

Thematic and spatial analysis of technologies using CorText Manager and RISIS patent database

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE
AND INNOVATION POLICY STUDIES



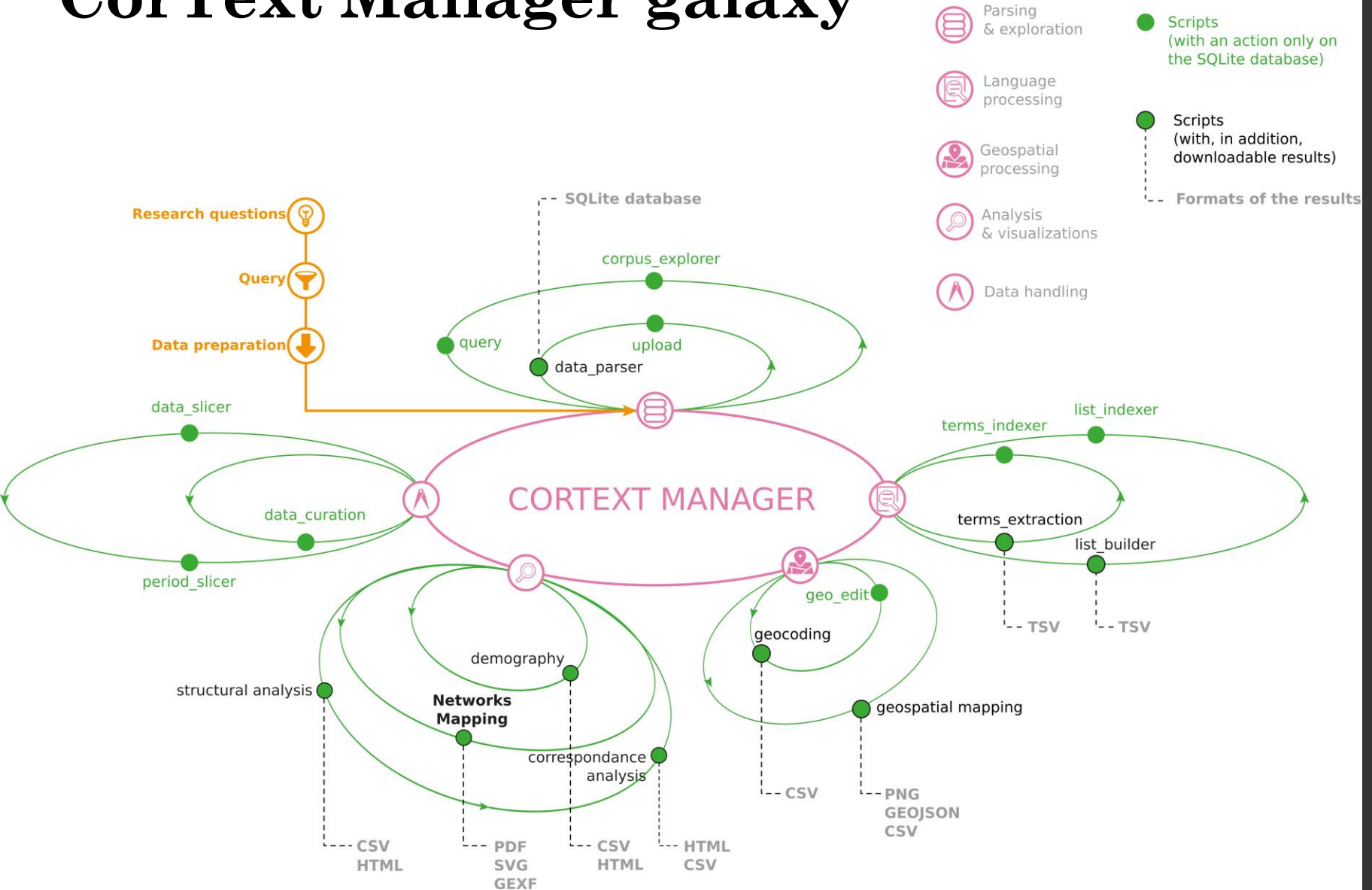
RISIS training sessions, 2021

Lionel VILLARD

LISIS, IFRIS, INRAE, CorText, ESIEE Paris

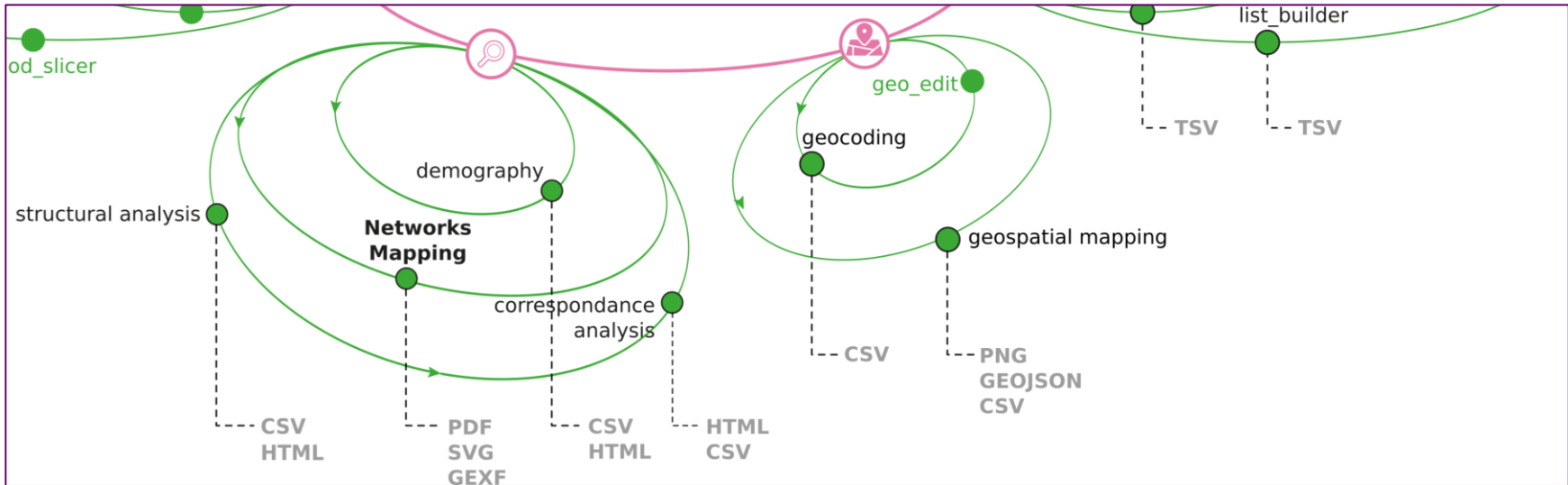
A global roadmap for users

CorText Manager galaxy



Dealing with « locations »: geocoding and mapping

1/ From data to spatial analysis



Examples of inventor addresses in a list of patents

id	rank	address
36	0	Hamdard Univ, Fac Eastern Med & Surg, Karachi, Pakistan.
36	1	Herbion Pharmaceut Pvt Ltd, Dept Med Affairs & Training, Karachi, Pakistan.
36	2	Herbion Pharmaceut Pvt Ltd, Dept Res & Dev, Karachi, Pakistan.
40	0	Ocean Univ China, Sch Med & Pharm, Key Lab Marine Drugs, Minist Educ China, Qingdao 266003, Peoples R China.
40	1	Qingdao Natl Lab Marine Sci & Technol, Lab Marine Drugs & Bioprod, Qingdao 266200, Peoples R China.
40	2	Univ Calif San Diego, Scripps Inst Oceanog, Ctr Marine Biotechnol & Biomed, La Jolla, CA 92093 USA.
40	3	Inst Invest Cient & Serv Alta Tecnol, Apartado 0816-02852, Clayton, Panama.
42	0	Dali Univ, Clin Med Coll, Dali, Yunnan, Peoples R China.
42	1	Dali Univ, Affiliated Hosp 1, Dali 671000, Yunnan, Peoples R China.
43	0	Bordeaux Univ Hosp CHU, Dept Paediat & Adult Congenital Cardiol, Ave Magellan, F-33600 Pessac, France.
43	1	Fdn Bordeaux Univ, Electrophysiol & Heart Modeling Inst, IHU Liryc, F-33600 Bordeaux, France.
43	2	INSERM, Ctr Rech Cardiothorac Bordeaux, U1045, F-33000 Bordeaux, France.
43	3	Aix Marseille Univ, CNRS, INSERM, URMITE, IHU Mediterranee Infect, Marseille, France.

