CorText Manager et espaces géographiques

[Rechercher les lieux, Les lieux de la recherche]

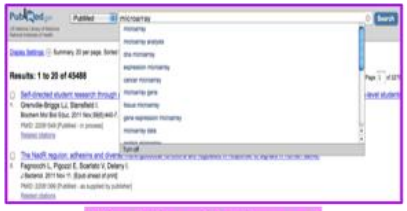
Lionel Villard, 14/01/2021

CorText Manager : variété des sources de données

scientific productions

Web Of Science ISI

Microsoft Academic Search



Medline Pubmed

. . .

specific databases

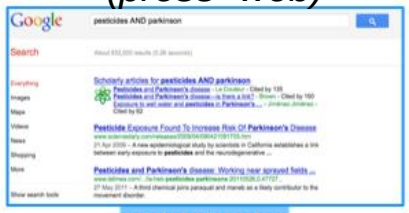
rare disease database

projects database

clinical trials database

. . .

media productions (press+web)



web crawler

Factiva, press articles archive

online forums

. . .

Capacity to collect, parse and handle corpora from various arenas

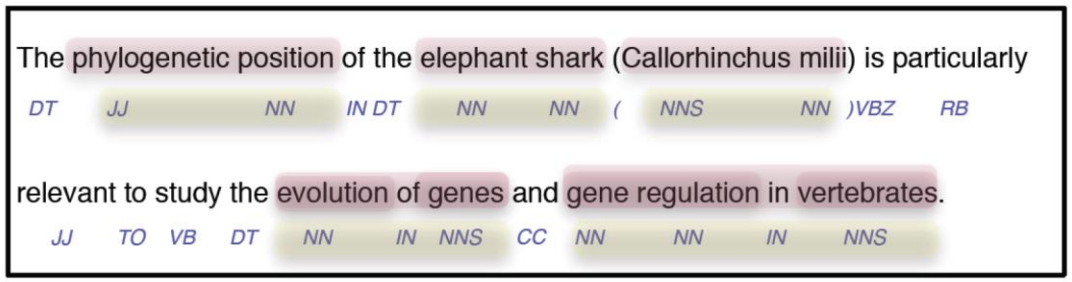
CorText Manager : le paysage semantique

Le Manager permet déjà de répondre à :

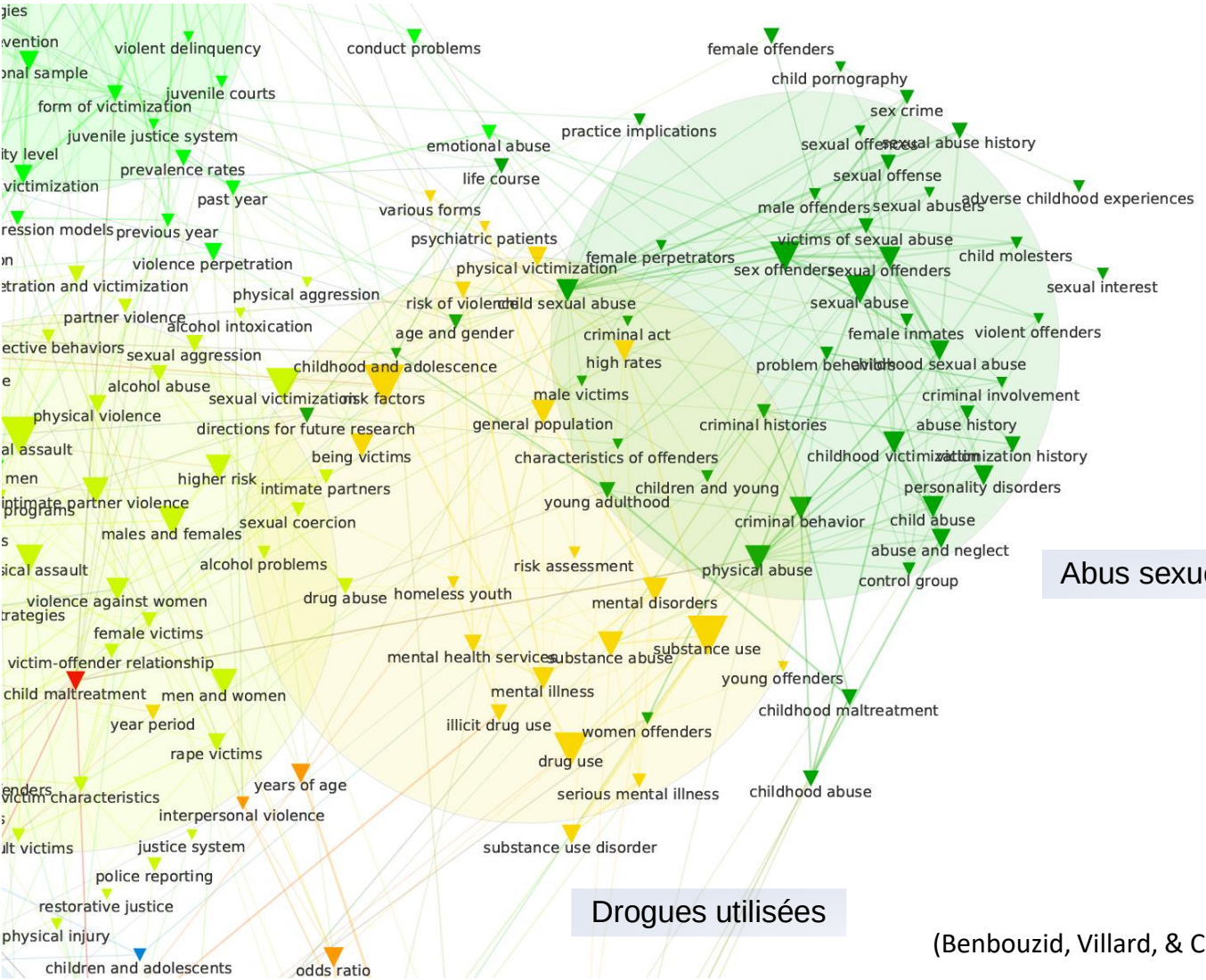
- **Quoi** : semantic landscapes, avec une analyse des contenus textuels;

Le paysage socio-sémantique est reconstruit notamment par une EXTRACTION TERMINOLOGIQUE qui s'effectue en plusieurs étapes :

- Étiquetage morpho-syntaxique : associer aux mots des textes disponibles dans les corpus les informations grammaticales (le genre, le nombre…)
- Extraction des groupes nominaux (noms et adjectifs...)



