Spatial Data Representation and OpenStreetMap Integration in OpenRefine

Graduate

Labian Gashi

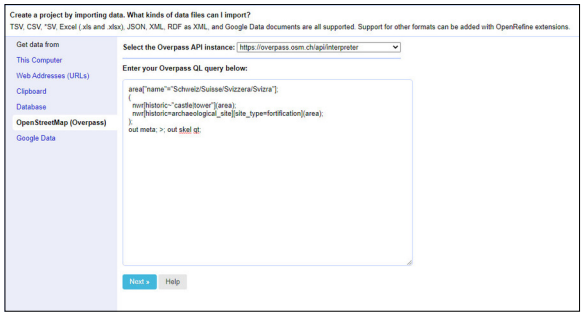
Introduction: The need to integrate geospatial data into large data-driven applications has increased rapidly over the last years. Along with this demand, new tools have emerged. OpenRefine is a low code development platform and a data integration tool for working with potentially messy data. The web application provides various features for data transformation and cleaning, including the blending of data with web services and external data. It also includes its own scripting language, called the "General Refine Expression Language" (GREL) - hence a low code development platform - which allows users to manage data with functions such as string processing or web scraping. OpenRefine also has an architecture that allows users to install new extensions, and developers to add functionality using Java.
Currently OpenRefine cannot process geodata and does not support services for accessing geodata sources such as OpenStreetMap (OSM), except probably the Geonames project. OSM is the most popular free map in the world.

Definition of Task: In this thesis we incorporate spatial data representations in OpenRefine as defined by the OGC Simple Feature Access geometry types, and by encoding them as a Well-Known Text (WKT) or as columns containing latitude/longitude coordinates. We extend OpenRefine to import OSM data, to process spatial data, and to export this data in the GeoJSON format.
Two extensions are built for achieving these goals: "OSM Extractor" and "GeoJSON Export". Both utilize OpenRefine's extension architecture and are developed according to the given development guidelines.
The OSM Extractor extension allows OpenRefine users to directly use OSM data in OpenRefine. For this an OSM Overpass service (which can be configured) is accessed by using the Overpass Query language. The retrieved OSM elements (Nodes, Ways, Relations) are then postprocessed into proper geometries, and a new OpenRefine project with those geometries and with columns (from Tags) is created.
The GeoJSON Export adds another option to export an OpenRefine project in this file format. The user can choose from a OpenRefine project which columns to include in the export, and if the geometry attribute is taken from a WKT column or from latitude/longitude columns or both.

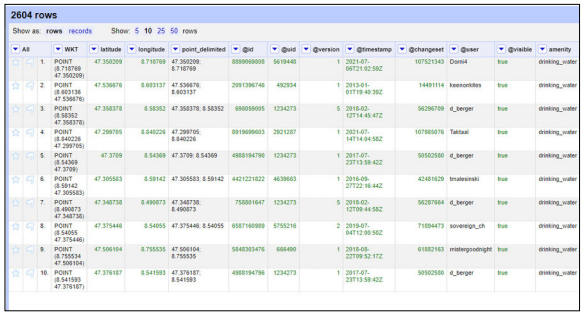
Result: The OSM Extractor extension has been successfully implemented and is able to fully integrate OSM data represented as a geometry of type Point, LineString, MultiLineString or MultiPolygon. OSM tags (attributes) are also included whereas main tags are sorted to the beginning. It also includes a GREL function named "interiorPoint()" that extracts the center point of a geometry, this allows for easier integration e.g. with Wikidata.
The GeoJSON export creates a file while letting the user choose the columns as described before. In addition, the decimal digits of the coordinates can be overridden.
Future versions of OSM Extractor could save the Overpass query like a database connection. Or it could allow synchronization to keep an OpenRefine project up-to-date with OSM. Contributions to the OpenRefine core project are also in the works.

OpenRefine - A data integration tool
Taken from OpenRefine's official website

The "OpenStreetMap (Overpass)" import option that allows the users to retrieve data from OpenStreetMap
Own presentment



An OpenRefine project, created using the "OSM Extractor" extension
Own presentment



Examiner
Prof. Stefan F. Keller

Co-Advisor
Claude Eisenhut, Eisenhut Informatik AG, Burgdorf, BE

Subject Area
Software and Systems