Improving data quality of UN/LOCODE using open source tools: case study

Stefano Sabatini

UN/LOCODE Conference 2016

Geneva, April 28th 2016

Using UN/LOCODE

Standard for maritime industry communication

- Container Terminals/Depots (UN/EDIFACT)
- Customs (European Entry Summary Declaration - ENS)

```
LOC+11+ITGOA:139:6+VTE:TER:ZZZ'
```

LOCODE Search

Search by Country and Subdivision

-- Country --

-- Please Select --

Search

Search by location (free text)

Search

Search by LOCODE (free text)

Search

You searched locodes from country IT, subdivision GE.

Locode	NameWoDiacritics	SUName	Function
ITZBB	Borzonasca	Genova	-----6--
ITZEO	Campo Ligure	Genova	-----6--
ITCZU	Casarza Ligure	Genova	--3--6--
ITCAX	Casella	Genova	--3-----
ITZEY	Cicagna	Genova	-----6--
ITGOA	Genova	Genova	12345---

Data Quality in UN/LOCODE

Fields with possible issues in the code list:

- Name
- Subdivision
- Coordinates

How to detect these issues?

Data quality: geography (example)

Data quality: geography

With a spreadsheet (example: LibreOffice Calc)

- Split coordinates in their components DDMMx DDDMMx [=MID($K2;1;2)]
- Check: Direction, Out of bounds degrees (not in [0,180] or [0,90]) and minutes (not in [0,59])
- Mind your country bounding box (Italy: latitude [35,47], longitude [6,18])

I	J	K	L	M	N	O	P	Q	R
Dat	IATA	Coordinates	Remarks	latd	latr	latdir	lond	lonm	londir
901		4496N 10470E		44	96	N	104	70	E
901		4466N 00798E		44	66	N	007	98	E
907		4499N 09690E		44	99	N	096	90	E
901		4695N 01191E		46	95	N	011	91	E
1001		4497N 01044E		44	97	N	010	44	E

Data quality: geography - Results

- Direction: 1 error [was: E → now: W]
- Coordinates - Degrees: 4 errors on latitude, 55 on longitude

Leading zeroes problems

IT GHE 4544N 12230E → 4528N 01215E

- Coordinates – Minutes: 11 errors on latitude, 24 on longitude

Decimal instead of DMS format

IT TTO 4496N 10470E → 4457N 01028E

Data quality: names

- Order alphabetically by name
- If any similarities check subdivision and coords

Example:

- IT GLA was in Ferrara (FE),
- IT GDD is the same as IT GGR (with wrong coordinates and name)

B	C	D	E	F	G	H	I	J	K
Countr	Locatior	Name	NameWoDiacriti	Subdivisior	Statu	Functior	Dat	IATA	Coordinates
IT	GLA	Gallo	Gallo	CN	RL	--3-----	607		4443N 11330E
IT	GDD	Gallo d'Alba	Gallo d'Alba		RL	-----6--	901		4466N 00798E
IT	GGR	Gallo di Grinzane	Gallo di Grinzane	CN	RL	--3-----	701		4439N 00758E

Data quality: spellcheck

Other possible problems:

- typographical mistakes (diacritics)
- incomplete names
- metadata in the name

B	C	D	F	G	H	I	J	K
Count	Locatio	Name	Subdivisio	Stat	Functio	Da	IAT	Coordinates
IT	BMM	Basalghelle Di Mansue'	TV	RL	--3-----	1207		4549N 02352E
IT	BED	Bedizzole(Brescia)		RQ	0-------	9307		
IT	SRF	Cascina Madonna Rosignano Monf.To	AL	RL	--3-----	1207		4504N 00824E
IT	DDO	Dogana do Ortonovo		RL	--3-----	9805		
IT	BTS	Isorella, Brescia	BS	RL	--3-----	607		4518N 01019E

