Open Data
with a focus on Belgium

Brussels, 31/03/2017

Jan Wijffels

BNOSAC - jwijffels@bnosac.be

bnosac

Overzicht

bnosac

open analytical helpers

Government data - overview

Global open data index

- Belgium ranking: `http://index.okfn.org/place/belgium`
- Belgium ranked 35 (2017/03/27)
- In general: a lot of data lacking, scatterred & non-integrated

Belgium

43% open

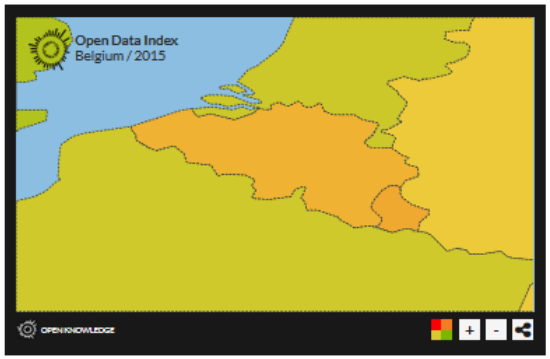
Belgium is ranked #35 in the 2015 Index

Belgium's overall Index ranking is up from #53 in 2014

See other years 2014 2013

Share or embed this page

</> Embed | Twitter | Facebook | Google+



Rank	Dataset	Breakdown	Location (URL)	Format	Info	Prev. (2014)	Score
8	Government Spending		n/a	n/a		#15 10%	10%
20	Election Results		http://www.elections.fgov.be/i...	XLSX		#51 45%	70%
20	Water Quality		http://www.milieurapport.be/Up...	n/a		n/a	35%
21	National Map		http://www.ngi.be /topomapviewe...	SHP		#59 30%	60%
25	Government Budget		http://www.begroting.be /FR/Doc...	xls		#59 45%	70%
26	Procurement tenders		https://enot.publicprocurement...	XML		n/a	60%
28	Weather forecast		http://www.meteo.be/meteo /view...	n/a		n/a	45%
30	Pollutant Emissions		http://www.irceline.be/	n/a		#36 45%	45%
31	Land Ownership		n/a	vector ...		n/a	30%
33	National Statistics		http://stat.nbb.be/Index.aspx?...	Excel, ...		#23 70%	70%
39	Legislation		http://www.belgielex.be	n/a		#33 45%	45%
58	Company Register		n/a	n/a		#19 55%	15%
61	Location datasets		n/a	n/a		#57 0%	0%

b nosac

open analytical helpers

Belgium: Statistics Belgium

StatBel - Overview

Open data from Statistics Belgium `http://statbel.fgov.be/nl/statistieken/opendata`

- Statistical sectors & spatial maps
- Net Income data (from FOD Finances)
- Census 2011: `http://census2011.fgov.be/index_nl.html`
- 'Licentie open data' (= Creative Commons Attribution 2.0) `https://creativecommons.org/licenses/by/2.0`

Figure: Statistics Belgium open data

StatBel - maps

- administrative boundaries **at the level of the statistical sector**
- R package by BNOSAC: `https://github.com/bnosac/BelgiumMaps.StatBel`

```
library(BelgiumMaps.StatBel)
data(BE_ADMIN_SECTORS)
bxl <- subset(BE_ADMIN_SECTORS, TX_RGN_DESCR_NL %in% "Brussels Hoofdstedelijk Gewest")
plot(bxl, main = "NIS sectors in Brussels")
```