1/ From data to spatial analysis

Common steps to fulfil the gap between textual information and spatial analysis are:

1/ Dealing with toponyms and finding coordinates

- Extract **toponyms**, if not done yet (e.g. NER, Gate);
- **Classify** toponyms (building names, neighbourhoods, postal codes, city names, region names...), if not done yet;
- Solve ambiguities and convert toponyms into **coordinates** (geocoding);

2/ Projection and mapping

- **Projection** of coordinates onto shapes;
- **Map aggregated locations** and (eventually) cross results with other variables.

1.1/ Geocode with CorText Manager

To classify elements in an address, we are using **LibPostal**: an address parser and normalizer, which is a multilingual, open source, Natural Language Processing based engine, to classify geographical elements in worldwide street addresses. LibPostal has been trained on OpenStreetMap.

(<https://github.com/openvenues/libpostal>)

Classification

Standardisation

Input	Output (may be multiple in libpostal)
One-hundred twenty E 96th St	120 east 96th street
C/ Ocho, P.I. 4	calle 8 polígono industrial 4
V XX Settembre, 20	via 20 settembre 20
Quatre vingt douze R. de l'Église	92 rue de l eglise
ул Каретный Ряд, д 4, строение 7	улица каретный ряд дом 4 строение 7
ул Каретный Ряд, д 4, строение 7	ulitsa karetnyy ryad dom 4 stroyeniye 7
Marktstraße 14	markt strasse 14

```

7. address_parser
-bash-3.2$ ./src/address_parser
Loading models...

Welcome to libpostal's address parser.

Type in any address to parse and print the result.

Special commands:
.exit to quit the program

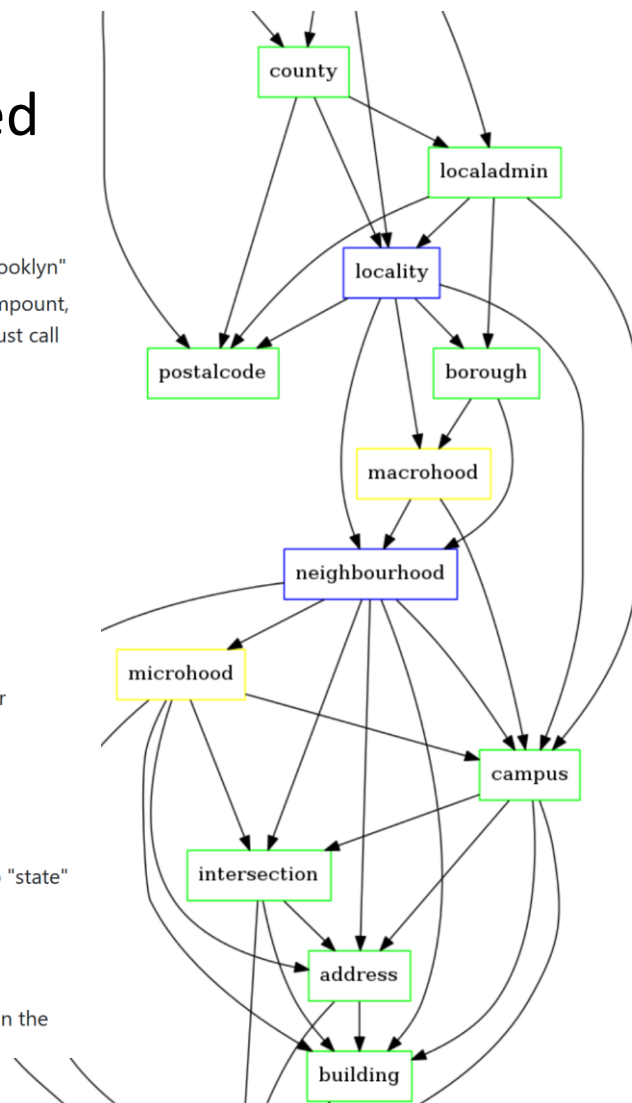
>

