Tutorial on RISIS CORTEXT Geospatial services

Ludovic Villard, Juan Pablo Ospina Delgado, Luis Daniel Medina

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

WP6

RISIS CORTEXT GEOSPATIAL SERVICES

January 2020

Lionel VILLARD – Université Gustave Eiffel

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

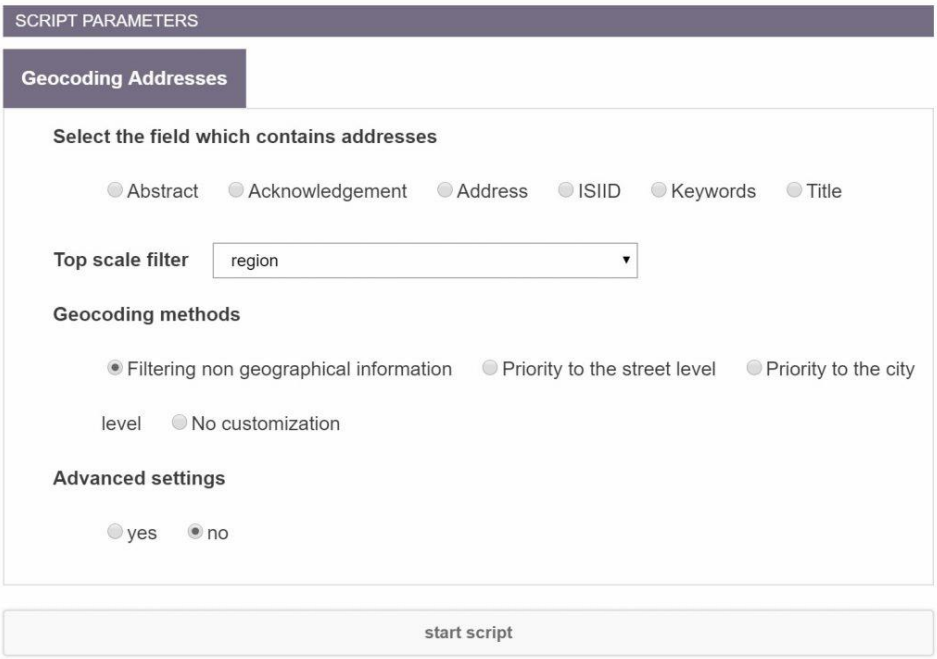
1. RISIS CorText Geocoding service

Modern geocoding engines "tackled the problems of assigning valid geographic codes [longitude and latitude coordinates] to far more types of locational descriptions [than older methods] such as street intersections, enumeration districts (census delineations), postal codes (zip codes), named geographic features, and even freeform textual descriptions of locations." (Goldberg et al., 2007)

The CorText geocoding engine has been built to manipulate semi-structured addresses written by humans. So, it is able to solve complex situations as:

- Different formats that rely on national postal services (or data providers), that largely vary across country;
- Non-geographic information (building names, lab names, person names…), that have ambiguities and could be multi-located;
- Ambiguous toponyms (e.g. Is "Paris" one the Paris in Canada or the capital city in France? Is "Osaka" in Japan, the region name or the city name?);
- Alternative and vernacular toponym names.

1.1. Parameters



- *Select the field:* select the field which contains the list of address. It should be formatted a least with a city name and a country name (e.g. "Paris, France"). Ideally, country names should follow ISO standard. As the country boundaries are an important information to locate addresses, CorText Geocoding service is able to deal with aliases for country names (e.g. USA, US, United States of America). When needed, intermediary names (state, region, county….) are also useful to help reducing ambiguities. Postal codes (but not postal boxes) are powerful and non-ambiguous information and are supported for 11 countries.

