Matching and Conflation of Open Government Data with OpenStreetMap Data

Enhanced OSM Integration Workflow due to better POIs Matching

Student

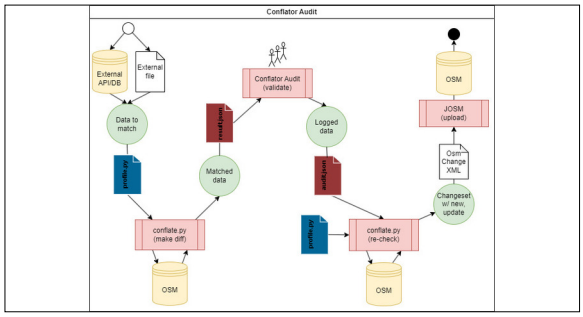
Claudio Bertozzi

Initial Situation: The integration of external geospatial datasets into OpenStreetMap (OSM) is a crucial task for enhancing the richness and accuracy of digital maps. As the use of geospatial data expands across various sectors, there is a growing need to automate the process of identifying and reconciling discrepancies between diverse datasets. This project was initiated to address these challenges, particularly focusing on the integration of data from the AllThePlaces (ATP) project into OSM. Traditionally, manual efforts to align such data with existing OSM features are time-consuming. A more automated and scalable solution would be desirable to handle the increasing volume of data and to ensure that updates could be managed efficiently. Previous projects had laid the groundwork with advanced matching algorithms, but the integration of these into a cohesive system that could handle real-world data in a robust manner had yet to be realized.

Objective: The primary objective of the project was to create a service that could automatically generate "diff" files highlighting differences between OSM and external data sources such as ATP, called DiffedPlaces. These diff files would be used for further analysis, validation and integration into OSM. To achieve this, the project sought to integrate existing tools and methods, such as the OSM Conflator, with advanced Point of Interest (POI) matching algorithms developed in previous projects. The system was designed to be modular, scalable and capable of handling continuous data updates. QuackOSM with DuckDB was chosen for the persistence layer. DuckDB provides fast processing of large geospatial datasets while being lightweight (in-process). In addition, the project uses Docker containerization to ensure that each component of the service can operate independently, making it easy to deploy, maintain and scale. The ultimate goal was to create a reliable and efficient framework for geospatial data integration that supports the community and the maturity of OSM.

Result: The project successfully developed a service that automates the process of generating diff files, enabling the efficient integration of external geospatial data into OSM. Using the capabilities of the OSM Conflator and incorporating the advanced POI matching algorithms from a previous project, the service was able to identify and highlight discrepancies between OSM and ATP datasets. Using QuackOSM with DuckDB proved effective for data storage and processing, supporting rapid querying and manipulation of large datasets up to country level. Docker containerization provided the necessary modularity and scalability, allowing the system to be easily deployed and managed across different environments. The project also established a web-based download service to allow users to easily access and review the generated diff files. Despite some challenges with data import and the need for further refinement of the matching algorithms, the DiffedPlaces service represents a significant step forward in the automation of geospatial data integration. It provides a robust foundation for future enhancements, including extending its applicability to a wider range of datasets, improving the accuracy of the algorithms, and implementing periodic automated execution. This project not only addresses current needs, but also positions the DiffedPlaces service as a valuable tool for the ongoing maintenance and improvement of OSM.

Current workflow showing interaction of data collection, merging, re-check, and storage in OSM using the OSM Conflator
Own presentment



New workflow with DiffedPlaces service for data collection and re-check, integrated with OSM Conflator
Own presentment

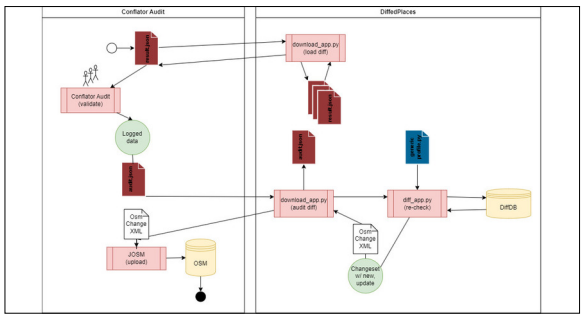


Illustration of accurate OSM data with community collaboration, highlighting the results of the DiffedPlaces service
OpenAI DALL·E. (2024). Digital map with accuracy

Advisor
Prof. Stefan F. Keller

Subject Area
Data Science, Computer Science