```

1.1/ Geocode with CorText Manager

Types of classified entity and ontology used

- **house:** venue name e.g. "Brooklyn Academy of Music", and building names e.g. "Empire State Building"
- **category:** for category queries like "restaurants", etc.
- **near:** phrases like "in", "near", etc. used after a category phrase to help with parsing queries like "restaurants in Brooklyn"
- **house_number:** usually refers to the external (street-facing) building number. In some countries this may be a compound, hyphenated number which also includes an apartment number, or a block number (a la Japan), but libpostal will just call it the house_number for simplicity.
- **road:** street name(s)
- **unit:** an apartment, unit, office, lot, or other secondary unit designator
- **level:** expressions indicating a floor number e.g. "3rd Floor", "Ground Floor", etc.
- **staircase:** numbered/lettered staircase
- **entrance:** numbered/lettered entrance
- **po_box:** post office box: typically found in non-physical (mail-only) addresses
- **postcode:** postal codes used for mail sorting
- **suburb:** usually an unofficial neighborhood name like "Harlem", "South Bronx", or "Crown Heights"
- **city_district:** these are usually boroughs or districts within a city that serve some official purpose e.g. "Brooklyn" or "Hackney" or "Bratislava IV"
- **city:** any human settlement including cities, towns, villages, hamlets, localities, etc.
- **island:** named islands e.g. "Maui"
- **state_district:** usually a second-level administrative division or county.
- **state:** a first-level administrative division. Scotland, Northern Ireland, Wales, and England in the UK are mapped to "state" as well (convention used in OSM, GeoPlanet, etc.)
- **country_region:** informal subdivision of a country without any political status
- **country:** sovereign nations and their dependent territories, anything with an [ISO-3166 code](https://www.iso.org/obp/ui/#iso:code:3166).
- **world_region:** currently only used for appending "West Indies" after the country name, a pattern frequently used in the English-speaking Caribbean e.a. "Jamaica. West Indies"

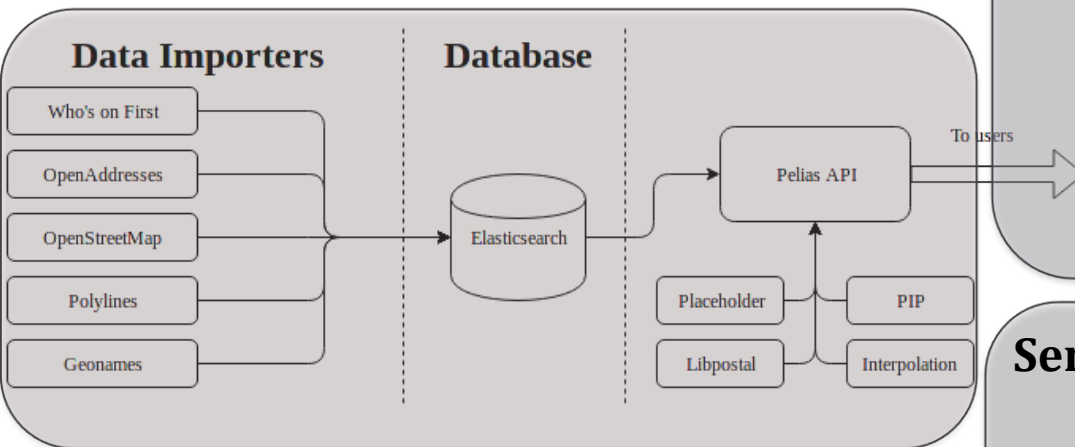


See the full picture here: <https://github.com/whosonfirst/whosonfirst-placetypes>

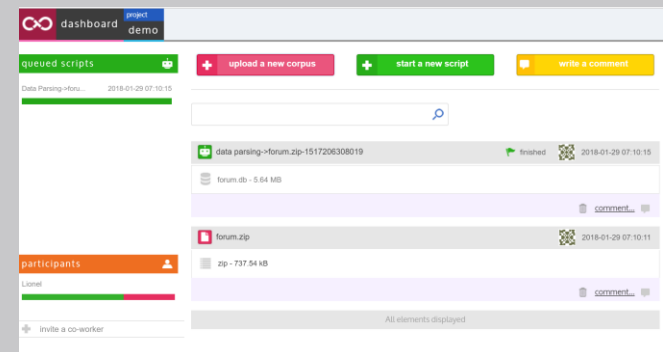
1.1/ Geocode with CorText Manager

We are providing two types of access: at the services (API) and application (CorText Manager) levels:

Backend (engine)



Web app access: through original methods developed inside CorText Manager



Services access (API):

- Toponym's classifier
- Reverse geocoding service
- Structured Geocoding (e.g. postal codes) service
- **Full addresses geocoding engine**
- ...

1.2/ Short insights on methods

For a given address, the geocoding engine try to solve **ambiguities*** by :

- Classifying toponyms and using the ontology;
- External variables (popularity criterium: number of inhabitants) to order which candidate is the best.