- *Top scale filter:* remove from results all geocoded addresses that have been geocoded above this scale. If having centroids at the "*region*" scale is not an option, "*county*" will provide a better precision.
- *Geocoding methods:* choose which geocoding method to use according to your needs. See the above definitions. "Filtering organisation names" is recommended as it will provide the best ratio between the quality of the results and the coverage. As geocoding addresses is slow, "Filtering organisation names" is also the fastest option. To expand the coverage you could consider to choose 0,4 as a "*Confidence threshold*". It will also increase the proportion of false negatives.

1.2. New added variables

Depending the method chosen, users will get different results. The Cortext Geocoding engine is not only able to geocode your addresses (longitude and latitude coordinates), but also enrich your corpus with three new variables that can be used in other CorText scripts:

- *geo_city*: name of the "city", chosen between locality name, localadmin name, neighbourhood and county name. More generally geo_city is the locality toponym name, rather than the official administrative name of a city;
- *geo_region*: region name identified by the hierarchy or identified directly in the address if it was the only information associated with the country name. Regional layer is feed with different geographic elements that rely on the administrative divisions of the country (e.g. state in US, department in France, prefecture in Japan);
- *geo_country*: standardised country name (ISO standard).

Villard, Lionel & Ospina, Juan Pablo (2018). *CorText Geocoding service*, ESIEE Paris, Paris-Est University. https://docs.cortext.net/geocoding-addresses/

2. RISIS CorText Urban and Rural Area (URA) tool

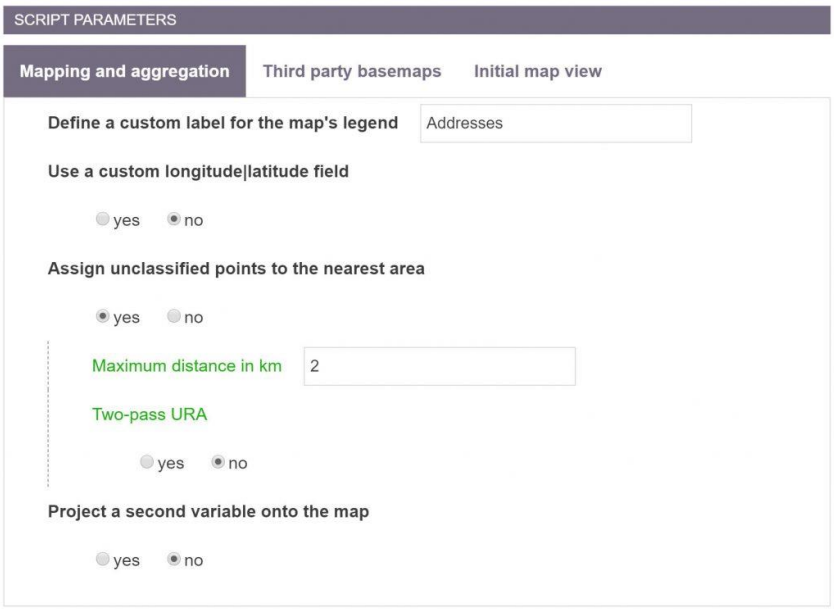
Geographical phenomena are by nature distributed. To deliver geographical coordinate to an address is one thing, but for the purpose of spatial analysis it is often necessary to aggregate geographical information, and associated data, on a more suitable scale. It is the case, for example, for analysing inter-urban or regional spatial dynamics based on human activities.

The CorText Geospatial Exploration Tool offers an original way to solve this, by combining in one unique tool a large variety of well-established sources of shapes. CorText Geospatial Exploration Tool is designed to work after CorText Geocoding service. It proposes basemaps at four different scales:

- URA boundaries: for Urban and Rural Areas, which includes a worldwide coverage of urban areas and, in the meantime, regionalized boundaries (states, departments, provinces, prefectures…) of the areas which are outside of the highest concentrations of inhabitants;
- UA boundaries: for Urban Area, is a subset extracted from the previous geographical layer, which includes only urban areas, to make it simpler (if needed);
- Regional boundaries: a worldwide layer for regional analysis;
- Country boundaries.