Data quality: metadata

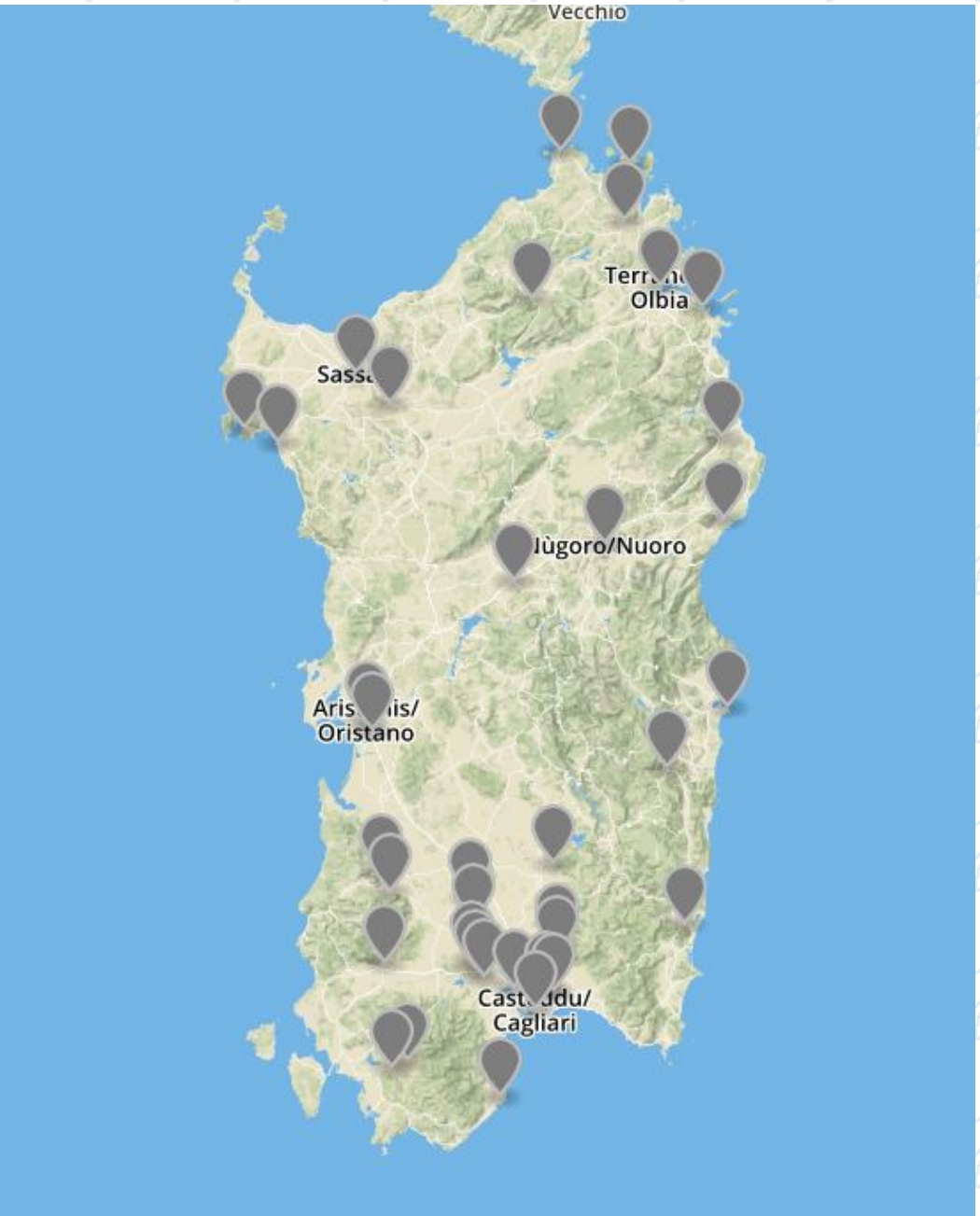
Other metadata issues:

- Missing subdivision
- Missing coordinates
- Missing both subdivision & coordinates

What to do with unidentifiable locations?