StatBel - maps

- spatially aggregated to level of municipality, district, province, region
- enriched with NIS/NUTS codes

```
data(BE_ADMIN_MUNTY)
data(BE_ADMIN_DISTRICT)
data(BE_ADMIN_PROVINCE)
plot(BE_ADMIN_MUNTY, main = "Belgium municipalities/districts/provinces")
plot(BE_ADMIN_DISTRICT, lwd = 2, add = TRUE)
plot(BE_ADMIN_PROVINCE, lwd = 3, add = TRUE)
```


```
str(BE_ADMIN_MUNTY@data)
```

```
'data.frame':   589 obs. of  20 variables:
 $ CD_MUNTY_REFNIS     : chr  "11001" "11002" "11004" "11005" ...
 $ TX_MUNTY_DESCR_NL   : chr  "Aartselaar" "Antwerpen" "Boechout" "Boom" ...
 $ TX_MUNTY_DESCR_FR   : chr  "Aartselaar" "Anvers" "Boechout" "Boom" ...
 $ CD_DSTR_REFNIS      : int  11000 11000 11000 11000 11000 11000 11000 11000 11000 11000 ...
 $ TX_ADM_DSTR_DESCR_NL: chr  "Arrondissement Antwerpen" "Arrondissement Antwerpen" "Arrondissement Antwerpen" "A
 $ TX_ADM_DSTR_DESCR_FR: chr  "Arrondissement d^^92Anvers""| __truncated__ "Arrondissement d^^92Anvers""| __trunc
 $ CD_PROV_REFNIS      : int  10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 ...
 $ TX_PROV_DESCR_NL    : chr  "Provincie Antwerpen" "Provincie Antwerpen" "Provincie Antwerpen" "Provincie Antwe
 $ TX_PROV_DESCR_FR    : chr  "Province d^^92Anvers""| __truncated__ "Province d^^92Anvers""| __truncated__ "Prov
 $ CD_RGN_REFNIS       : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
 $ TX_RGN_DESCR_NL     : chr  "Vlaams Gewest" "Vlaams Gewest" "Vlaams Gewest" "Vlaams Gewest" ...
 $ TX_RGN_DESCR_FR     : chr  "Région flamande" "Région flamande" "Région flamande" "Région flamande" ...
 $ nuts0               : chr  "BE" "BE" "BE" "BE" ...
 $ nuts1               : chr  "BE2" "BE2" "BE2" "BE2" ...
 $ nuts2               : chr  "BE21" "BE21" "BE21" "BE21" ...
 $ nuts3               : chr  "BE211" "BE211" "BE211" "BE211" ...
```

StatBel - income

- Tax income in CSV files, packaged in R at

  `https://github.com/weRbelgium/BelgiumStatistics`

- At **the level of the commune** from 2005 up to 2014: nr residents, nr fiscal declarations, total net taxable income, roerend/onroerend/professioneel inkomen, tax deductions, municipality, regional and state taxes
- At **the level of the statistical sector** Total net taxable income from years 2005 up to 2013