**same names different countries, same names and different geographical entities (region and city), acronyms, misspellings, vernacular names, multilingual names*

With **three options** to fit the needs of the researches conducted by our users : from meso scale (regional level) to smaller spaces (venue, building streets).

1.2/ Short insights on methods

Top scale filter

Geocoding methods

Filtering non geographical information Priority to the street level Priority to the city level No customization

Advanced settings

yes no

Output columns

label longitude latitude confidence accuracy layer city

region country iso3

Max number of addresses to process

Confidence threshold

start script

Chose the Geocoding method: "Filtering non geographical information" (country boundary defined but slow), "Priority to the street" (high quality but slow), "Priority to the city" (aggregated scale but slow), "No customization" (fast geocoding).

1.3/ Examples of results

- Located at different scales (points or centroids of shapes)
- Located with different types of geospatial entities (venue, locality, postal code, neighbourhood...)

address	label	accuracy	layer	city	region	country	confidence	longitude
Hamdard Univ, Fac Eastern Med & Surg, Karachi, Pakistan.	Hamdard University, Pakistan	point	venue	Karachi City	Sindh	Pakistan	0.89	67.008122
Herbion Pharmaceut Pvt Ltd, Dept Med Affairs & Training, Karachi, Pakistan.	Herbion Pakistan Pvt Ltd, Pakistan	point	venue	Karachi City	Sindh	Pakistan	0.9	67.097107
Herbion Pharmaceut Pvt Ltd, Dept Res & Dev, Karachi, Pakistan.	Herbion Pakistan Pvt Ltd, Pakistan	point	venue	Karachi City	Sindh	Pakistan	0.9	67.097107
Ocean Univ China, Sch Med & Pharm, Key Lab Marine Drugs, Minist Educ China, Qingdao, China		centroid	locality	Qingdao	Shandong	China	1	120.363522
Qingdao Natl Lab Marine Sci & Technol, Lab Marine Drugs & Bioprod, Qingdao, China		centroid	locality	Qingdao	Shandong	China	1	120.363522
Univ Calif San Diego, Scripps Inst Oceanog, Ctr Marine Biotechnol & Biomed, La Jolla, Anaheim, CA, USA		centroid	neighbourho	Anaheim	California	United State	0.6	-117.86889
Inst Invest Cient & Serv Alta Tecnol, Apartado 0816-02852, Clayton, Panama.	Panama City, Panama	centroid	locality	Panama City	Panama	Panama	1	-79.51973
Dali Univ, Clin Med Coll, Dali, Yunnan, Peoples R China.	Dali, China	centroid	locality	Dali	Yunnan	China	1	100.190243
Dali Univ, Affiliated Hosp 1, Dali 671000, Yunnan, Peoples R China.	Dali, China	centroid	locality	Dali	Yunnan	China	1	100.190243
Bordeaux Univ Hosp CHU, Dept Paediat & Adult Congenital Cardiol, Ave Magell: 33600, Pessac, France		centroid	postalcode	Pessac	Gironde	France	1	-0.676303
Fdn Bordeaux Univ, Electrophysiol & Heart Modeling Inst, IHU Liry, F-33600 Bo 33600, Pessac, France		centroid	postalcode	Pessac	Gironde	France	1	-0.676303
INSERM, Ctr Rech Cardiothorac Bordeaux, U1045, F-33000 Bordeaux, France.	33000, Bordeaux, France	centroid	postalcode	Bordeaux	Gironde	France	1	-0.587624
Aix Marseille Univ, CNRS, INSERM, URMITE, IHU Mediterranee Infect, Marseille, Marseille, France		centroid	locality	Marseille	Bouches-du-l	France	1	5.401581

1.4/ Refine directly online

Based on the confidence score and for not geocoded addresses you may want to refine the results.