2.1. Parameters

The first step is to select which field contains the geographical coordinates (***Use a custom longitude | latitude field)***. This field is produced by CorText Geocoding service as ***geo_longitude_latitude*** which is selected by default.



If you want to use your own list of longitude/latitude coordinates, you can upload a csv file where a column holds the two values separated by a pipe (e.g. 104.068108|30.652751). If you have multiple values per document, you may use a separator (e.g. 104.068108|30.652751***4.89973|52.37243 in a cell). In that case, select **yes** and precise the name of the field.

CorText GeoSpatial Exploration tool uses shapes with a high level of precision: from 10 meters to 5 kilometers depending the layer and the source. For geographical space: distance matters! So, on one hand, at this level of precision a geocoded address may not fall into the shape for a small distance (as for an address in a park, next to a river or in a port, in an island… excluded from the urban areas). On the other hand, an address which is close to a boarder, may fall outside the right shape for only a few meters due to some simplification of the boundaries.

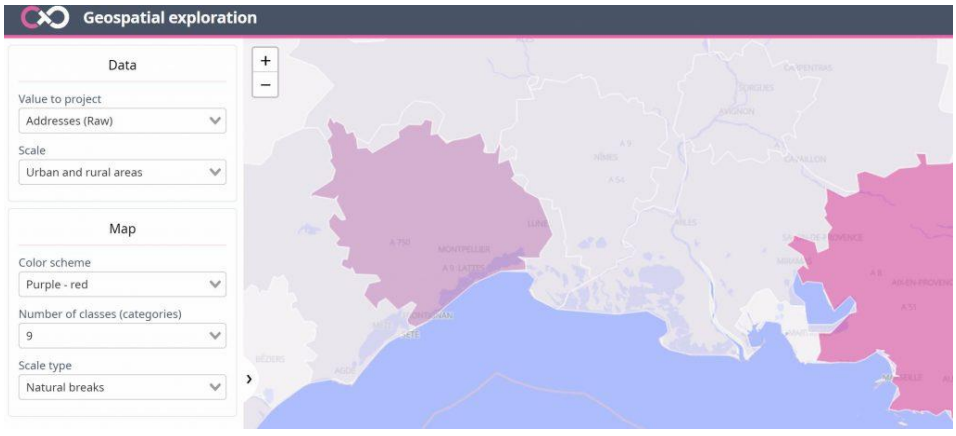
If you want to avoid these situations, you may choose the parameter ***Assign unclassified points to the nearest area***. If yes (which is the default behaviour), CorText Manager will look for each outlier point (geocoded address), in a given perimeter (by default 2 km), which are the closest centroids of shapes (urban or rural areas, regions…). After having selected the three closest centroids, CorText Manager will identify for the outlier which is the closest boundary between the three shapes selected by their centroids. Finally, the point will be affected to this area.

By default, the outlier points will be attributed to the nearest area found, whatever if it is a urban area or a rural area. For the URA layer, if ***Two-pass URA*** is selected as ***Yes***, the spatial join will proceed first with urban areas only, and finding the nearest point (geocoded address) in a given perimeter, and secondly with rural areas for the remaining points. This parameter creates a buffer zone around the urban areas of a specific distance (2 km if using the default value). It aggregates locations (geocoded addresses) with the urban areas if they are close to it, even if they would have fall in a rural area.

If ***Project a second variable onto the map*** is selected, a given field is used to tag the shapes with the ***Top N elements***. It could be useful for semantic clusters, or to draw profiles with classes for regions or urban and rural areas. This parameter enriches the legend of a map with this Top N elements and produces a csv file for each layer.

2.2. Interactions with the choropleth maps

Several interactions are accessible to explore the data in different ways.