Candidates for removal

Case study: Sardinia

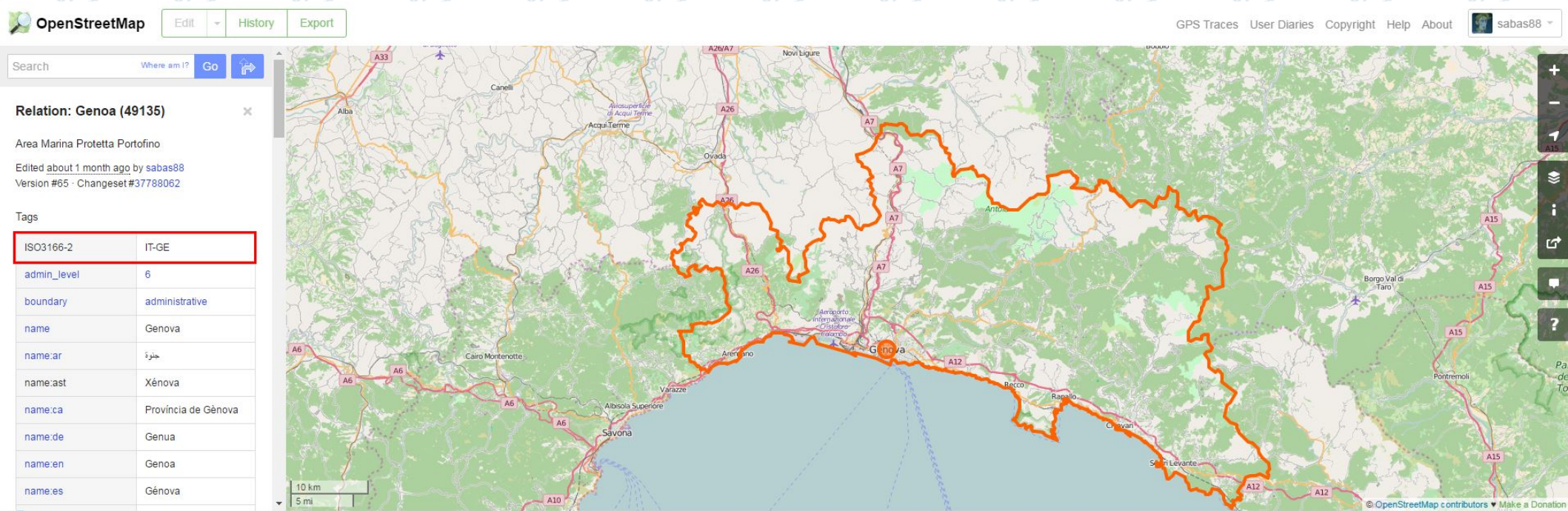


- Sample of 37 LOCODEs with coordinates
- Test against OpenStreetMap
- Following figures are in kilometers

Median	0.993445
Minimum	0.160789
Maximum	119.808256
1st quartile	0.729795
3rd quartile	1.641366
Interquartile range	0.911571
Outliers	119.808256, 46.487778, 11.30125

Tools: OpenStreetMap

- Geographic Database
- Open License
- Crowdsourcing (with a bit of open government data)
- “Ground truth”



OSM: Example usage

- Search for a place
- Get coordinates
- Other possible hints?