CorText Manager : le paysage semantique

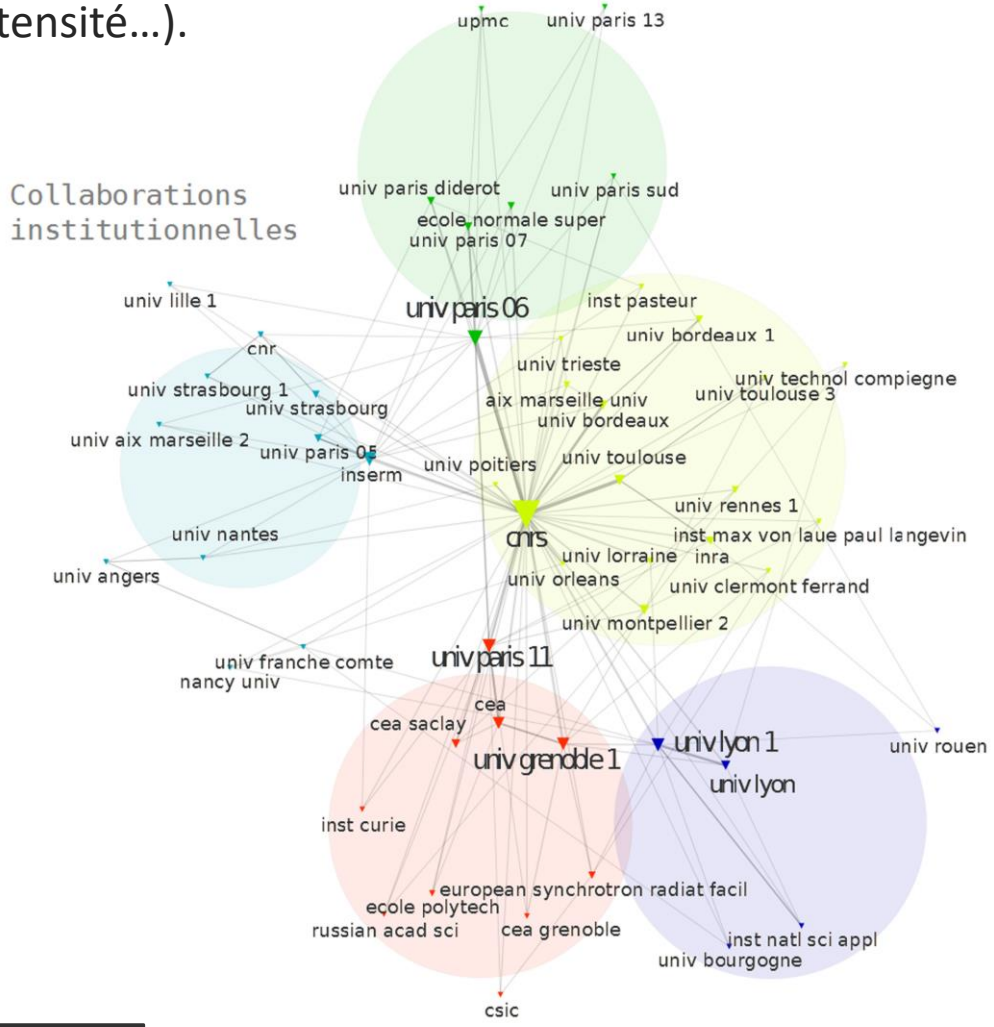


Articles sur les enquêtes portant sur la victimisation

(Benbouzid, Villard, & Cointet, 2016)

CorText Manager : l'analyse de réseaux sociaux

- **Qui** : Social Network Analysis, permet de savoir qui portent les dynamiques étudiées, et dans des quelles positions (positions dans le réseau, communauté, intensité…).



Quelles sont les positions relatives des **acteurs** et leurs principales collaborations ?

Nombre de publications scientifiques partagées entre les acteurs (top50)

Position centrale (connecteur) du CNRS. Paris 11 également fortement connectée.

Des universités importantes qui ont des collaborations privilégiées.

Collaborations étroites entre les universités Lyonnaises, Parisiennes et Grenobloises.

Un groupe d'universités de seconde division dans les nanobiotechnologies (Nantes, Angers, Strasbourg…) mais qui collaborent pourtant beaucoup entre elles.

Pourquoi l'espace géographique comme outil dans le CorText Manager ?

La dimension temporelle :

- **Quand** : est à associer aux autres dimensions d'analyse.

Les deux méthodes possibles dans CorText Manager pour travailler la dimension géographique :

- Full text **toponyme detection :** NER (spacy);
- **Enrichissement à partir de dictionnaires :** avec des données semi-structurées, le corpus peut être alors enrichi avec des informations extérieures comme, par exemple, la liste des pays par continent.

Il manquait :

- **Où** : cartographie des lieux des documents (contenus, personnes ou organisations) et projection dans leurs espaces géographiques (villes, régions…)

Pourquoi l'espace géographique comme outil dans le CorText Manager ?

70% of text documents contain place name references (Hill, 2006), but with a lack of:

- variables on spaces that can be compute directly in statistical analysis (e.g., longitude and latitude);
- freely and open services accessible for research purposes in a context of large dataset treatments (e.g., that are able to convert addresses or toponyms from full text into geographical maps).

And different types of geographical information:

- Images or pictures of places
- **Toponyms in full text**
- **Addresses**
- Structured metadata
- **Geographical coordinates**

(Hill, 2006)

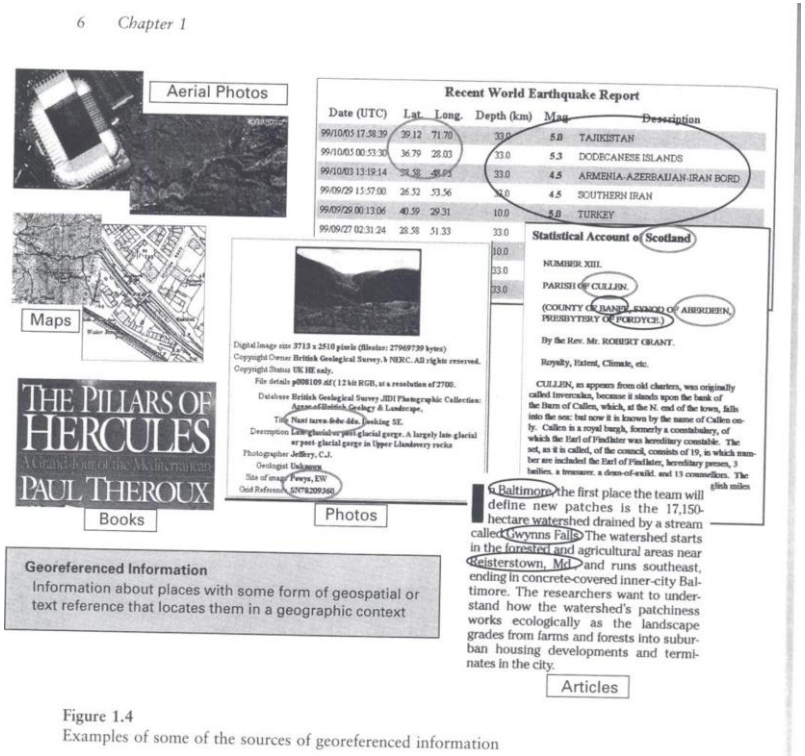


Figure 1.4
Examples of some of the sources of georeferenced information

➢ Need to build our own open source geocoding engine that fit for large datasets

Background: easily bringing geography in S&T datasets

Adding a layer of geographical information that can be computed in order to conduce spatial analysis, and to display results onto maps.

Common problems with addresses:

- Different formats that rely on national postal service (and data providers), that largely vary;
- Non-geographic information (building names, lab names, person names…), that have ambiguities and could be multi-located;
- Ambiguous toponyms;
- Costs.

```
AB Background: This study measured how myelodysplastic syndrome (MDS) patients value transfusion ir
   Methods: 47 MDS patients were interviewed, US (n = 8), France (n = 9), Germany (n = 9) and the U
   Discussion: The mean age was 67 years (range: 29-83); 45% male, 70% retired; 40% had secondary/h
   Conclusion: Patients value TI, suggesting an important role for new treatments aiming to achieve
C1 [Szende, Agota; Schaefer, Caroline] Covance, Leeds, W Yorkshire, England.
   [Schaefer, Caroline; Goss, Thomas F.] Covance, Gaithersburg, MD USA.
   [Heptinstall, Kathy] MDS Fdn, Crosswicks, NJ USA.
   [Knight, Robert] Celgene Corp, Summit, NJ USA.
   [Luebbert, Michael; Deschler, Barbara] Univ Freiburg, Freiburg, Germany.
   [Fenaux, Pierre] Univ Paris 13, Hop Avicenne, Bobigny, France.
   [Mufti, Ghulam J.] Kings Coll Hosp London, London, England.
   [Killick, Sally] Royal Bournemouth Hosp Fdn Trust, Bournemouth, Dorset, England.
   [List, Alan F.] Univ S Florida, Tampa, FL USA.
```

Examples of raw

WOS export

```
odeling a
LA English
DT Article
DE EATING DISORDERS; ONLINE HEALTH COMMUNITIES; E-HEALTH; SOCIAL INFLUENCE
   MODEL; AGENT-BASED COMPUTER SIMULATION
ID OPINION; NETWORK
AB This article presents an agent-based model of a health-related Internet forum. If recent literature demonstrates the re
C1 [Casilli, Antonio A.] Telecom ParisTech, F-75013 Paris, France.
   [Casilli, Antonio A.] EHESS, CNRS, Ctr Edgar Morin, F-75009 Paris, France.
   [Rouchier, Juliette] Aix Marseille Univ, Fac Econ & Gest, Grp Rech Econ Quantitat GREQAM, F-13236 Marseille, France.
   [Tubaro, Paola] Univ Greenwich, Old Royal Naval Coll, Dept Int Business & Econ, London SE10 9LS, England.
```

Geocoding

Bring a new layer of computer-processable information to understand the spatial dynamics of user's datasets.