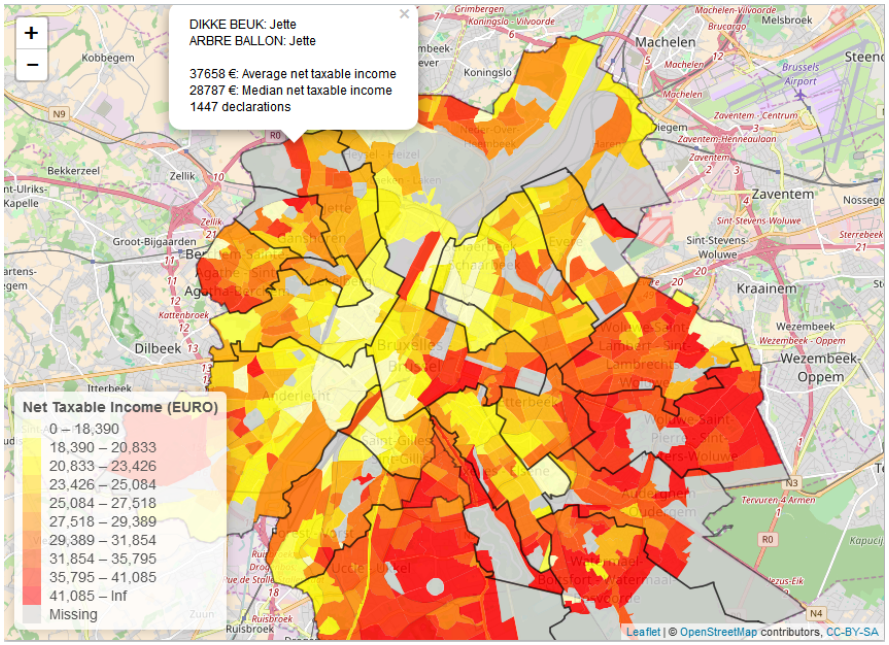


Figure: Income statistics

StatBel - income

```
load("data/TF_PSNL_INC_TAX_SECTOR.RData")
str(TF_PSNL_INC_TAX_SECTOR)

'data.frame':   178787 obs. of  13 variables:
 $ CD_YEAR                   : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
 $ CD_MUNTY_REFNIS           : int  11001 11001 11001 11001 11001 11001 11001 11001 11001 11001 ...
 $ CD_SECTOR                 : chr  "A00-" "A01-" "A020" "A030" ...
 $ CD_REFNIS_SECTOR          : chr  "11001A00-" "11001A01-" "11001A020" "11001A030" ...
 $ MS_NBR_NON_ZERO_INC       : int  1203 1509 449 245 204 1457 424 16 269 33 ...
 $ MS_TOT_NET_TAXABLE_INC    : int  34861256 52985912 16851901 7074668 12979300 46130749 13135454 NA 10658222 NA
 $ MS_AVG_TOT_NET_TAXABLE_INC: int  28979 35113 37532 28876 63624 31661 30980 NA 39622 NA ...
 $ MS_MEDIAN_NET_TAXABLE_INC : int  21962 25261 25591 22651 28844 23261 23281 NA 22332 21510 ...
 $ MS_INT_QUART_DIFF         : int  21856 28117 32363 24246 40705 26453 26275 NA 36305 24229 ...
 $ MS_INT_QUART_COEFF        : int  100 111 126 107 141 114 113 NA 163 113 ...
 $ MS_INT_QUART_ASSYM        : int  22 31 35 27 39 28 38 NA 40 53 ...
 $ TX_SECTOR_DESCR_NL        : chr  "AARTSELAAR-CENTRUM" "DE LEEUWERIK" "BUERSTEDE" "YSSELAAR" ...
 $ TX_SECTOR_DESCR_FR        : chr  "AARTSELAAR-CENTRUM" "DE LEEUWERIK" "BUERSTEDE" "YSSELAAR" ...


library(BelgiumStatistics)
data(TF_PSNL_INC_TAX_MUNTY)
str(TF_PSNL_INC_TAX_MUNTY)

'data.frame':   5301 obs. of  41 variables:
 $ CD_YEAR                   : int  2005 2006 2007 2008 2009 2010 2011 2012 2013 2005 ...
 $ CD_MUNTY_REFNIS           : int  11001 11001 11001 11001 11001 11001 11001 11001 11001 11002 ...
 $ MS_NBR_NON_ZERO_INC       : int  8065 8155 8153 8285 8236 8280 8145 8117 8075 265752 ...
 $ MS_NBR_ZERO_INC           : int  381 261 266 428 530 517 427 375 348 21131 ...
 $ MS_TOT_NET_TAXABLE_INC    : num  2.61e+08 2.67e+08 2.80e+08 2.86e+08 2.94e+08 ...
 $ MS_TOT_NET_INC            : num  2.63e+08 2.71e+08 2.83e+08 2.90e+08 2.99e+08 ...
 $ MS_NBR_TOT_NET_INC        : int  8065 8155 8153 8285 8236 8280 8145 8117 8075 265752 ...
 $ MS_REAL_ESTATE_NET_INC    : num  2857128 2945478 2954362 2996274 2990553 ...
 $ MS_NBR_REAL_ESTATE_NET_INC: int  815 830 822 814 808 857 879 912 1059 18674 ...
 $ MS_TOT_NET_MOV_ASS_INC    : num  614864 300369 323008 438377 1625597 ...
 $ MS_NBR_NET_MOV_ASS_INC    : int  94 100 104 129 143 155 205 370 338 2036 ...
 $ MS_TOT_NET_VARIOUS_INC    : num  454013 616257 637938 440221 477309 ...
 $ MS_NBR_NET_VARIOUS_INC    : int  160 156 140 136 146 118 108 113 120 4620 ...
 $ MS_TOT_NET_PROF_INC       : num  2.59e+08 2.67e+08 2.80e+08 2.86e+08 2.94e+08 ...
 $ MS_NBR_NET_PROF_INC       : int  8016 8112 8115 8253 8190 8248 8120 8080 8044 264303 ...
 $ MS_SEP_TAXABLE_INC        : num  14564091 13441786 21357545 18657486 21023480 ...
 $ MS_NBR_SEP_TAXABLE_INC    : int  1680 1891 1766 1833 1773 1757 1789 1847 1803 57077 ...
```