FIND COORDINATES

Just click on the map

COO
44.407144,8.934159

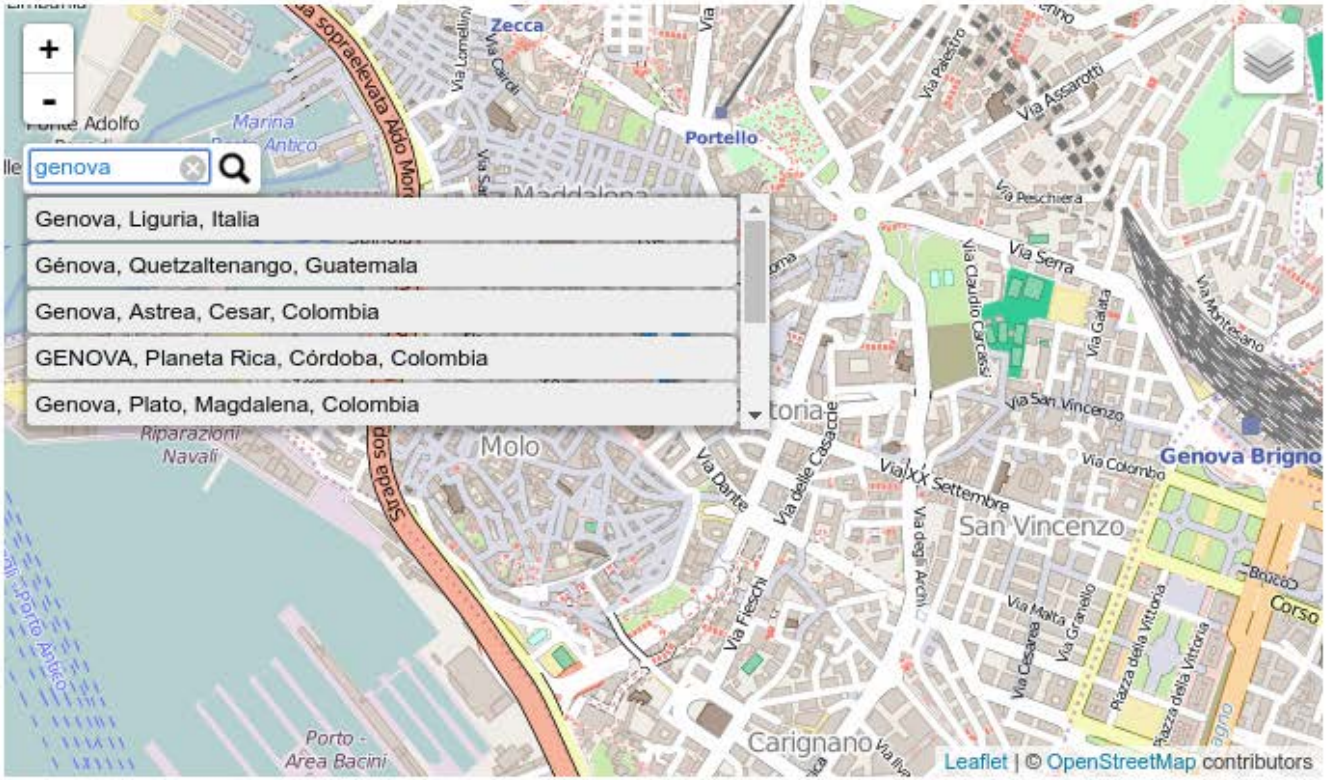
LAT
44.407144

LON
8.934159

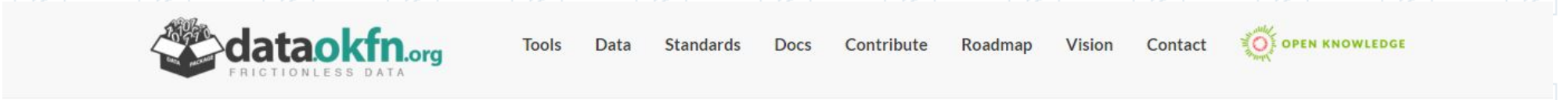
DMS
44° 24' 25" N 8° 56' 2" E

OLC
8FPCCW4M+VM

LOCODE
4424N 00856E



Reuse examples: OKFN core dataset



http://data.okfn.org/data/core/un-locode

UN-LOCODE Codelist

UN-LOCODE datapackage — Read more

Download Data

Metadata | Report an Issue

github.com/datasets/un-locode

OPEN DATA

Sources
UNECE

Data Files

Code list	Download	[Local: CSV – JSON]
Country codes	Download	[Local: CSV – JSON]
Function classifiers	Download	[Local: CSV – JSON]
Status indicators	Download	[Local: CSV – JSON]
Subdivision codes	Download	[Local: CSV – JSON]

Code list

Data Table | 102690 records | Search data ... | Go »

Change	Country	Location	Name	NameWoD...	Subdivisio...	Status	Function	Date	IATA	Coordinat...	Remarks
	AU	AVA	Alexandra	Alexandra	VIC	RL	-----6--	1201		3711S 14542E	
	AU	ALX	Alexandria	Alexandria	NSW	RL	--3-----	201		3354S 15113E	
	AU	AXL	Alexandria	Alexandria	NT	AI	---4----	9601			
	AU	ASP	Alice Springs	Alice Springs	NT	AI	---4----	9601			
	AU	ALB	Allambie Heigh...	Allambie Heigh...	NSW	RL	--3-----	1207		3345S 15115E	
	AU	AFD	Allansford	Allansford	VIC	RQ	--3-----	1101		3824S 14235E	
	AU	ABH	Alpha	Alpha	QLD	AI	---4----	9601			

