellung



Referent
Prof. Stefan F. Keller

Korreferent
Dr. Ralf Hauser, PrivaSphere AG, Zürich, ZH

Themengebiet
Application Design, Software

Projektpartner
Sascha Brawer, Rapperswil, SG