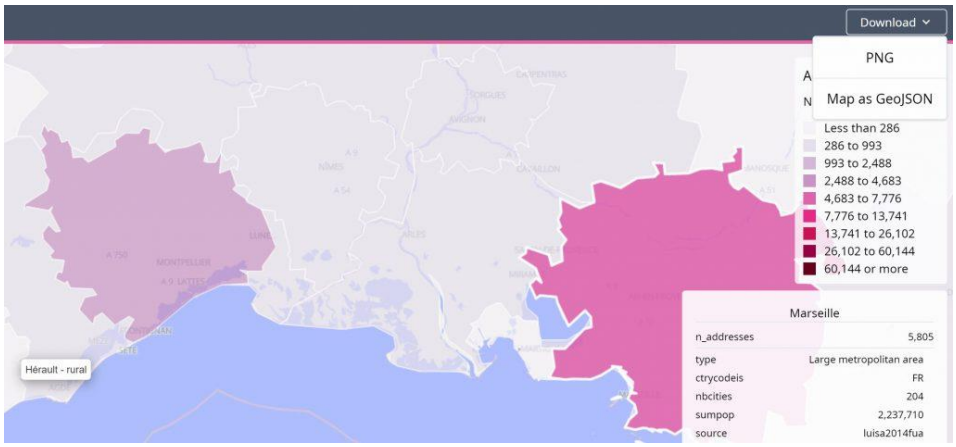


Two methods are accessible to classify the values projected onto the maps:

- ***Quantiles***: the effectif of shapes represented in map is divided in N classes. Each class receives the same amount of shapes;
- ***Natural breaks***: Jenks natural breaks optimization (Jenks, 1977) is made to classify data for choropleth map in a more human readable way. It tends to "minimize each class's average deviation from the class mean, while maximizing each class's deviation from the means of the other" classes. We are using here a recent optimization (Schnurr, 2016) of the original method.

Due to the precision of the layers, it would have been too heavy to load it into a browser. The full layers are used for the geospatial join (to ensure the highest possible accuracy), but the shapes visualized onto maps have been widely simplified. We have used the Douglas-peucker algorithm (Douglas Thomas K., 1973) to reduce the number of points of the shapes' boundaries. Up to 90% of the points (in some portions of the URA layer) have been removed and replaced by lines, with two constraints that the algorithm should follow:

- none of the shapes should completely disappear (so, the minimal shape is a triangle);
- take in account shared lines (e.g. boundaries between two countries, to avoid holes between them).