RISIS²

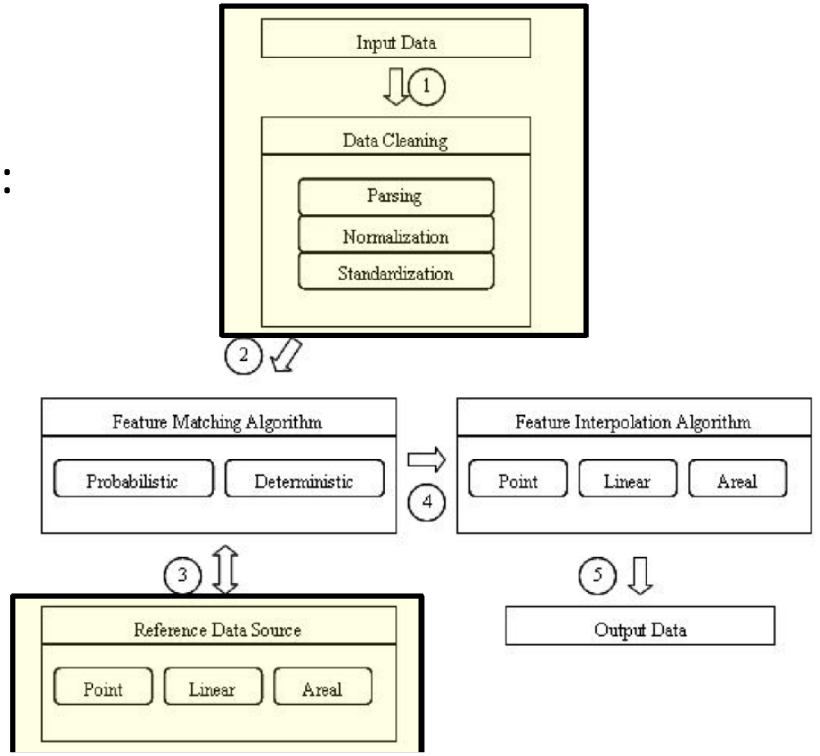
Research infrastructure for research and innovation policy studies

1/ Geocoding: global overview

The typical steps needed to move from documents, and raw addresses related to documents, to maps are:

- **Normalization**: cleaning, parsing (components are more essential than others, like city names or postal code), normalisation (same geographical object with the same name, like abbreviations or ambiguities);
- **Matching**: comparing the normalized address with one or several reference databases.

Full workflow can be illustrated as follow:



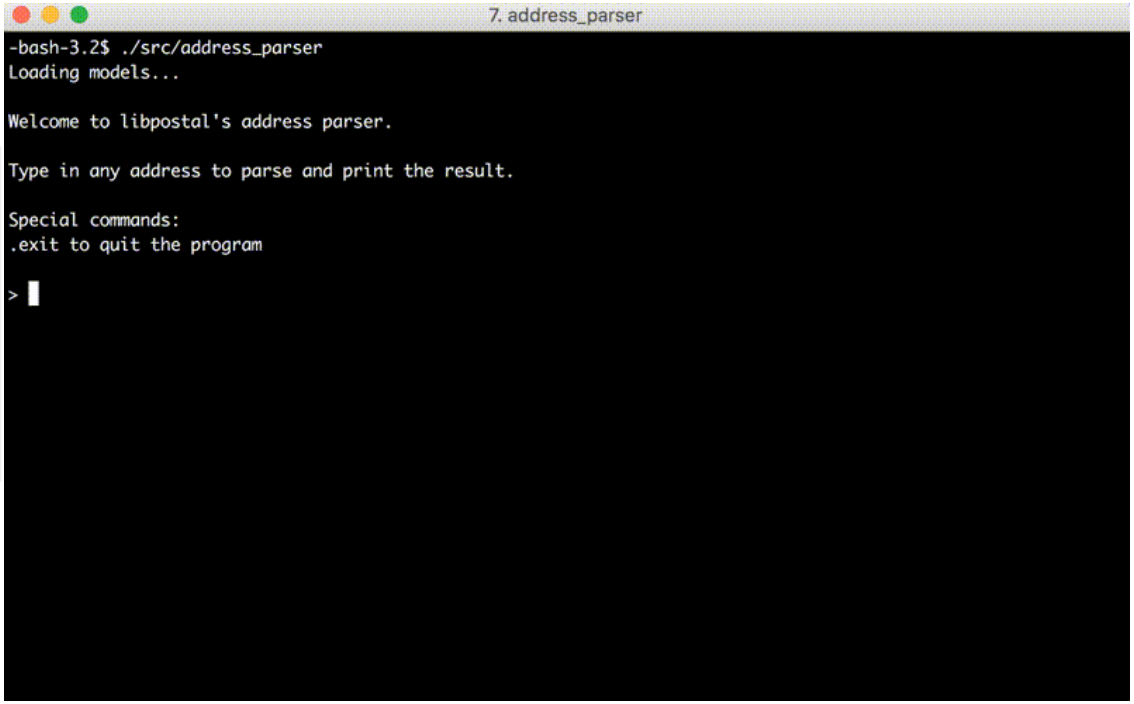
(Goldberg, 2008)

1/ Classifying geo-objects and toponym ambiguities

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 каретныи ряд дом 4 строение 7
ул Каретный Ряд, д 4, строение 7	ulitsa karetnyy ryad dom 4 stroyeniye 7
Marktstraße 14	markt strasse 14

1/ Classifying geo-objects and toponym ambiguities

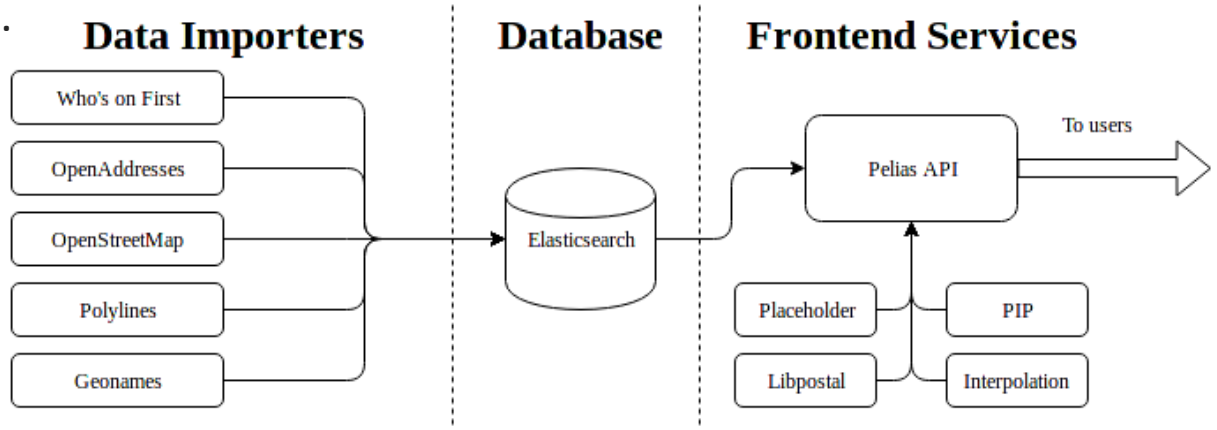
Types of object identified (classified/tagged, but also used in the matching step to solve ambiguities)

- **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 compount, 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.
- **world_region**: currently only used for appending “West Indies” after the country name, a pattern frequently used in the English-speaking Caribbean e.g. “Jamaica, West Indies”

1.1/ Pelias: data sources and sizes

Open sources (external) reference databases:

- enable geocoding **large datasets**;
- enhancing datasets with external links and new layers of information (of different types).