StatBel - land occupation

- Land occupation from Cadaster. From 1983-2015
- At level of municipality
- containing surface, parcels, cadastral income, taxable surface
- http://statbel.fgov.be/nl/statistieken/cijfers/leefmilieu/geo/bodembezetting_kadaster

```
library(BelgiumStatistics)
data(TF_EAE_LAND_OCCUPTN_2015)
str(TF_EAE_LAND_OCCUPTN_2015)

'data.frame':   15903 obs. of  31 variables:
 $ CD_YEAR               : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ CD_RUB_LVL_3          : chr  "1AE" "1BC" "1DI" "1F" ...
 $ CD_RUB_LVL_2          : chr  "1TOT" "1TOT" "1TOT" "1TOT" ...
 $ CD_RUB_LVL_1          : chr  "4TOT" "4TOT" "4TOT" "4TOT" ...
 $ CD_RUB_LVL_0          : chr  "6TOT" "6TOT" "6TOT" "6TOT" ...
 $ TX_RUB_FR_LVL_3       : chr  "Terres agricoles" "Pâtures, prés" "Jardins et parcs" "Vergers" ...
 $ TX_RUB_FR_LVL_2       : chr  "Parcelles non-bâties" "Parcelles non-bâties" "Parcelles non-bâties" "Parcelles n
 $ TX_RUB_FR_LVL_1       : chr  "Superficie cadastrée" "Superficie cadastrée" "Superficie cadastrée" "Superficie
 $ TX_RUB_NL_LVL_3       : chr  "Akkerland" "Grasland" "Tuinen en parken" "Boomgaarden" ...
 $ TX_RUB_NL_LVL_2       : chr  "Onbebouwde percelen" "Onbebouwde percelen" "Onbebouwde percelen" "Onbebouwde per
 $ TX_RUB_NL_LVL_1       : chr  "Gekadastreerde oppervlakte" "Gekadastreerde oppervlakte" "Gekadastreerde opperv
 $ CD_MUNTY_REFNIS       : int  11001 11001 11001 11001 11001 11001 11001 11001 11001 11001 ...
 $ TX_MUNTY_DESCR_NL     : chr  "Aartselaar" "Aartselaar" "Aartselaar" "Aartselaar" ...
 $ TX_MUNTY_DESCR_FR     : chr  "Aartselaar" "Aartselaar" "Aartselaar" "Aartselaar" ...
 $ CD_ADM_DSTR_REFNIS    : int  11000 11000 11000 11000 11000 11000 11000 11000 11000 11000 ...
 $ TX_ADM_DSTR_DESCR_NL  : chr  "Arrondissement Antwerpen" "Arrondissement Antwerpen" "Arrondissement Antwerpen"
 $ TX_ADM_DSTR_DESCR_FR  : chr  "Arrondissement d^^92Anvers""| __truncated__ "Arrondissement d^^92Anvers""| __tru
 $ CD_PROV_REFNIS        : int  10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 ...
 $ TX_PROV_DESCR_NL      : chr  "Provincie Antwerpen" "Provincie Antwerpen" "Provincie Antwerpen" "Provincie Antw
 $ TX_PROV_DESCR_FR      : chr  "Province d^^92Anvers""| __truncated__ "Province d^^92Anvers""| __truncated__ "Pr
 $ CD_RGN_REFNIS         : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
 $ TX_RGN_DESCR_NL       : chr  "Vlaams Gewest" "Vlaams Gewest" "Vlaams Gewest" "Vlaams Gewest" ...
 $ TX_RGN_DESCR_FR       : chr  "Région flamande" "Région flamande" "Région flamande" "Région flamande" ...
 $ MS_NUM_PRCL           : int  446 265 108 3 36 47 11 11 115 553 ...
 $ MS_TXBL_SUR           : num  259.492 161.318 16.103 0.175 14.532 ...
 $ MS_NON_TXBL_SUR       : num  12.831 7.258 3.646 0.461 10.102 ...
 $ MS_TOT_SUR            : num  272.323 168.576 19.748 0.636 24.634 ...
```

