Best Practices to Improve Data Quality of UN/LOCODE for the UN/LOCODE Focal Points

Drafted by Stefano Sabatini

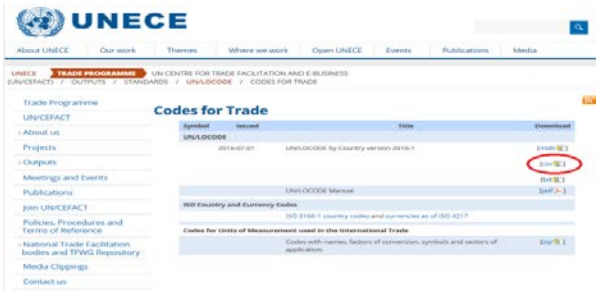
In this presentation:

1. **Error detection**

2. **Metadata improvements**

Error detection

Task 1: Preparing the file for working

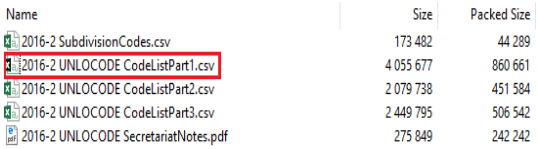


Approach 1

1. Download CSV package from the UNECE Website
   a. Go to http://www.unece.org/cefact/codesfortrade/codes_index.html
   b. Click [cvs]
   c. Open http://www.unece.org/fileadmin/DAM/cefact/locode/loc161csv.zip

2. Select the Code List part with the country you're dealing with in the unzipped package.

   (i.e. 2016-1 UN/LOCODE CodeListPart1.csv)



Name	Size	Packed Size
2016-2 SubdivisionCodes.csv	173 482	44 289
2016-2 UNLOCODE CodeListPart1.csv	4 055 677	860 661
2016-2 UNLOCODE CodeListPart2.csv	2 079 738	451 584
2016-2 UNLOCODE CodeListPart3.csv	2 449 795	506 542
2016-2 UNLOCODE SecretariatNotes.pdf	275 849	242 242

3. Open it in a spreadsheet software (like Microsoft Excel or Libreoffice Calc)
4. When importing it, you need to:
   a. Choose 'Windows-1252' as character encoding (when you notice garbled characters caused by diacritics)
   b. Add the header in the first row
      - Copy the following texts in the first cell

Change,Country,Location,Name,NameWoDiacritics,Subdivision,Function,Status,Date,IATA,Coordinates,Remarks

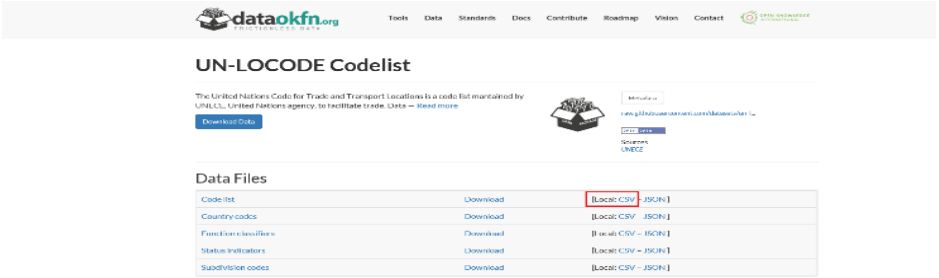
      - Split the rows in columns ("Text to columns" function) using comma (,) as separator

   c. When being asked to choose the Data format for each column, select the Data column and explicitly set it to 'Text' (this solves the disappearance of the leading zeroes in the column); do this also for Subdivision and Status (if not for all of them)

Error detection

Task 1: Preparing the file for working

Approach 2:



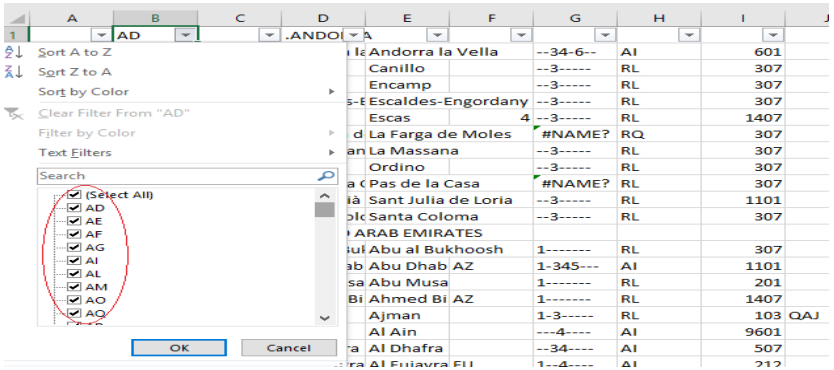
1. Use the package from http://data.okfn.org/data/core/un-locode :
2. Select Code list and click Download
3. When opening it, the character set is UTF-8, the separator is the comma