Data Source	Description
Who's on First	Who's on First is an open-data directory of worldwide administrative places. Created by *Mapzen*.
Geonames	Geonames is an aggregation of many authoritative and non-authoritative datasets. It contains information on everything from country borders to airport names to geographical features.
OpenAddresses	OpenAddresses is a collection of over 300 million addresses around the world. Data in OpenAddresses only comes from national, state, and local governments, so this data is highly authoritative.
OpenStreetMap	OpenStreetMap is a community-driven, editable map of the world.
Tiger	US Census Bureau TIGER data
Transit	Public transit feeds from around the world
Polylines	Road network info derived from OSM

Dataset	Size
LibPostal	~ 2.2GB
WOF with venues	~ 120GB
Geonames world	~ 3GB
OA global	~ 35GB
OSM planet	~ 34GB
Polylines	~ 2.3GB
Elasticsearch index	~ 330GB
All disk usage	~ 530 GB (200 GB + 330 GB)
Recommended free space	1 TB

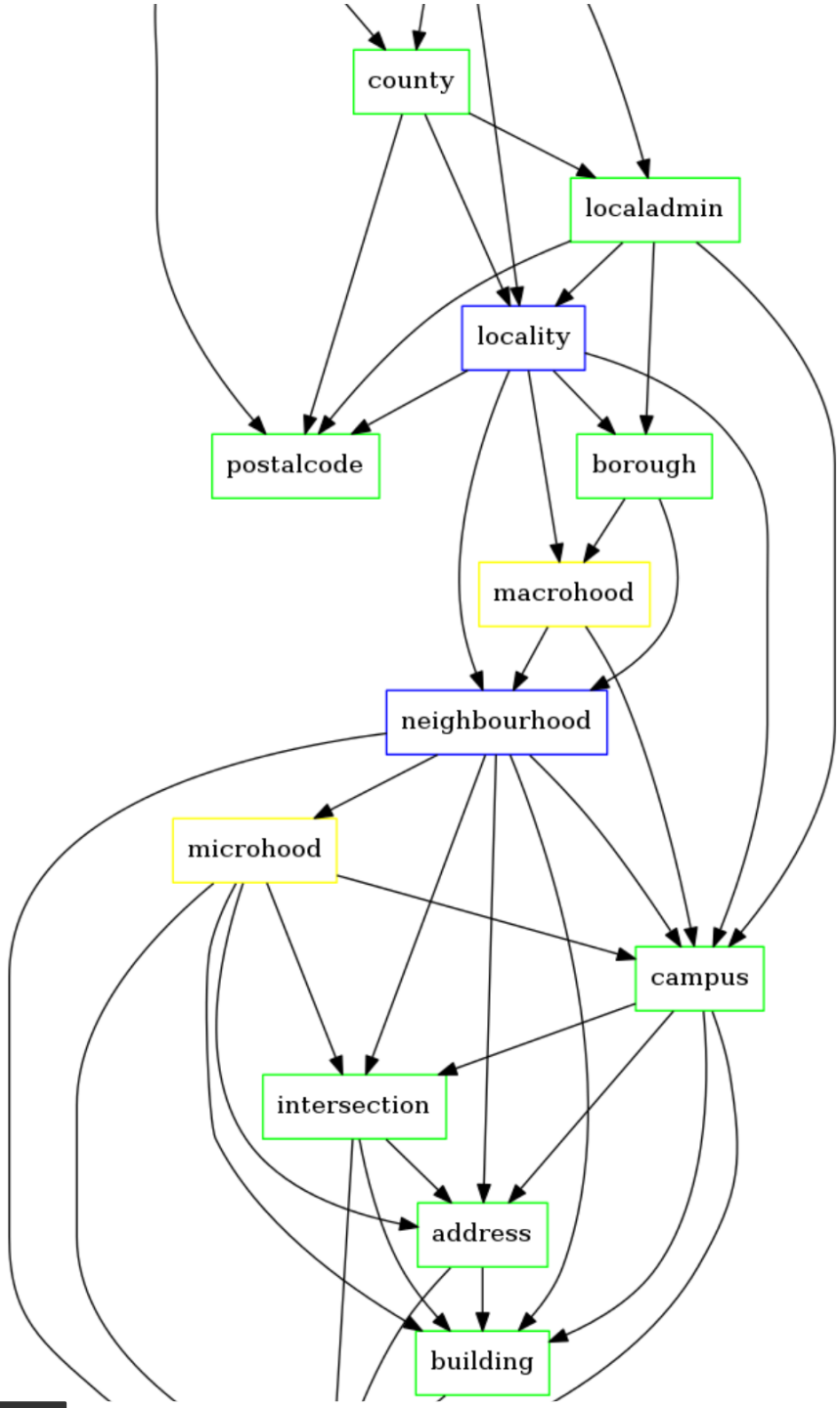
1.1/ Geo-objects ontology

WhosOnFirst is the glue of all (external) reference databases.

Based on the links between objects, this ontology is used:

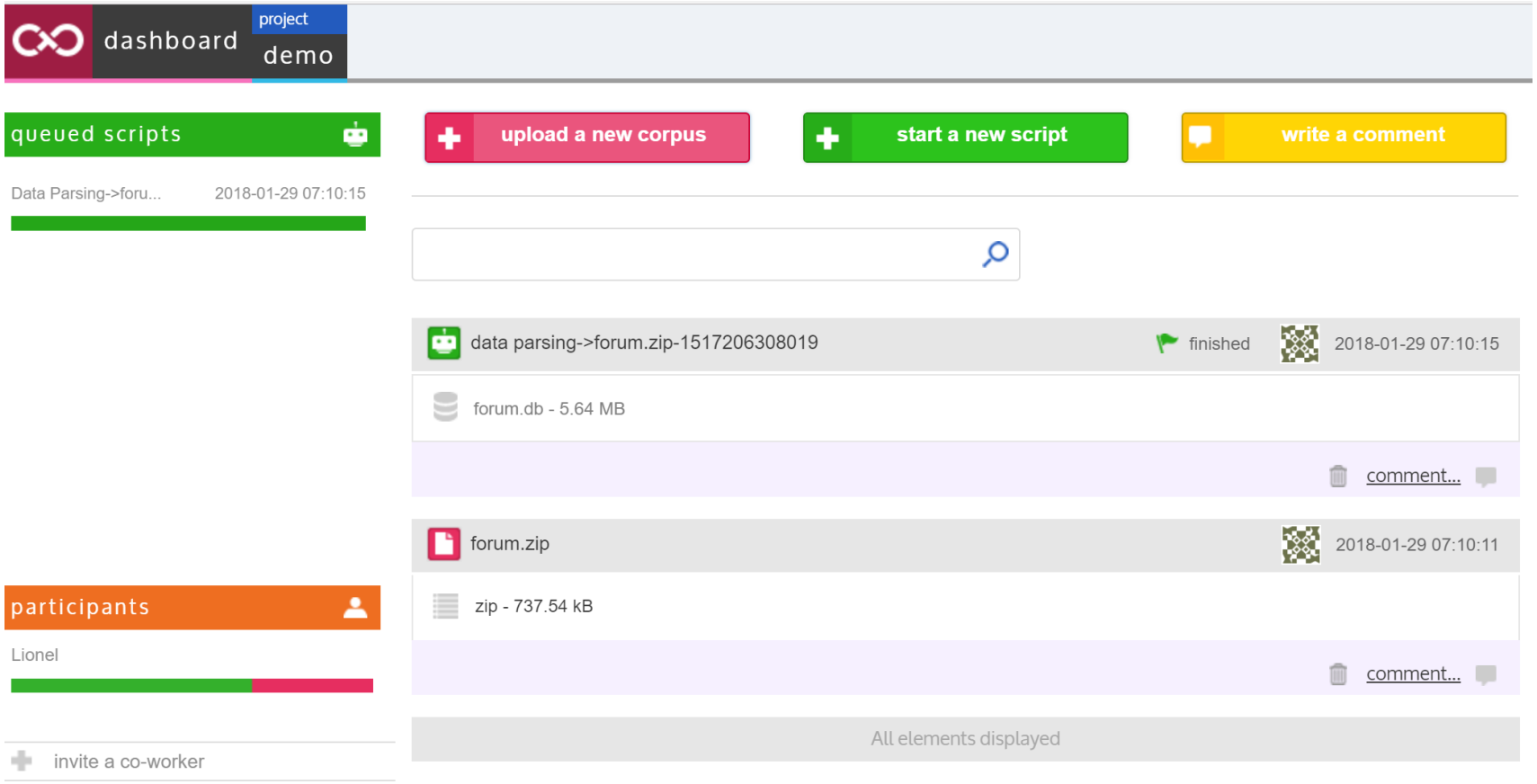
- to compare the classified (tagged) objects in an address,
- with the classification and names contained in the reference databases,
- and to retrieve the best candidates.

