

# 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.

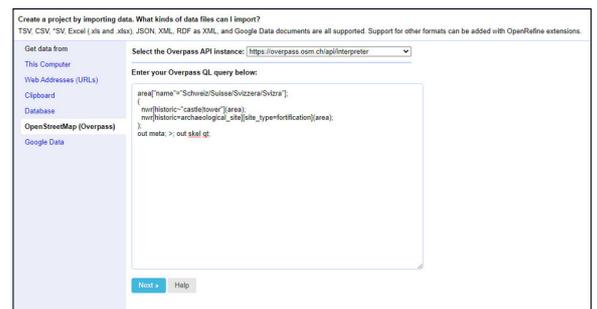
The GeoJSON export creates a file while letting the user choose the columns as described before. In addition, the decimal digits of the coordinates can be overridden.

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 presentation



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

2604 rows										
id	WKT	latitude	longitude	print_details	@id	@lat	@version	@timestamp	@changeset	@user
1	POINT (8.18789 47.36029)	47.36029	8.18789	819090000	5919446	1	2021-07-08T21:02:50Z	107212143	Caner	true
2	POINT (8.536676 6.603137)	6.603137	8.536676	20912901746	492354	1	2015-03-01T19:40:30Z	14491114	kaaronides	true
3	POINT (8.58352 47.28975)	47.28975	8.58352	888090005	1234273	5	2016-03-12T14:45:47Z	56236769	a_benger	true
4	POINT (8.58352 47.28975)	47.28975	8.58352	888090005	2921207	1	2017-07-14T14:54:52Z	107905076	Takusa	true
5	POINT (8.54369 47.3709)	47.3709	8.54369	4888194796	1234273	1	2017-07-23T15:59:42Z	50502560	a_benger	true
6	POINT (8.30563 8.59142)	8.59142	8.30563	4421221022	4639663	1	2016-06-27T02:16:44Z	42401629	Immanuel	true
7	POINT (8.49073 47.34873)	47.34873	8.49073	758001647	1234273	5	2016-03-07T09:41:50Z	56236764	a_benger	true
8	POINT (8.375446 8.54055)	8.54055	8.375446	658716009	5755216	2	2016-07-04T12:00:50Z	71039473	kovalevich	true
9	POINT (8.586164 8.755035)	8.755035	8.586164	5646203476	666460	1	2016-06-27T09:52:17Z	61882163	materpoodlight	true
10	POINT (8.541893 47.376187)	47.376187	8.541893	4888194796	1234273	1	2017-07-23T15:59:42Z	50502560	a_benger	true

**Examiner**  
Prof. Stefan F. Keller

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

**Subject Area**  
Software and Systems