Reuse examples: Whosonfirst

https://whosonfirst.mapzen.com/

unlc

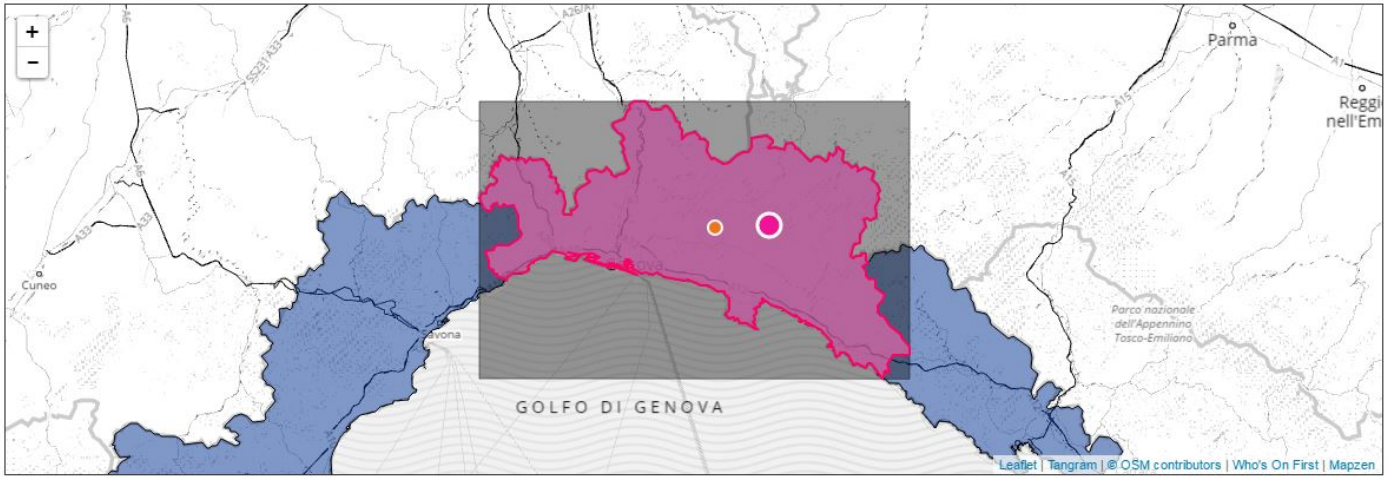
subdivision	IT-GE

who's on **first** jump to a **random place**

Search for a place Search

*remember — this is a simple full-text search tool and **not** a proper geocoder*

Genova 856/852/19/85685219.geojson



Genova is a **region** and its consensus geometry is derived from **quattroshapes**. Its **label centroid** is derived from **mapshaper**. *Take a **screenshot of this map** (this may require a few seconds to complete)*

Properties — *some notes about sources and names*

view raw

edtf

cessation	uuuu
inception	uuuu

geom

area	0.207549
bbox	8.57159513351,44.2166374522,9.57489305955,44.6764224636

Hierarchy

the **continent** of Europe
the **country** of Italy
the **macroregion** of Liguria
the **region** of Genova