   Note: This package has removed

- the country rows (example: ,"AD",,".ANDORRA",,,,,,,,)
- the name equivalence rows

  (example: "=","AE","","Ruwais = Ar Ruways","Ruwais = Ar Ruways","",,"",,"","","","")

4. Filter the data (select the first row and apply the Filter function) to select a country in question



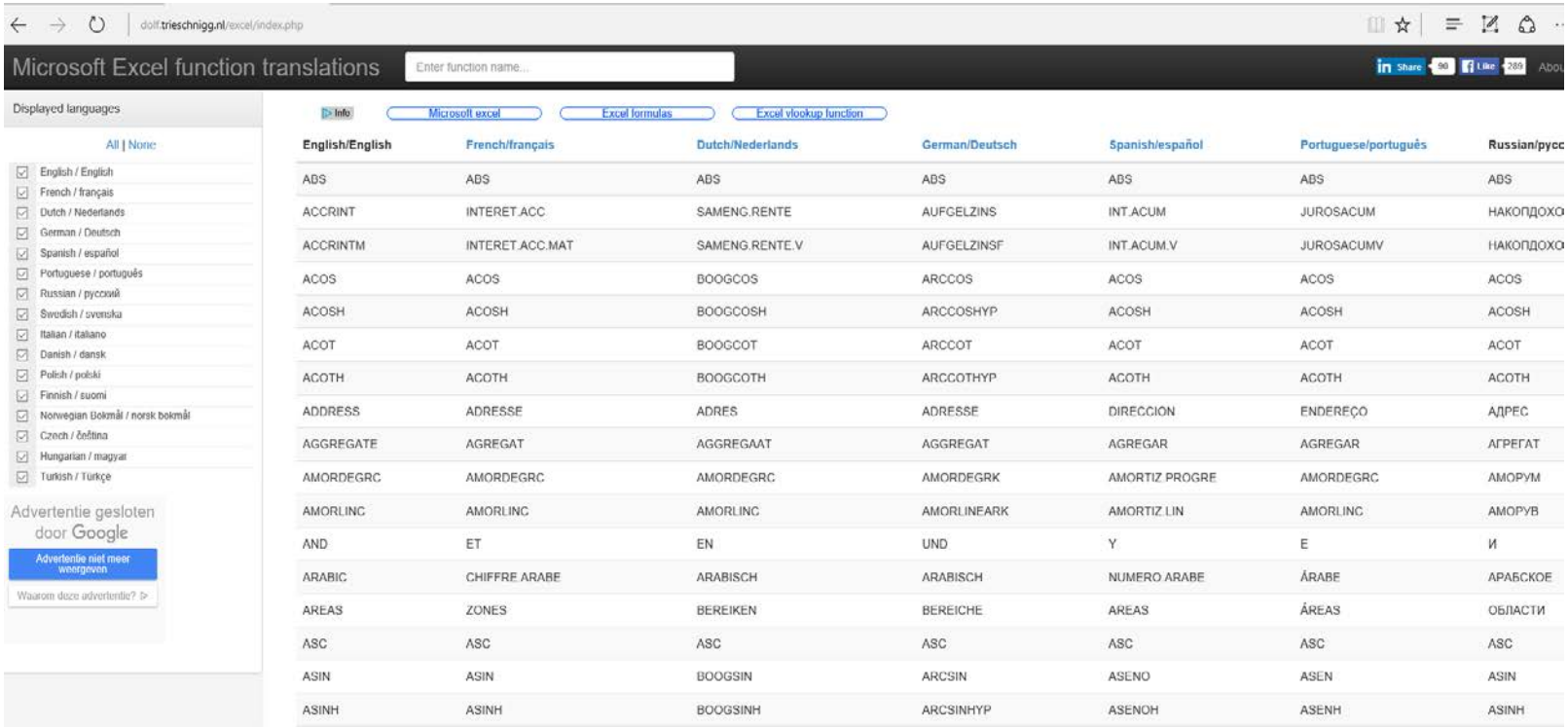
Error detection

Task 2: Checking Coordinates

Note:

- In this step, we are going to check for common errors, not for coordinates precision.
- For your reference, Excel can change the function names based on your language.