<https://explorer.cortext.net/geo/refine/0b2a89bacc53907bce0aeb640f6649c5/2747510003210>

The screenshot displays the Cortext Explorer interface for refining geocoded locations. The main map shows a street grid in Brazzaville, Congo, with a red location pin on 'Ave. de France' near the 'UBA agence de Poto-Poto'. A search bar at the top right contains the text 'Brazzaville, Rep Congo.' and a 'Search' button. Below the search bar, a 'Search results' panel lists several entries:

- Brazzaville, Congo locality
- Congo
- Africa
- Brazzaville Department, Congo region
- Congo
- Africa
- Brazzaville Department, Congo region
- Congo
- Africa
- Brazzaville, Congo county
- Congo

In the bottom left corner, a 'Coordinates' panel shows the following information:

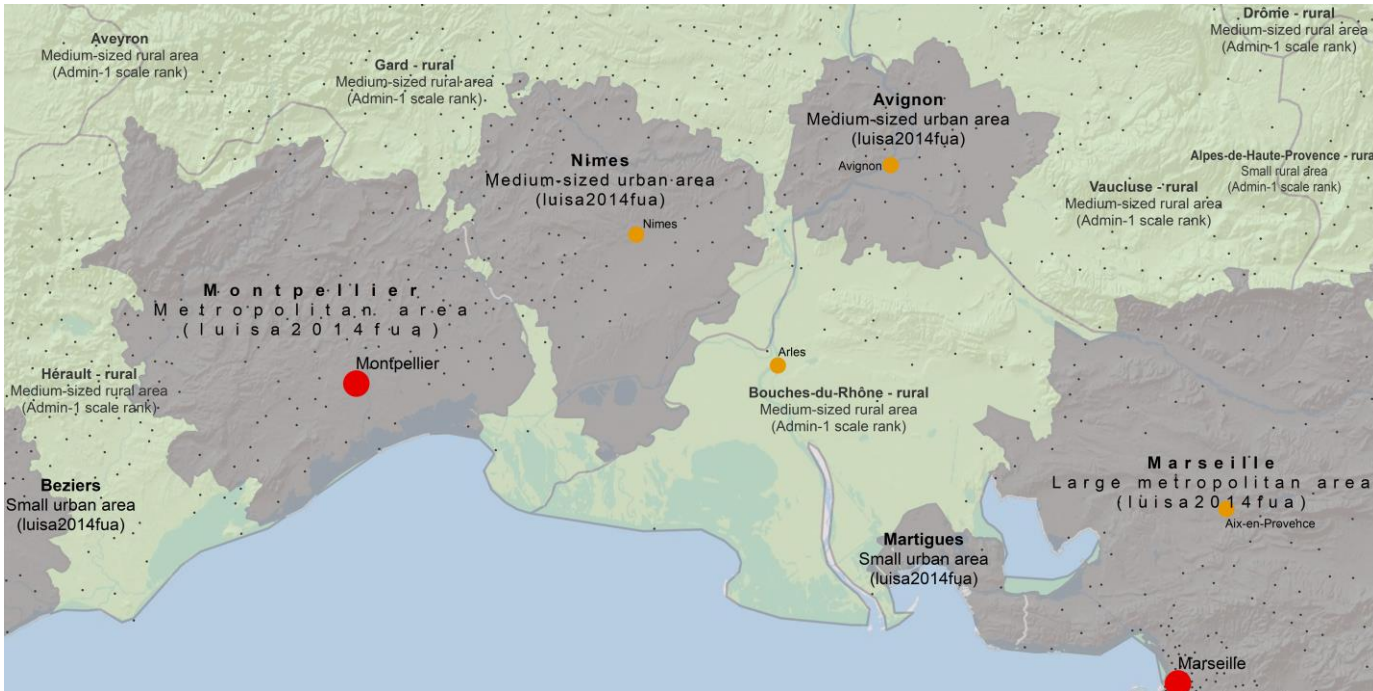
- Coordinates: [15.28318, -4.26613]
- City/Place: Brazzaville
- Region: Brazzaville Department
- Country: Congo
- A 'Save' button is located at the bottom of this panel.

The map interface includes a zoom control on the top left and a title bar that reads 'Refine geocoded locations'.

2.1/ Projection onto geospatial shapes, short insight on methods

Aggregate longitudes and latitudes in Urban & Rural Areas (URA) or NUTS3...