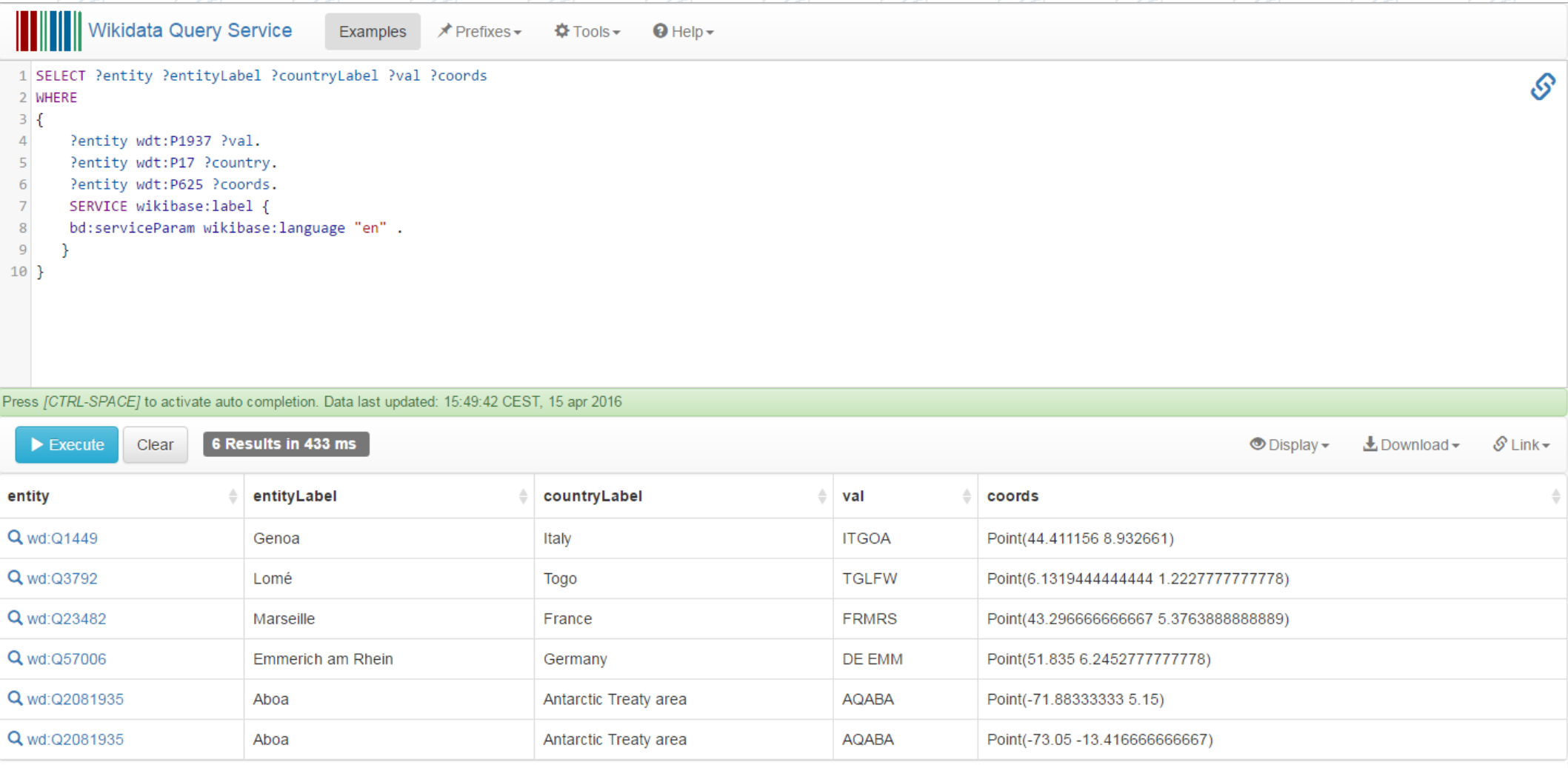
Other

See all the descendants of Genova
Raw data (GeoJSON)

Log

Reuse examples: Wikidata

https://query.wikidata.org



entity	entityLabel	countryLabel	val	coords
wd:Q1449	Genoa	Italy	ITGOA	Point(44.411156 8.932661)
wd:Q3792	Lomé	Togo	TGLFW	Point(6.1319444444444 1.2227777777778)
wd:Q23482	Marseille	France	FRMRS	Point(43.296666666667 5.3763888888889)
wd:Q57006	Emmerich am Rhein	Germany	DE EMM	Point(51.835 6.2452777777778)
wd:Q2081935	Aboa	Antarctic Treaty area	AQABA	Point(-71.88333333 5.15)
wd:Q2081935	Aboa	Antarctic Treaty area	AQABA	Point(-73.05 -13.416666666667)

Conclusions

DMRs generated in the first pass: 167 (2 days)

Room of improvement:

- Detect duplicates
- Detect wrong coordinates and Subdivisions
- Check Function field
- Add missing coordinates and Subdivisions

Possible solutions to improve and mantain UN/LOCODE:

- Cross-referencing with open data (Wikidata, OSM) and (open) government data
- Feedback from the general public and communities of interest (via DMR and outreach)

Thank you!

Stefano Sabatini
sabatst@coscon.it