This website can help search the function translations: http://dolf.trieschnigg.nl/excel/index.php



English/English	French/français	Dutch/Nederlands	German/Deutsch	Spanish/español	Portuguese/português	Russian/русс
ABS	ABS	ABS	ABS	ABS	ABS	ABS
ACCRINT	INTERET.ACC	SAMENG.RENTE	AUFGELZINS	INT.ACUM	JUROSACUM	НАКОПДОХО
ACCRINTM	INTERET.ACC.MAT	SAMENG.RENTE.V	AUFGELZINSF	INT.ACUM.V	JUROSACUMV	НАКОПДОХО
ACOS	ACOS	BOOGCOS	ARCCOS	ACOS	ACOS	ACOS
ACOSH	ACOSH	BOOGCOSH	ARCCOSHYP	ACOSH	ACOSH	ACOSH
ACOT	ACOT	BOOGCOT	ARCCOT	ACOT	ACOT	ACOT
ACOTH	ACOTH	BOOGCOTH	ARCCOTHYP	ACOTH	ACOTH	ACOTH
ADDRESS	ADRESSE	ADRES	ADRESSE	DIRECCION	ENDEREÇO	АДРЕС
AGGREGATE	AGREGAT	AGGREGAAT	AGGREGAT	AGREGAR	AGREGAR	АГРЕГАТ
AMORDEGRC	AMORDEGRC	AMORDEGRC	AMORDEGRK	AMORTIZ.PROGRE	AMORDEGRC	АМОРУМ
AMORLINC	AMORLINC	AMORLINC	AMORLINEARK	AMORTIZ.LIN	AMORLINC	АМОРУВ
AND	ET	EN	UND	Y	E	И
ARABIC	CHIFFRE.ARABE	ARABISCH	ARABISCH	NUMERO.ARABE	ÁRABE	АРАБСКОЕ
AREAS	ZONES	BEREIKEN	BEREICHE	AREAS	ÁREAS	ОБЛАСТИ
ASC	ASC	ASC	ASC	ASC	ASC	ASC
ASIN	ASIN	BOOGSIN	ARCSIN	ASENO	ASEN	ASIN
ASINH	ASINH	BOOGSINH	ARCSINHYP	ASENOH	ASENH	ASINH

Error detection

Task 2: Checking Coordinates

Step 1. Conversion from sexagesimal format into decimal one

Suppose you have the first row at position 2, the Coordinates column at position K.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		AD		.ANDORRA															
2		AD	ALV	Andorra la	Andorra la Vella		--34-6--	AI	601		4230N 00131E								
3		AD	CAN	Canillo	Canillo		--3-----	RL	307		4234N 00135E								

1. Position yourself in cell M2 from here we are going to split the coordinates in their components:

		example: 4230N 00131E
a.	Latitude_degrees (M2): =MID($K2,1,2)	42
b.	Latitude_minutes (N2): =MID($K2,3,2)	30
c.	Latitude_direction (O2): =MID($K2,5,1)	N
d.	Longitude_degrees (P2): =MID($K2,7,3)	001
e.	Longitude_minutes (Q2): =MID($K2,10,2)	31
f.	Longitude_direction (R2): =MID($K2,12,1)	E

2. In case you'd want to have a column with the decimal conversion of the coordinates (GPS coordinates, used in web mapping applications) you can add

- a column for latitude [S2] with this formula =$M2+($N2/60)*(IF($O2="S“,-1,1)) and
- another for longitude [T2] with this formula =$P2+($Q2/60)*(IF($R2="W",-1,1))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		AD		.ANDORRA																
2		AD	ALV	Andorra la	Andorra la Vella		--34-6--	AI	601		4230N 00131E		42	30	N	001	31	E	42.5	1.516667
3		AD	CAN	Canillo	Canillo		--3-----	RL	307		4234N 00135E									

Error detection

Task 2: Checking Coordinates

Step 2: Checking

1. Minutes errors detection:

For this step you should check the two minutes column (N2 and Q2): If greater than 59, they need to be fixed (because in sexagesimal notation numbers go from 0 to 60, and 60 equals 0)

2. Bounding box test:

- Create a rectangle enclosing your area of interest : bounding box.