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



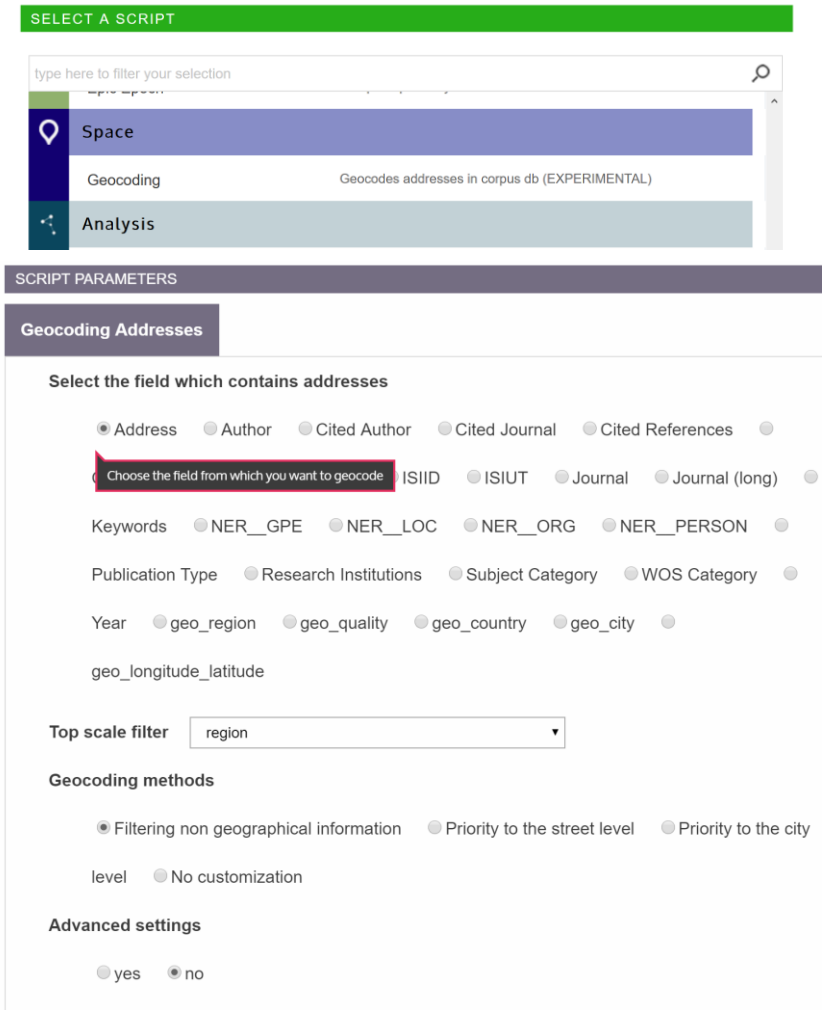
[Demo]

https://managerv2.cortext.net/project/1447



[Demo]

https://managerv2.cortext.net/project/1447



1.2/ Results

We are promoting this two steps approach, with four options to fit the needs of the researches conducted by our users: from macro scales (e.g. regional level) to smaller geographical spaces (e.g. building names or neighborhood). Within a given address, tagged toponyms with the ontology are used to calibrate the results. For the remaining ambiguities (e.g. when having “Paris” without any other information), the geocoding engine uses external variables (popularity criterium, like number of inhabitants) to decide which candidate is the best (e.g. and to choose “Paris, France” instead “Paris, Texas, USA”).

Meso geocoding:

- **Filtering organisation names**: organisation names, street names, person names and postal boxes are removed in order to reduce ambiguity (e.g. for multi-located compagnies or laboratories), and to retrieved more aggregated geographical information. Addresses can be located from the postal codes scale (sub-city scale) to urban areas and metropoles (sub regional scale, as for county). Depending of the tagged elements in the address, the geocoding engine will decide which candidate and scale are the best.
- **Priority on city scale**: tagged addresses will be searched in the specific sub-area (meso area as regions or counties) and retrieved coordinates when it is possible at the city scale. It tends to reduce the variety of geographical objects retrieved and to narrow the scales on centroid coordinates of cities.

These two approaches produce aggregated coordinates (shapes located by a centroid), with a less spread spatial distribution.

address proposed	found	longitude	latitude	confidence	accuracy	layer	city	region	country	iso3
Ctr Hosp Univ, F-37044 Tours, France	Tours, France	0.689776	47.386419	1.00	centroid	locality	Tours	Indre-et-Loire	France	FRA
ENSIC, F-54001 Nancy, France	Nancy, France	6.178289	48.690303	1.00	centroid	locality	Nancy	Meurthe-et-Moselle	France	FRA
Univ Bretagne Sud, F-56100 Lorient, France	Lorient, France	-3.377392	47.750565	1.00	centroid	locality	Lorient	Morbihan	France	FRA
Thales Naval Nederland, NL-7550 GD Hengelo, Netherlands	Hengelo, Netherlands	6.793795	52.260996	1.00	centroid	locality	Hengelo	Overijssel	Netherlands	NLD
IRCCyN, F-44321 Nantes 03, France	Nantes, France	-1.554004	47.228973	1.00	centroid	locality	Nantes	Loire-Atlantique	France	FRA
CNRS, F-75700 Paris, France	Paris, France	2.3488	48.85341	1.00	centroid	locality	Paris	Paris	France	FRA

1.2/ Results

Full addresses geocoding:

- **Priority on street scale**: tagged addresses will be searched in the specific sub-area and retrieved with a very fine-grained toponyms or building names (and POI). Street names, building names are prioritized :
  - Need a well and uniform full addresses coverage (with street names and/or building names);
  - Fit when you want to follow precise intra-urban locale dynamics;
  - Produce a spread spatial distribution of the coordinates.
- **No customisation**: full addresses are sent without any customisation (no pre-processing step, no toponym filters, non prioritisation of the scale) to the geocoding engine and let it decide which geographical object is the best;