StatBel - population structure

- Population information at level of the commune from 2009-2016
- Gender/Nationality/Age/Code civil/Married
- Population number more detailed at 1km2 grid

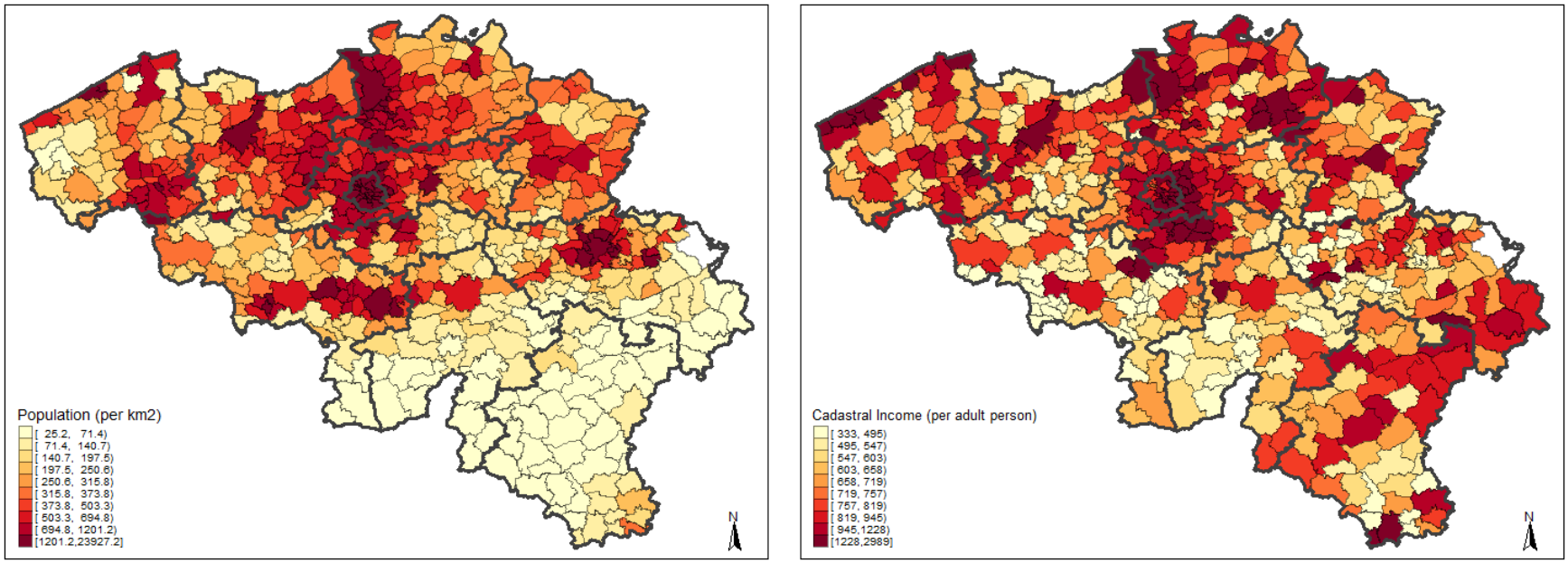


Figure: Population stats

StatBel - census

- Census data from Statbel
- `https://github.com/weRbelgium/BelgiumStatistics`

AGE	Age en années révolues au 01/01/2011	Leeftijd in verstreken jaren in 01/01/2011
BAT	Salle de bain	Badkamer
CAS	Situation sur le marché de l'emploi	Arbeidsmarktsituatie
COC	Pays de citoyenneté	Land van staatsburgerschap
DRM	Nombre de pièces par occupant	Aantal kamers per bewoner
EDU	Niveau d'instruction	Opleidingsniveau
FST	Position dans la famille	Positie in de familiekern
GEO	Lieu de résidence	Verblijfplaats
HAR	Conditions de logement	Huisvestingsregeling
HST	Position dans le ménage	Positie in het huishouden
IND	Secteur d'activité économique	Economische sector
LMS	Etat civil	Burgerlijke staat
LOC	Taille de la localité	Grootte van de agglomeratie
LPW	Lieu de travail	Plaats van tewerkstelling
NOC	Nombre d'occupants	Aantal bewoners
NOR	Nombre de pièces	Aantal kamers
OCS	Régime d'occupation des logements c	Bewoningssituatie
OWS	Régime de propriété	Type eigendom
POB	Lieu de naissance	Geboorteplaats