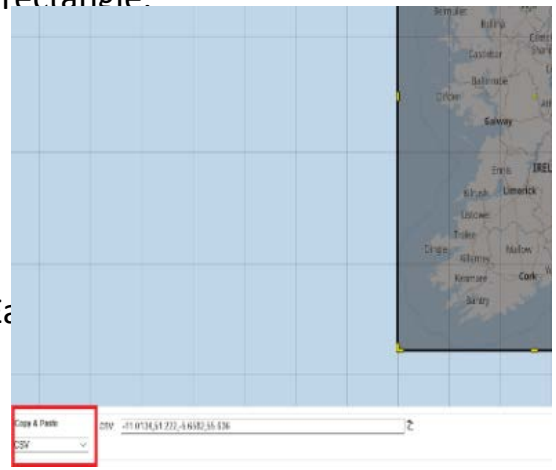
You can use this tool http://boundingbox.klokantech.com/ and select "CSV" format in the left-bottom corner, which gives you a string containing "western latitude, southern longitude, eastern latitude, northern longitude". i.e. the bounding box for Switzerland is with the parameters: 5.96,45.82,10.49,47.81.

- Search via the Filtering function of the spreadsheet for values outside this rectangle.

Example: Ireland has a rough bounding box -10.3,51.7,-5.7,55.5;

on the spreadsheet if we filter for latitude less than 51 or more than 55 we find one match, i.e. IE GOW, which should have 52 instead of 57 as latitude degree.

Note: It could be possible to find also direction problems (West instead of Ea



Error detection

Task 3: Location name checking

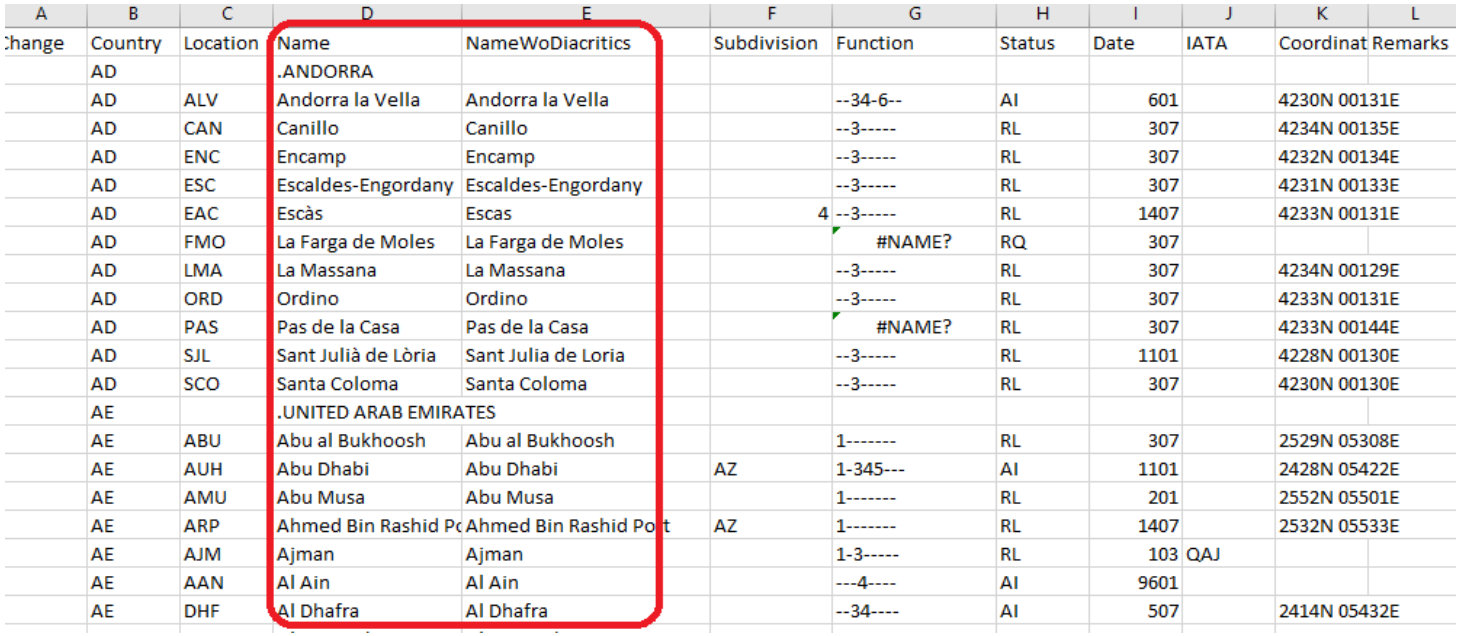
Order the spreadsheet by Name or NameWoDiacritics.