- urban areas (> 50 000 inhabitants);
- “rural areas” at the regional scale (regional shapes, excluding URA shapes).
- Worldwide coverage: 4 200 Urban areas and 4 428 Rural areas



2.1/ Projection onto geospatial shapes, short insight on methods

SCRIPT PARAMETERS

Mapping and aggregation Third party basemaps Initial map view

Define a custom label for the map's legend

Use a custom longitude|latitude field

yes no

Assign unclassified points to the nearest area

yes no

Maximum distance in km

Two-pass URA

yes no

Project a second field onto the map

yes no

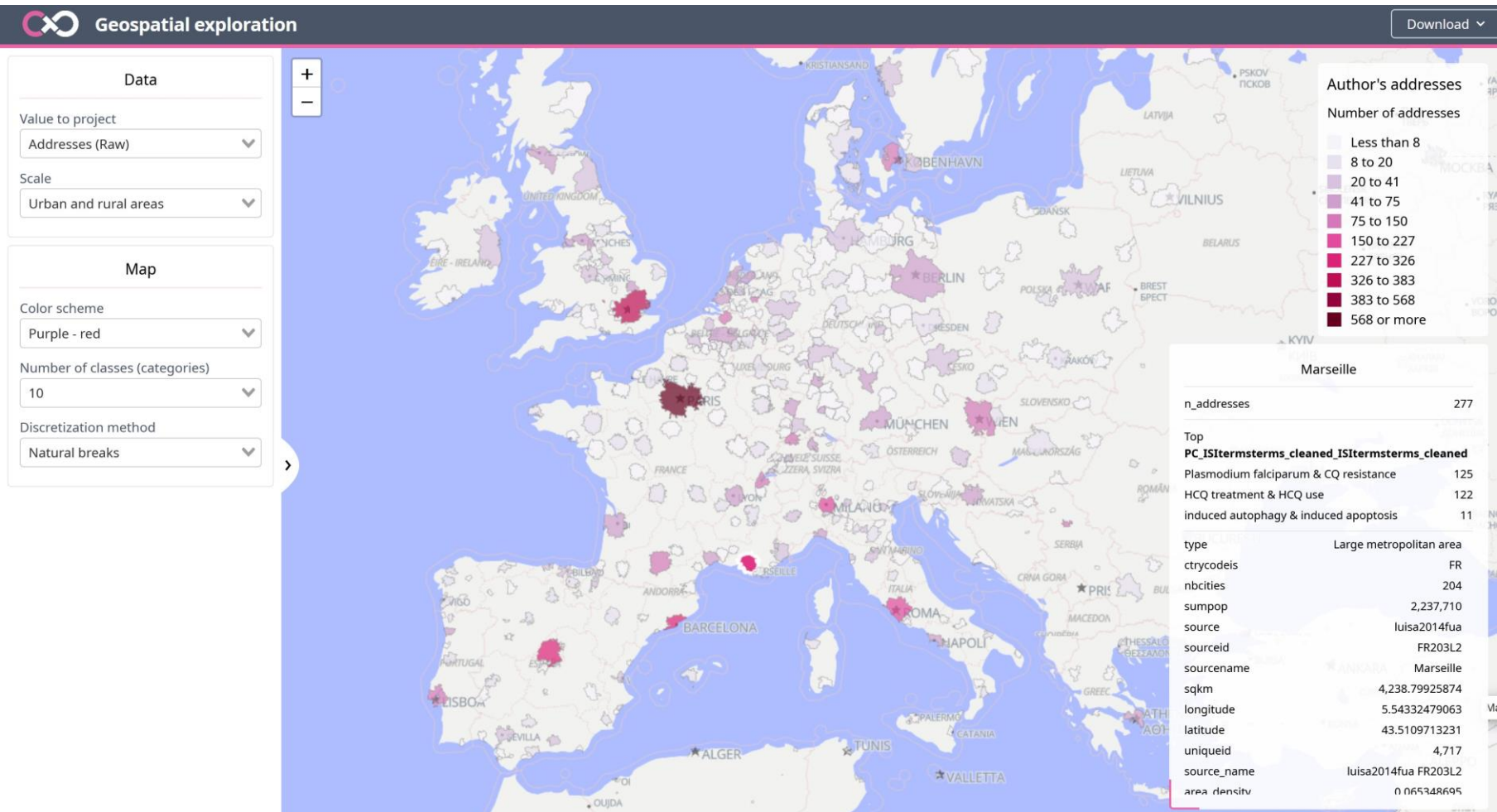
Choose (than counting)

yes no

start script

2.1/ Visualized URA and NUTS3 (maps)