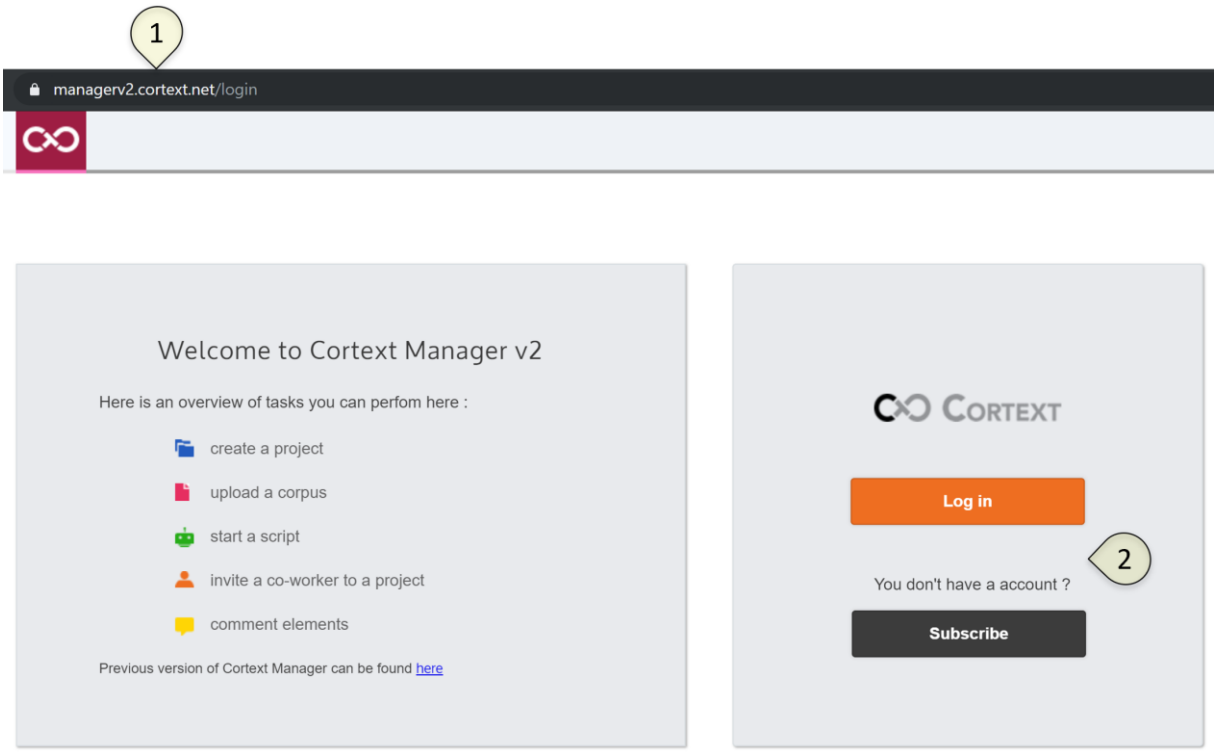


You may want to do further work with the maps outside CorText Manager: you can export the maps in geojson format. Keep in mind the layers and shapes you will download are the one which have been simplified. Furthermore, only shapes with a value are mapped. So, you will gather and download in the exported geojson file only shapes where at least one point from your dataset have been projected. Please ask us if you want the precise version of URA layer with all shapes.