For each name, check

a. for spelling correctness (abbreviations, apostrophes, diacritics)

b. Check based on section 3.3 of the UN/LOCODE manual (http://www.unece.org/fileadmin/DAM/cefact/locode/unlocode_manual.pdf)

c. check if there could be possible duplications comparing similar names with the assistance of Subdivision and Coordinates metadata



A	B	C	D	E	F	G	H	I	J	K	L
hange	Country	Location	Name	NameWoDiacritics	Subdivision	Function	Status	Date	IATA	Coordinat	Remarks
	AD		.ANDORRA								
	AD	ALV	Andorra la Vella	Andorra la Vella		--34-6--	AI	601		4230N 00131E	
	AD	CAN	Canillo	Canillo		--3-----	RL	307		4234N 00135E	
	AD	ENC	Encamp	Encamp		--3-----	RL	307		4232N 00134E	
	AD	ESC	Escaldes-Engordany	Escaldes-Engordany		--3-----	RL	307		4231N 00133E	
	AD	EAC	Escàs	Escas	4	--3-----	RL	1407		4233N 00131E	
	AD	FMO	La Farga de Moles	La Farga de Moles		#NAME?	RQ	307			
	AD	LMA	La Massana	La Massana		--3-----	RL	307		4234N 00129E	
	AD	ORD	Ordino	Ordino		--3-----	RL	307		4233N 00131E	
	AD	PAS	Pas de la Casa	Pas de la Casa		#NAME?	RL	307		4233N 00144E	
	AD	SJL	Sant Julià de Lòria	Sant Julia de Loria		--3-----	RL	1101		4228N 00130E	
	AD	SCO	Santa Coloma	Santa Coloma		--3-----	RL	307		4230N 00130E	
	AE		.UNITED ARAB EMIRATES								
	AE	ABU	Abu al Bukhoosh	Abu al Bukhoosh		1-------	RL	307		2529N 05308E	
	AE	AUH	Abu Dhabi	Abu Dhabi	AZ	1-345---	AI	1101		2428N 05422E	
	AE	AMU	Abu Musa	Abu Musa		1-------	RL	201		2552N 05501E	
	AE	ARP	Ahmed Bin Rashid Pc	Ahmed Bin Rashid Po t	AZ	1-------	RL	1407		2532N 05533E	
	AE	AJM	Ajman	Ajman		1-3-----	RL	103	QAJ		
	AE	AAN	Al Ain	Al Ain		---4----	AI	9601			
	AE	DHF	Al Dhafra	Al Dhafra		--34----	AI	507		2414N 05432E	

Metadata improvements

Status (Refresh older entries)

The objective of the Metadata improvements is to ensure identification of the locations, each one of them should have the correct metadata.

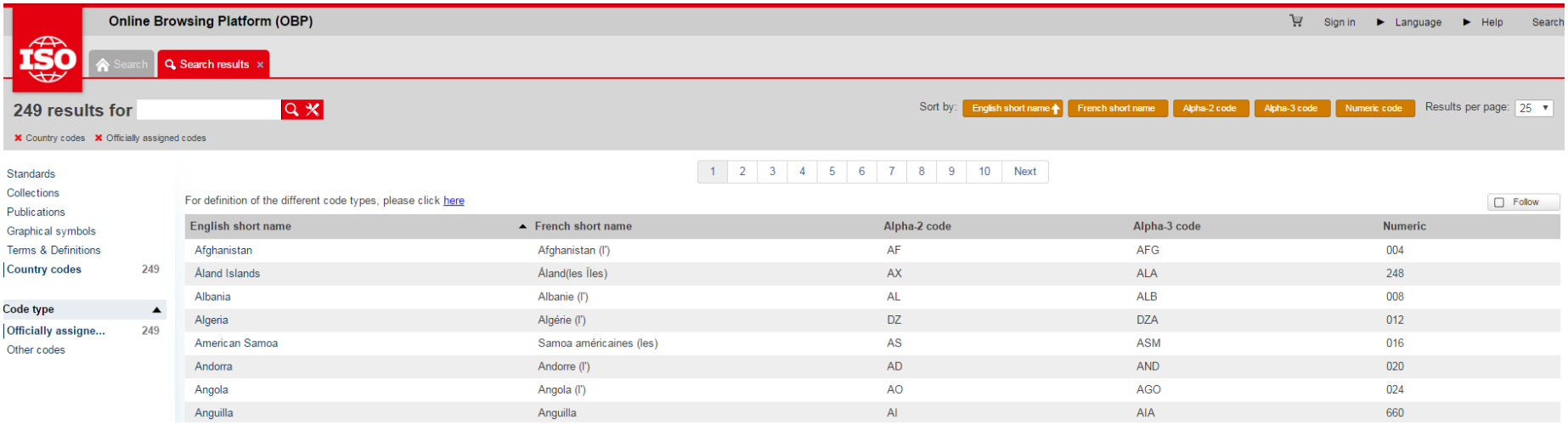
Priority in (re)checking entries should be given to:

- entries where Date column is empty or where Status is not yet Approved (start focusing on QQ and UR)
- entries where Date column is older than 20 years

Metadata improvements

Subdivision

Subdivision is based on ISO-3166-2 standard (https://www.iso.org/obp/ui/#search/code).



Metadata improvements

Function

Function	Description	Remark
0	Specifies that the functional use of a location is not known and is to be specified.	
1	Specifies that the location is a Port, as defined in UNECE Recommendation 16.	
2	Specifies that the location is a Rail terminal.	
3	Specifies that the location is a Road terminal.	
4	Specifies that the location is an Airport.	Authorized by IATA
5	Specifies that the location is a Postal exchange office.	Authorized by UPU
6	Value reserved for multimodal functions, ICDs etc.	
7	Value reserved for fixed transport functions (e.g. oil platform).	
B	Specifies that the location is Border crossing.	

- Function '0' indicates "function not known". Entries with function '0' should be updated with the correct function(s).
- To assist in the classification, each function should map to a general criterion, which could be established at country level or could be discussed to provide uniformity.

Example questions to which you would answer:

- What is a postal exchange office in your country? A "mail sorting center"?
- Is there any official registry for this function to use as reference?
- Should a location have the function "airport" only when a location has an IATA-classified airport?

Metadata improvements

Coordinates

When feasible, you should fill in the coordinates in the correct format: DDMMx DDDMMx

If you're given a coordinate, you should ask if it is in the correct projection (data should be in Web Mercator projection, EPSG:900913), especially when asking to people working with geographic systems, because in different parts of the world are using different coordinates (more precise) projection.

To convert from a decimal coordinate, you could use this simple algorithm (example: 41.894100, 12.502441):

- The whole part becomes the degree (41)
- Remove the whole part and multiply per 60
  (41.894100 – 41 = 0.8941; 0.8941*60 = 53.646)
- The whole part becomes the minute (53)
- The desired coordinate is 4153, since the starting number is positive, we would use direction N, so the final number is 4153N
- Repeating the steps for the second coordinate, gives 1230E.
- The conversion result is '4153N 01230E'

Data sources

When checking locations or processing DMRs, you could ask for confirmation to the responsible agency in your country.

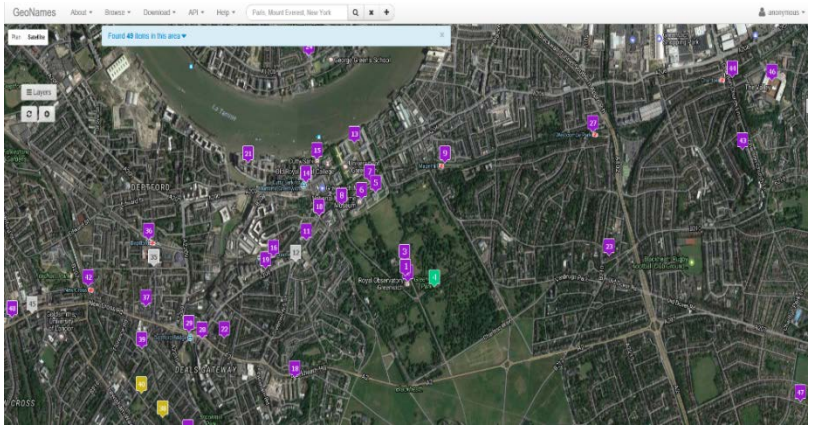
- **AA** Approved by competent national government agency
- **AC** Approved by Customs Authority
- **AF** Approved by national facilitation body
- **AI** Code adopted by international organisation (IATA or ECLAC)
- **AS** Approved by national standardisation body
- **RL** Recognised location
- **RN** Request from credible national sources for locations in their own country
- **RQ** Request under consideration
- **RR** Request rejected
- **QQ** Original entry not verified since date indicated
- **XX** Entry that will be removed from the next issue of UN/LOCODE