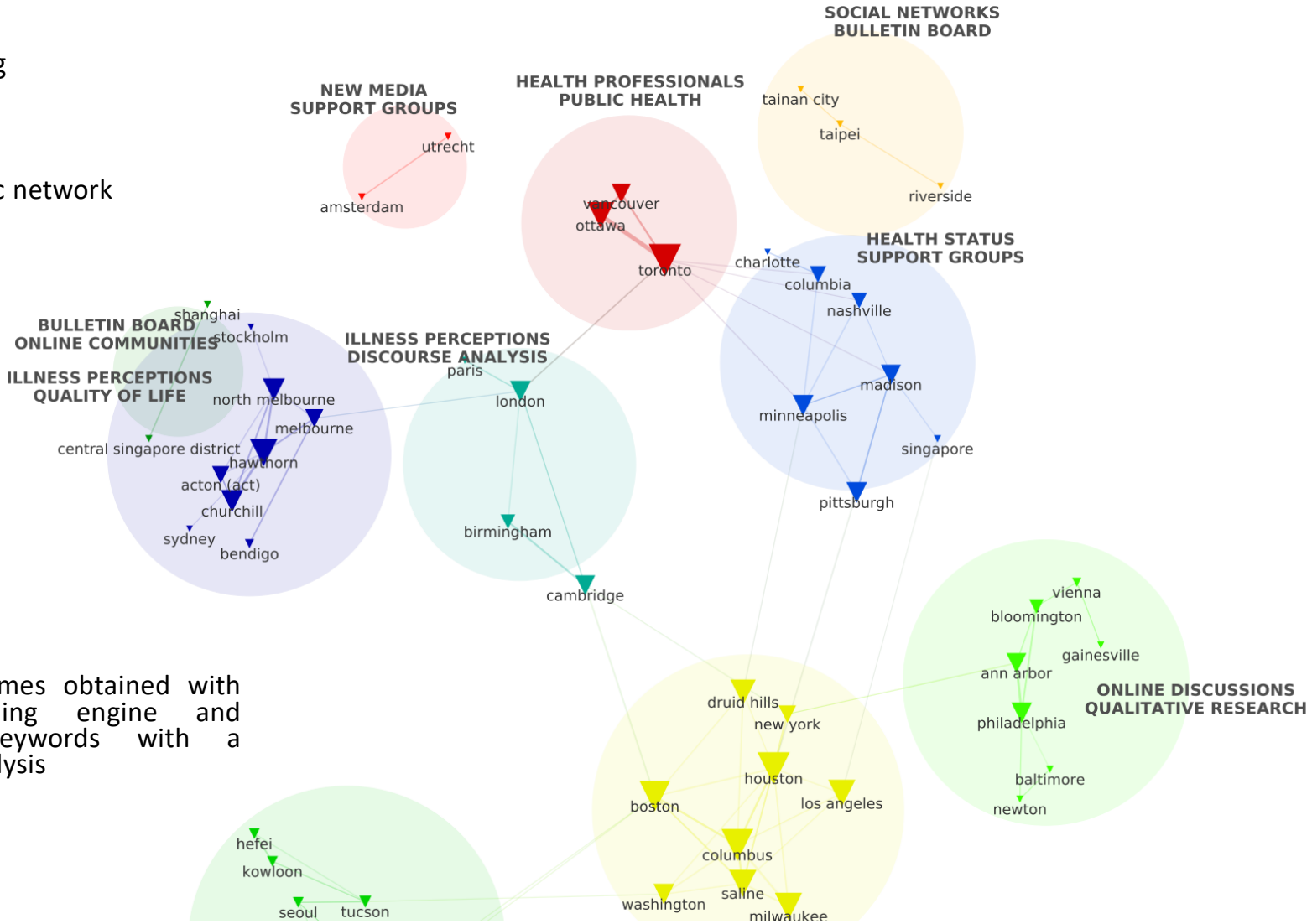
address proposed	found	longitude	latitude	confidence	accuracy	layer	city	region	country	iso3
Ctr Hosp Univ, F-37044 Tours, France	Hôpital Clocheville, Tours, France	0.680499	47.39007	1.00	point	venue	Tours	Indre-et-Loire	France	FRA
ENSIC, F-54001 Nancy, France	C.p.i.c. - Ensic, Nancy, France	6.183405	48.689548	0.67	point	venue	Nancy	Meurthe-et-Moselle	France	FRA
Univ Bretagne Sud, F-56100 Lorient, France	Université de Bretagne Sud, Lorient, France	-3.385535	47.742947	0.89	point	venue	Lorient	Morbihan	France	FRA
Thales Naval Nederland, NL-7550 GD Hengelo, Netherlands	Thales Nederland, Hengelo, Netherlands	6.772381	52.246779	0.73	point	venue	Hengelo	Overijssel	Netherlands	NLD
IRCCyN, F-44321 Nantes 03, France	IRCCyN, Nantes, France	-1.547316	47.250124	1.00	point	venue	Nantes	Loire-Atlantique	France	FRA
CNRS, F-75700 Paris, France	CNRS, Paris, France	2.264014	48.847535	1.00	point	venue	Paris	Paris	France	FRA
ST Microelect, F-38019 Grenoble, France	S.t. Microélectronics S.a., Grenoble, France	5.732003	45.193989	1.00	point	venue	Grenoble	Isère	France	FRA

1.3/ Integration with network analysis

Small teasing

Geo-Semantic network

with city names obtained with the geocoding engine and extracted keywords with a semantic analysis



Top 75 nodes (geo_city) by raw values Top 3 Neighbours, Top2 Terms Chi2 (WOS dataset on Internet forums)

Mapping

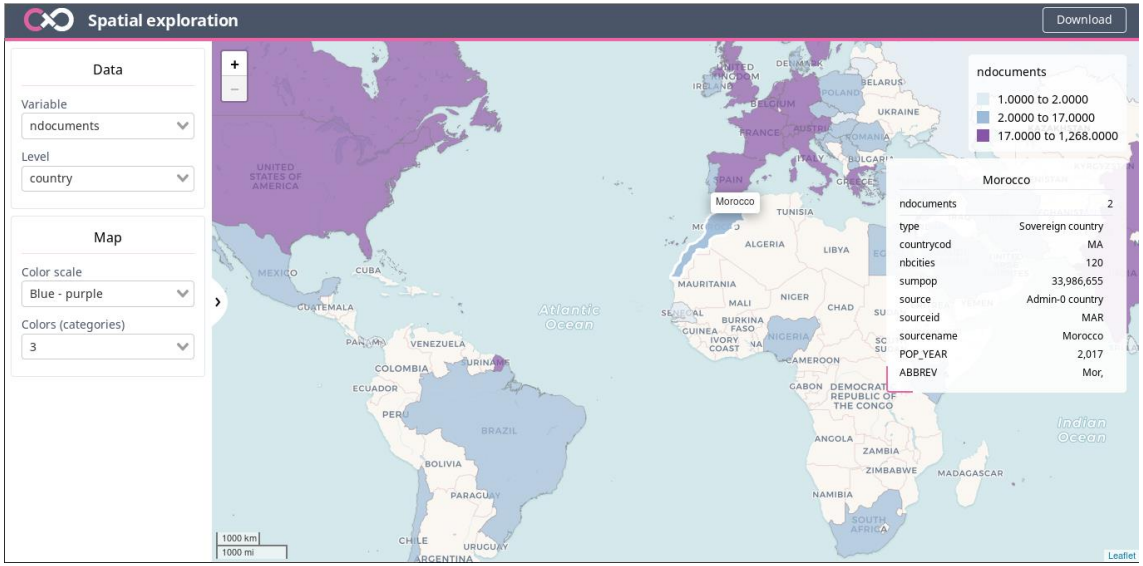
Projection of the geocoded data in geographical spaces, to allow their analyzes and to cross geographical spaces with the other analytical dimensions already available in CorText Manager (semantic or actors).

2.1/ Cartographie : projection des coordonnées dans des espaces géographiques

Short term goal (prototype)

A simple choropleth map to explore where are located the user's addresses aggregated at a given scale :

- urban,
- regional,
- a combinaison of urban and regional shapes (rurban),
- country



https://lu1sd4.github.io/map-visualization-demo

2.1/ Cartographie : projection des coordonnées dans des espaces géographiques

Europe

- **Functional Urban Areas (FUA)**: Kompil, Mert; Lavalle, Carlo; Aurambout, Jean-Philippe (2015): UI - Boundaries for the functional urban areas (LUISA Platform REF2014). European Commission, Joint Research Centre (JRC);
- Switzerland, Island and Norway: **Urban Audit** 2011-2014 (2015);

US

- US Census Bureau, Cartographic Boundary Shapefiles - **Urban Areas**, 2017;

Medium and large FUA for countries outside Europe from OECD (KR, JP, AU, CA…)

- Definition of **Functional Urban Areas** (FUA) for the OECD metropolitan database September (2013);

China

- **Chinese Urban Area** (Beijing City Lab, 2012): Long Y, Shen Y, 2014, Mapping parcel-level urban areas for a large geographical area;

Rest of the world

- Natural Earth (MODIS 2003 method) based

  on landscape images to capture **dense light areas** on earth.