<https://geomapping.cortext.net/#/map/c0801bef413b5d7bb44dcc17149f1bff>



3.2/ Cross geospatial and semantic landscape: profiling

plasmodium falciparum & cq resistance - salience (one vs all, symetrized deviation rate)2014_2020

lupus and COVID19

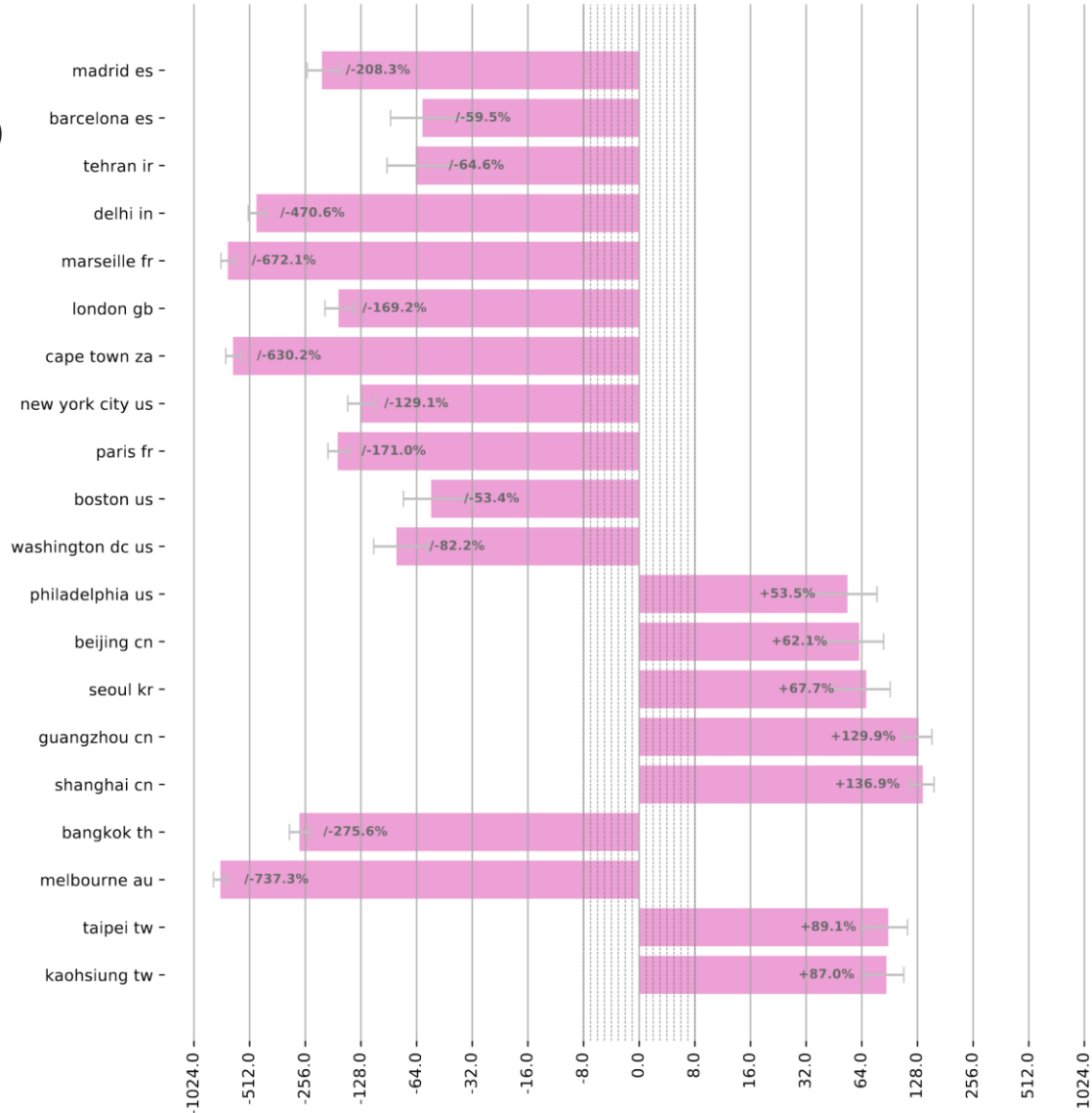
(Plasmodium falciparum & CQ resistance)



3.2/ Cross geospatial and semantic landscape: profiling

induced autophagy & induced apoptosis - salience (one vs all, symetrized deviation rate)2014_2020

cancers
(Induced autophagy
& induced apoptosis)



3.2/ Cross geospatial and semantic landscape: profiling

hcq treatment & hcq use - salience (one vs all, symetrized deviation rate)2014_2020

HCQ treatment
& HCQ use
malaria