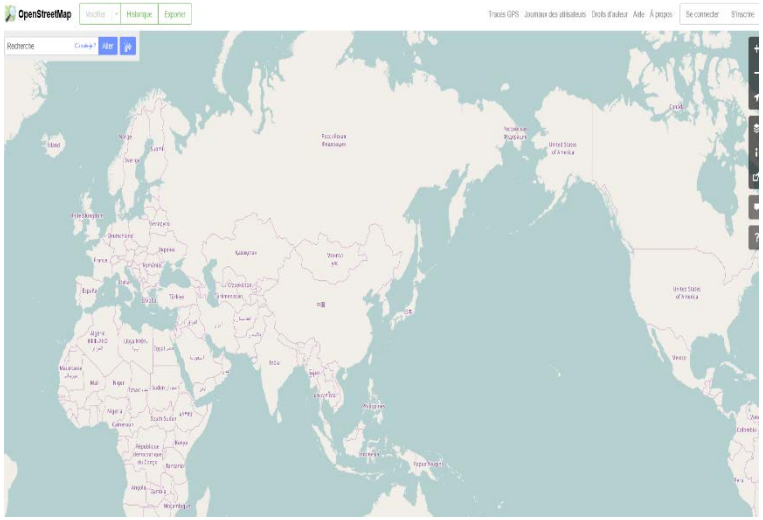
Entries with function(s) authorized by governmental bodies should be with status starting with ‘A’.

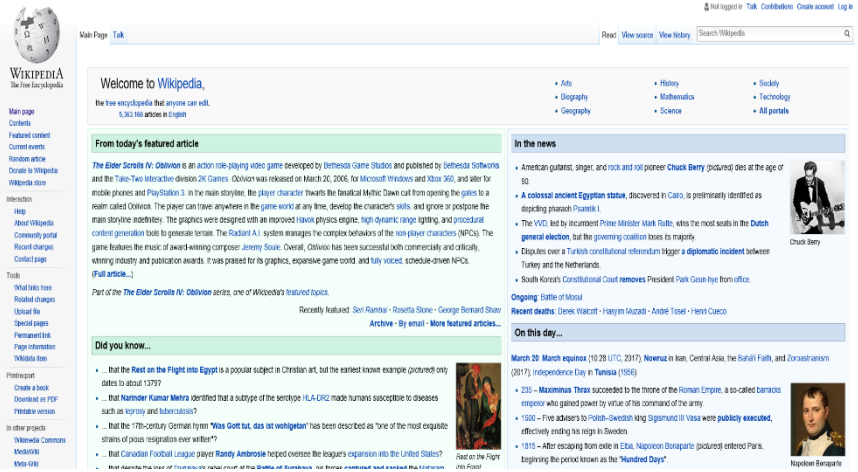
Data sources

Free data sources:

- http://www.geonames.org/v3/
- http://www.openstreetmap.org/
- https://en.wikipedia.org/wiki/Main_Page







Data sources

OpenStreetMap and Whosonfirst

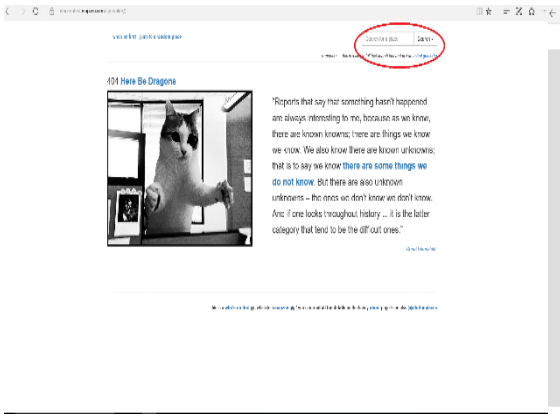
- OpenStreetMap (http://www.openstreetmap.org/)