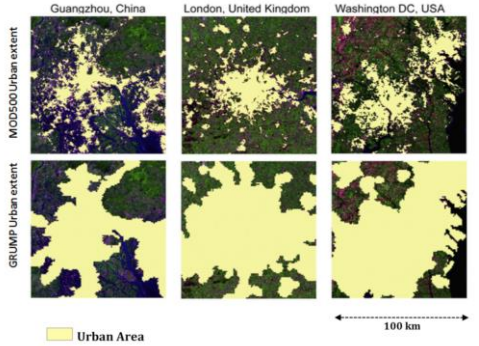
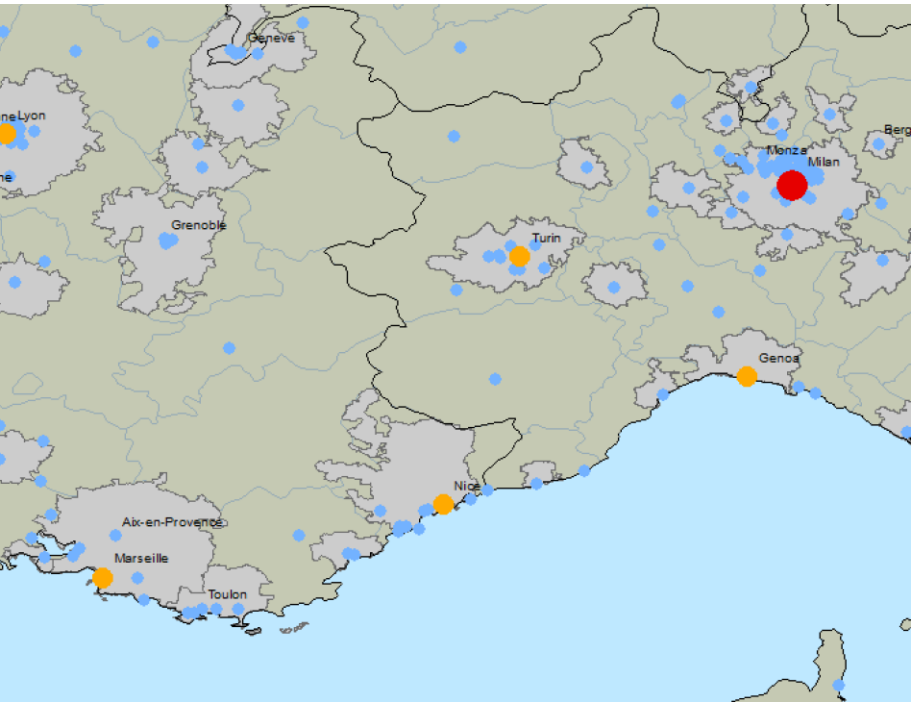
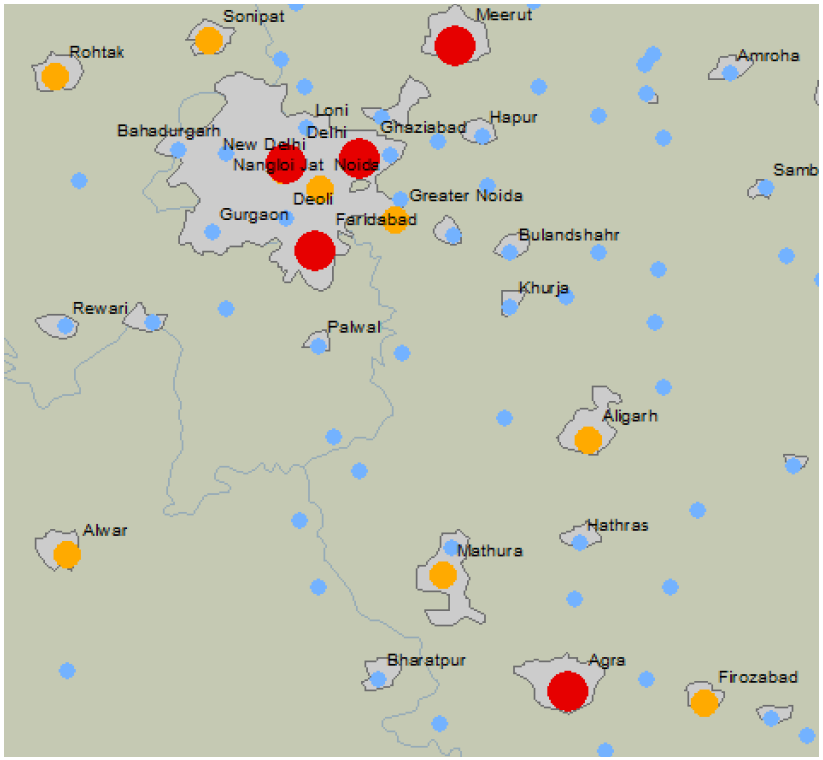


Figure 1. Comparison of MOD500 and GRUMP masks for three metropolitan areas: Guangzhou, China; London, United Kingdom; and Washington D.C.-Baltimore, U.S.A. [Schneider et al. 2010].

2.1/ Cartographie : projection des coordonnées dans des espaces géographiques



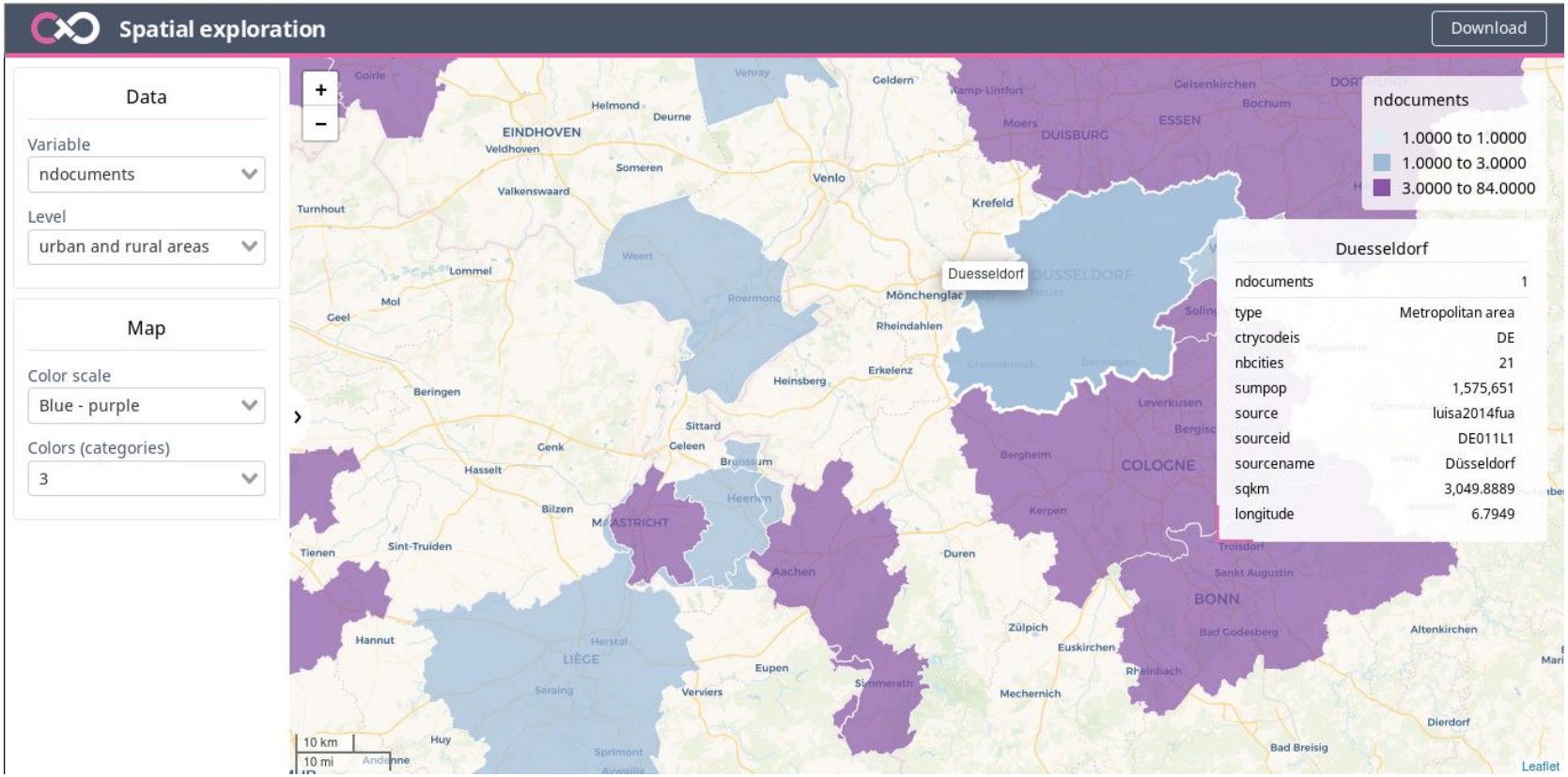
France and Italie (FUA)



India, around New Delhi (dense light areas)