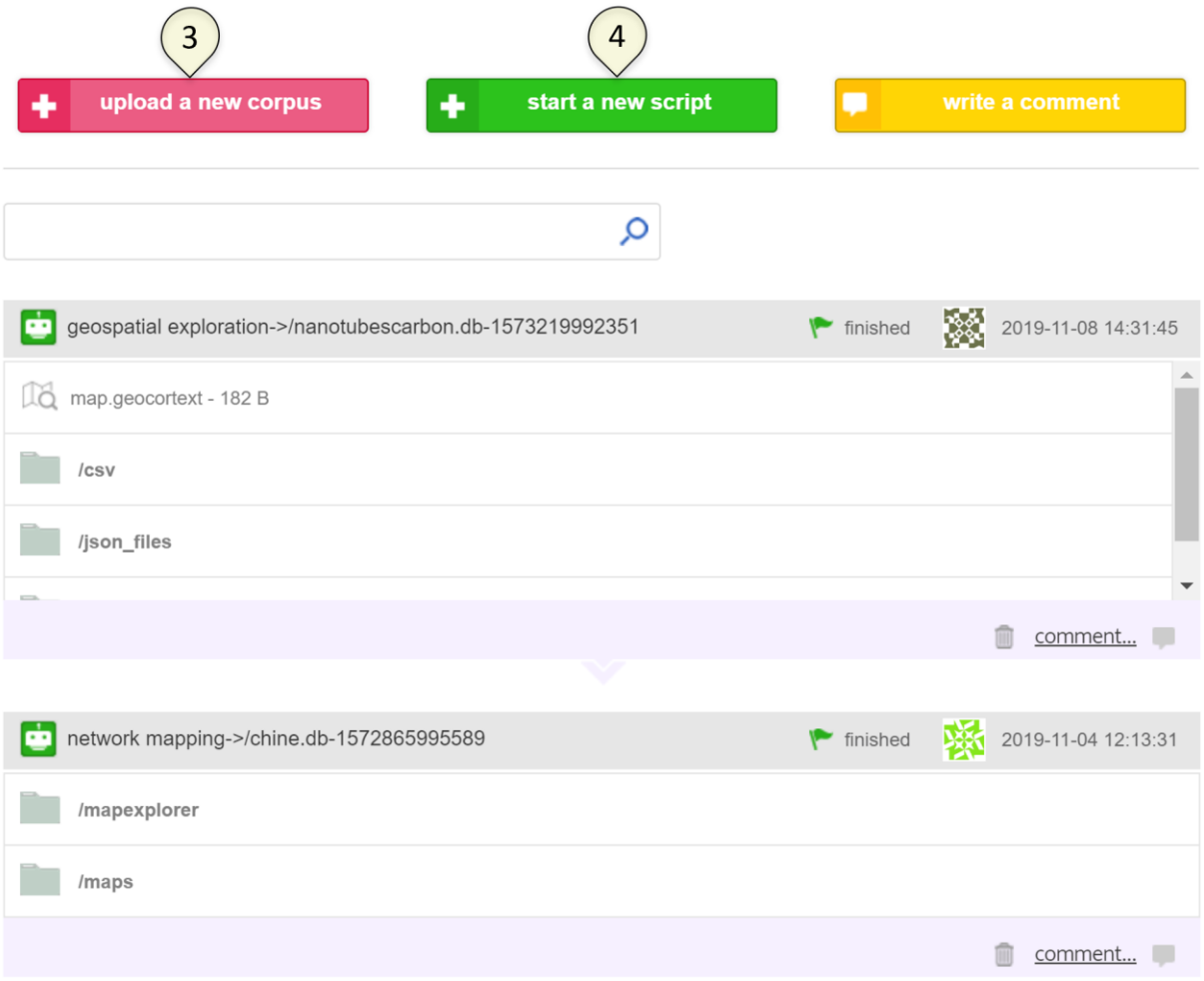
Villard, Lionel, Opsina, Juan Pablo & Medina, Luis Daniel (2019). *CorText Geospatial Exploration Tool*. ESIEE Paris, Paris Est University. https://docs.cortext.net/cortext-geospatial-exploration-tool/