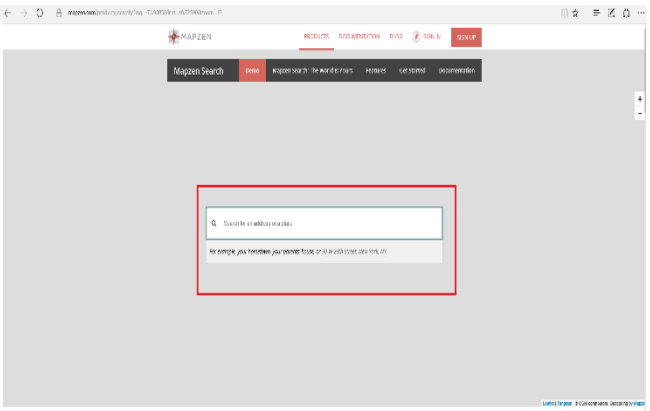
OpenStreetMap is a global geographic database built by the people on the Wikipedia model (crowdsourcing): everyone can log in and edit the data by themselves, and the data is available to download and process. Many services are built using it as source of data for map displays.

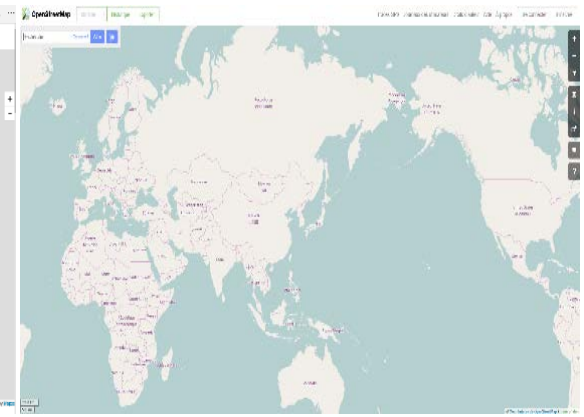
The main website provides services to the users, but can be used also as a browsing interface to look for places and check their descriptions (OSM uses a key=value data model for attributes).

- Whosonfirst (https://whosonfirst.mapzen.com/spelunker/)

Whosonfirst is a global gazetteer , essentially a big list of places with their geometry. It can be used to look up places and get information about their administrative structure (getting their subdivision for example). A geocoding service using this project is Mapzen search (https://mapzen.com/projects/search/).







Data sources

Wikipedia / Wikidata

- Wikipedia

Wikipedia, the free encyclopedia, is a reliable source of information and if is often one of the first results when searching on Google.

On Wikipedia there are different experiments of using UN/LOCODE:

- a proposal of displaying into the Infobox (https://en.wikipedia.org/wiki/Template_talk:Infobox_settlement/Archive_7#UN.2FLOCODE)

- a redirect template (https://en.wikipedia.org/wiki/Template:R_from_UN/LOCODE) that redirects a UN/LOCODE to the city (or location) page.

Wikipedia could be used to get information about infrastructure, location (articles could have coordinates in the top right corner of the page), and administrative subdivisions.

- Wikidata

The sister project Wikidata is a database of information editable like Wikipedia: each object has an identifier (Q plus a number) and has different 'predicates' (property –P plus a number- and value). The property "UN/LOCODE" exists as P1937.

Example: Genoa, Italy (https://www.wikidata.org/wiki/Q1449) has UN/LOCODE (https://www.wikidata.org/wiki/Property:P1937) ITGOA.

Each predicate can have 'qualifiers': one could add a start date (P580) or an end date (P582) to the predicate when it is added or deprecated.

The advantage of Wikidata is that in the future it could be used as a data source for Wikipedia, and also can be interrogated as a database to gain information with the Wikidata query service (https://query.wikidata.org/) using a language called SPARQL.

Data sources

Wikipedia / Wikidata

As an example, this query lists all the objects in Wikidata having a UN/LOCODE property, returning a list with the name and the coordinates.

```
SELECT ?entity ?entityLabel ?countryLabel ?val ?coords
WHERE
{
    ?entity wdt:P1937 ?val.
    ?entity wdt:P17 ?country.
    ?entity wdt:P625 ?coords.
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en".
    }
}
```

Applying this to a single country is simple, we change the country predicate to return all the locations in San Marino (https://www.wikidata.org/wiki/Q238).

```
SELECT ?entity ?entityLabel ?val ?coords
WHERE
{
      ?entity wdt:P1937 ?val.
    ?entity wdt:P17 wd:Q238.
    ?entity wdt:P625 ?coords.
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en".
    }
}
```