2.1/ Cartographie : projection des coordonnées dans des espaces géographiques

Short term goal (prototype)



https://lu1sd4.github.io/map-visualization-demo

2.1/ Cartographie : projection des coordonnées dans des espaces géographiques

Paris	Large metropolitan area	FR	1321	1167174	luisa2014fua	FR001L1	Paris	27836,483923	2,387258	48,845055
Marseille	Large metropolitan area	FR	204	2237710	luisa2014fua	FR203L2	Marseille	8027,588114	5,543325	43,510971
Lyon	Large metropolitan area	FR	296	1761319	luisa2014fua	FR003L2	Lyon	7513,175753	4,90652	45,756371
Lille	Metropolitan area	FR	173	1365148	luisa2014fua	FR009L2	Lille	3569,17427	2,988581	50,607707
Toulouse	Metropolitan area	FR	424	1118356	luisa2014fua	FR004L2	Toulouse	9951,531378	1,409504	43,51525
Bordeaux	Metropolitan area	FR	237	1061198	luisa2014fua	FR007L2	Bordeaux	10970,728706	-0,691193	44,780893
Nantes	Metropolitan area	FR	102	819249	luisa2014fua	FR008L2	Nantes	6844,917653	-1,587198	47,2513
Nice	Metropolitan area	FR	126	813868	luisa2014fua	FR205L2	Nice	5947,877674	7,02709	43,92805
Strasbourg	Metropolitan area	FR	240	745531	luisa2014fua	FR006L2	Strasbourg	4639,826194	7,567826	48,575384
Moselle - rural	Rural area	FR	477	654866	Admin-1 scale rank	FRA-5328	Moselle	10699,297314	6,783301	49,012937
Rouen	Metropolitan area	FR	315	653617	luisa2014fua	FR215L2	Rouen	6775,823917	1,076269	49,476495
Grenoble	Metropolitan area	FR	182	643945	luisa2014fua	FR026L2	Grenoble	5330,622737	5,695304	45,132627
Montpellier	Metropolitan area	FR	126	611187	luisa2014fua	FR010L2	Montpellier	4260,217293	3,777454	43,681041
Rennes	Metropolitan area	FR	182	604085	luisa2014fua	FR013L2	Rennes	8526,531411	-1,76992	48,081909
Pas-de-Calais - rural	Rural area	FR	482	553414	Admin-1 scale rank	FRA-5334	Pas-de-Calais	9175,0982	2,291491	50,493665
Vendée - rural	Rural area	FR	242	550178	Admin-1 scale rank	FRA-5353	Vendée	13401,807896	-1,401723	46,68128
Toulon	Metropolitan area	FR	35	549207	luisa2014fua	FR032L2	Toulon	1966,06404	6,088412	43,189129
La Réunion - rural	Rural area	FR	20	517497	Admin-1 scale rank	FRA-4601	La Réunion	2846,423975	55,544512	-21,121658
Saint-Etienne	Metropolitan area	FR	113	501714	luisa2014fua	FR011L2	Saint-Etienne	3665,138066	4,315259	45,445938
Nancy	Medium-sized urban area	FR	320	493691	luisa2014fua	FR016L2	Nancy	6466,570411	6,210893	48,639671
Clermont-Ferrand	Medium-sized urban area	FR	186	441872	luisa2014fua	FR022L2	Clermont-Ferrand	5461,451703	3,136571	45,817216
Tours	Medium-sized urban area	FR	122	440546	luisa2014fua	FR035L2	Tours	6194,367782	0,619915	47,37771
Nord - rural	Medium-sized rural area	FR	212	428774	Admin-1 scale rank	FRA-5330	Nord	6202,197987	3,297482	50,397772

Urban Areas and Regional Rural Areas are classified following the same rules (large, medium, small…) based on population (Geonames, 2010-2018) with:

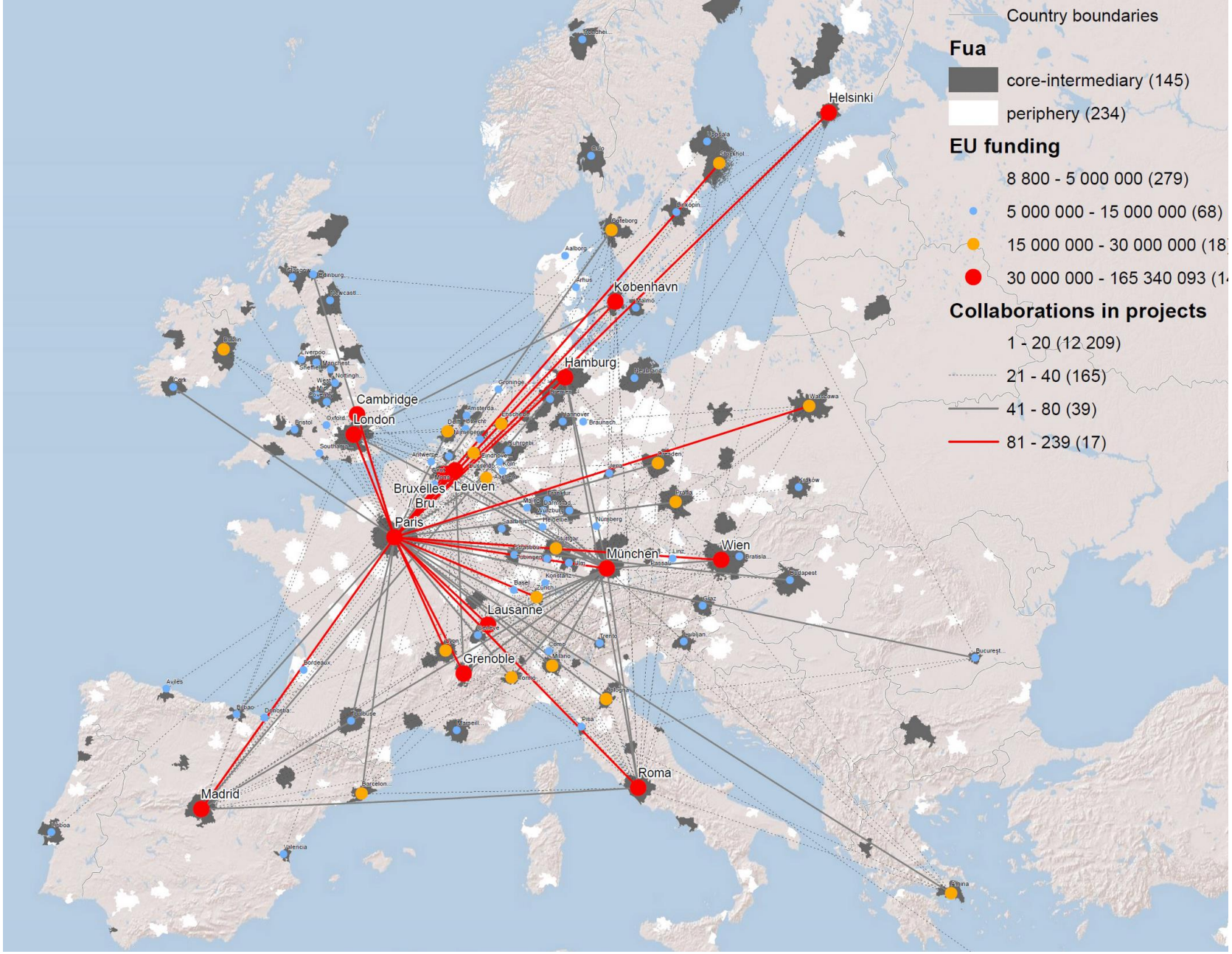
- Source ids, for interlinking;
- Core city names or source names when accessible;
- Centroid calculation…

2.2/ Cartographie des réseaux géographiques

Objectif à long terme

Offrir aux utilisateurs du Manager un cheminement d'analyse symétrique a celui appliqué aux contenus (analyse sémantique) avec mais appliqué à la géographie:

- **Filtering** : ce qui est important (noeux et liens) et ce qui fait liens (proximity measures);
- **Clustering géographique** : détection des zones d'agrégation spatiale (algo : DBScan avec Chameleon, autres ?);
- **Network Analysis**: à partir de ces agrégats, construire le réseau à partir des liens définis par des mesure de proximité ou encore une analyse centre-périphérie;
- **Project a 3rd variable onto the map**: associer aux clusters une troisième/seconde variable, comme celle des clusters sémantiques pour montrer en quoi ces espaces sont spécifiques (spécialisation) où quelles sont leurs caractéristiques (ranking, volumes).



Metropolisation, peripheries and funding of nano sciences & technologies (Villard et al., 2017). Avec l'utilisation du package flows (Beauguitte, Giraud, & Guerois, 2016).