3. How to access to the services

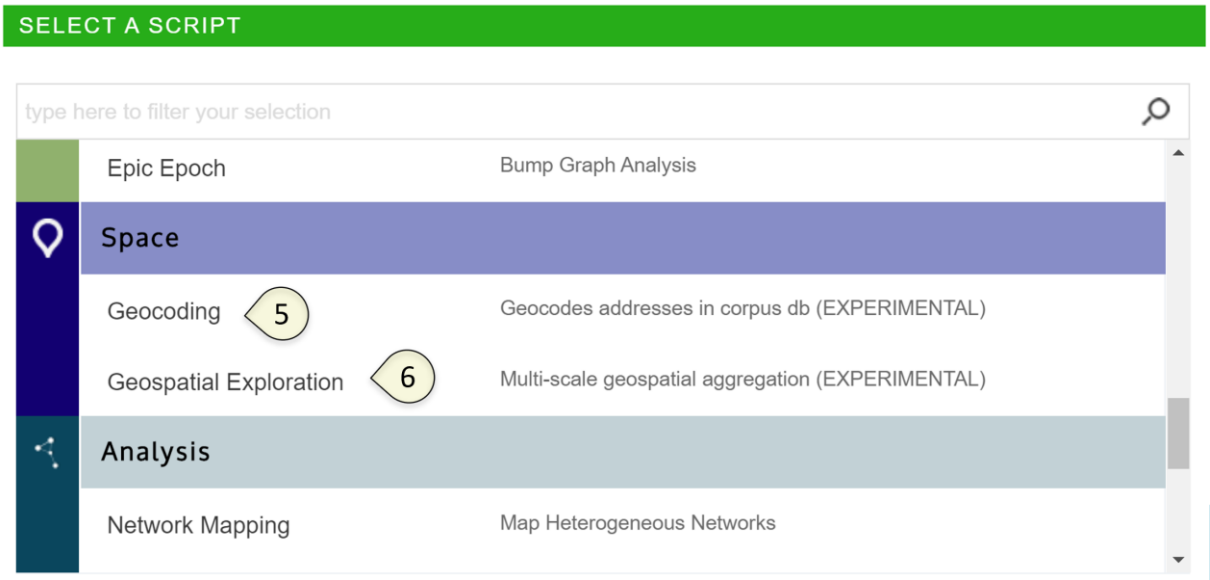
3.1. Log into CorText Manager



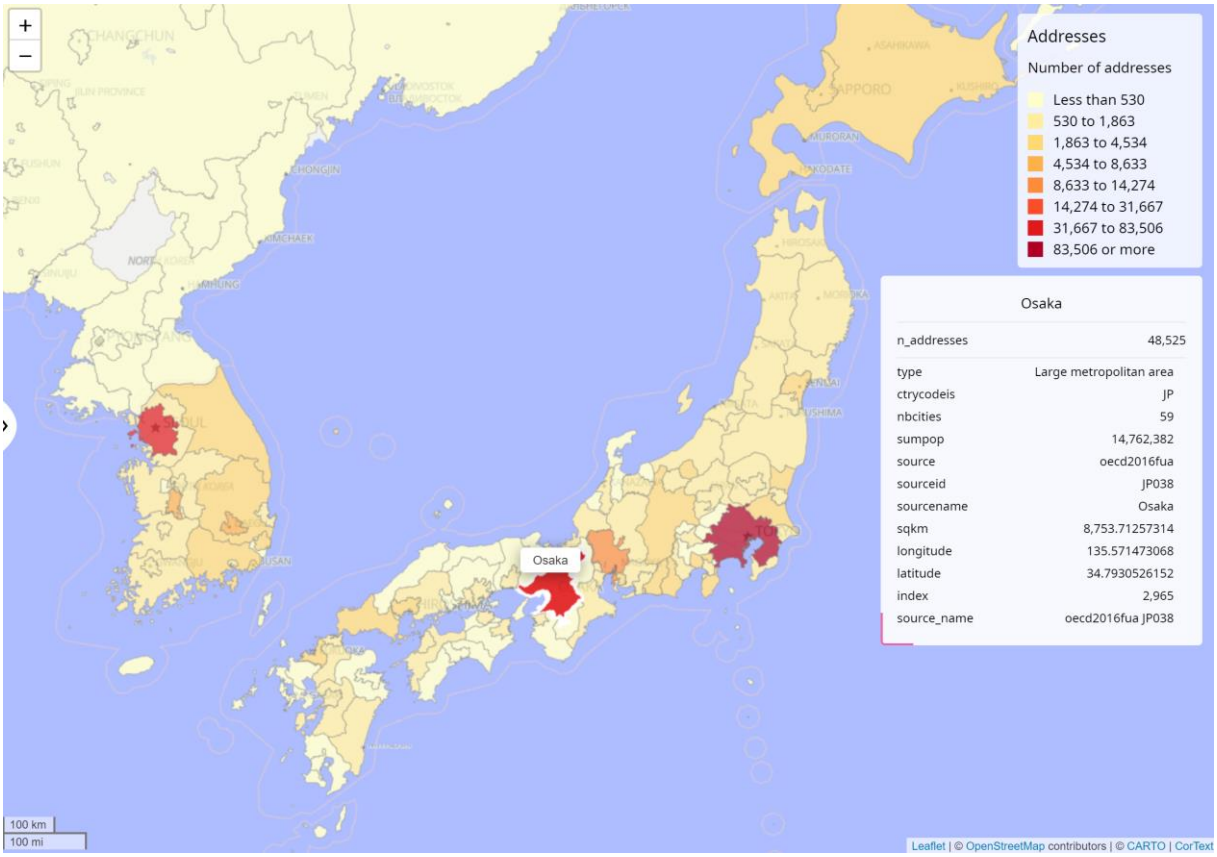
3.2. Load a dataset and run a script



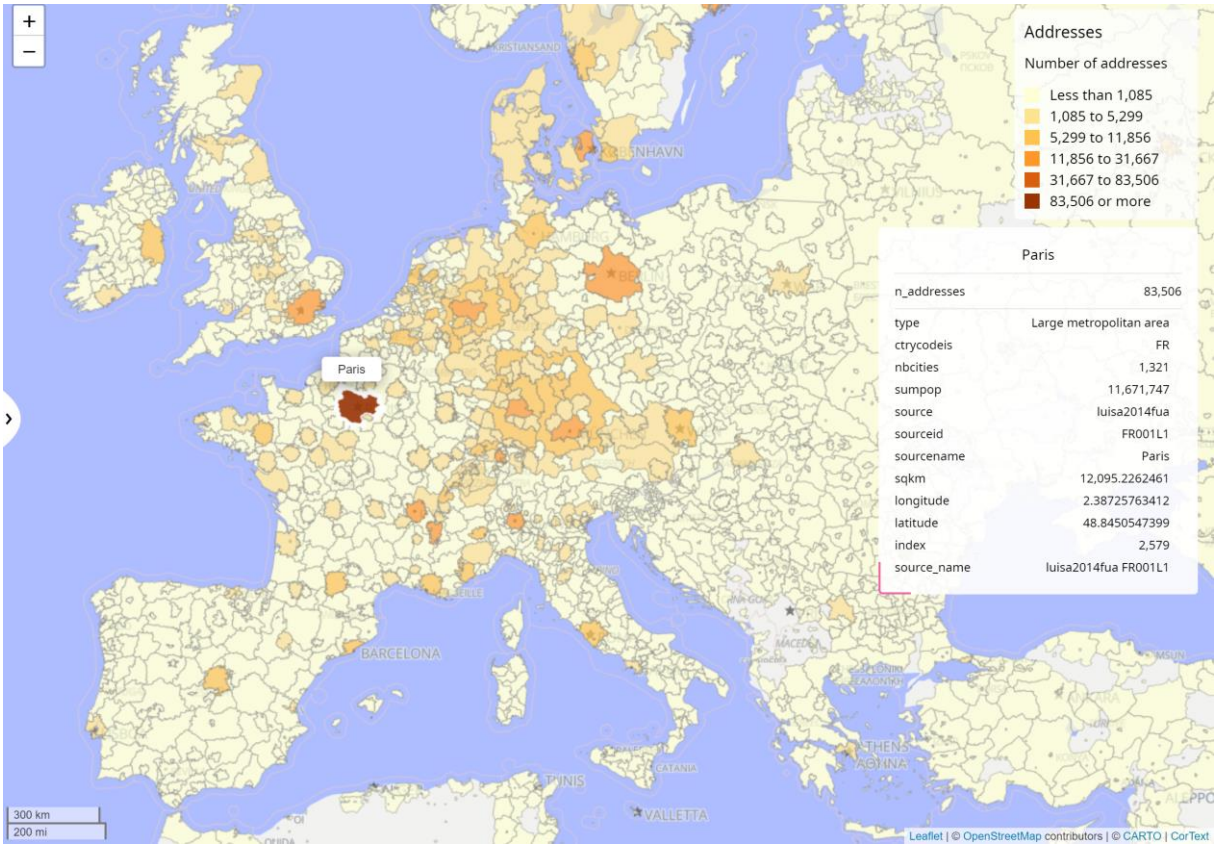
3.3. Choose CorText Geocoding or Cortext Geospatial exploration scripts



3.4. Navigate through your maps



Patents inventor's addresses per Urban and Rural Areas



Patents inventor's addresses per Urban and Rural Areas