POC	Logements par époque de constructior	Bouwperiode
ROY	Lieu de résidence habituelle au 01/01	Gewone verblijfplaats op 01/01/2010
SEX	Sexe	Geslacht
SFN	Taille du noyau familial	Grootte familiekern
SIE	Statut professionnel	Beroepsstatuut
SPH	Taille du ménage privé	Grootte particulier huishouden
TFN	Type de noyau familial	Type familiekern
TLQ	Type de local d'habitation	Type woonverblijf
TOB	Type de bâtiment	Type gebouw
TOH	Chauffage central	Centrale verwarming
TPH	Type de ménage privé	Type particulier huishouden
TSH	Modalités de jouissance du logement	Eigendomssituatie
YAE	Année d'immigration depuis 1980	Jaar van aankomst sinds 1980
YAT	Année d'immigration depuis 2000	Jaar van aankomst sinds 2000

StatBel - census

- Several statistics per province
- Some statistics more detailed than others

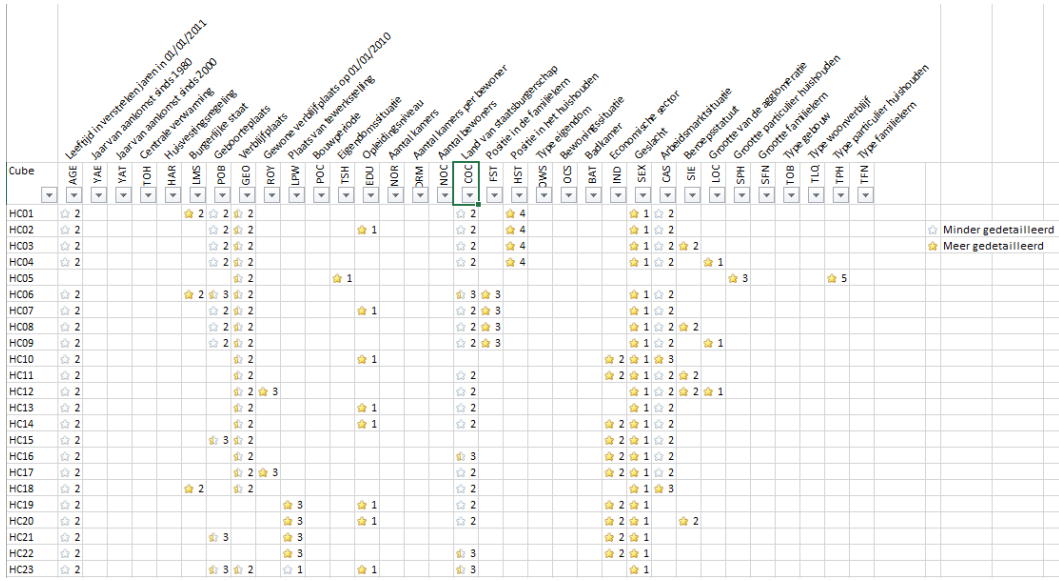


Figure: Census data combinations

```
library(BelgiumStatistics)
data(TF_CENSUS_2011_HC01_L)
str(TF_CENSUS_2011_HC01_L)

'data.frame':   130593 obs. of  60 variables:
 $ CD_GEO_LVL_0          : chr  "BE" "BE" "BE" "BE" ...
 $ CD_GEO_LVL_1          : chr  "BE2" "BE2" "BE1" "BE3" ...
 $ TX_GEO_DESCR_FR_LVL_1: chr  "Région flamande" "Région flamande" "Région de Bruxelles-Capitale" "Région wallonn
 $ TX_GEO_DESCR_NL_LVL_1: chr  "Vlaams Gewest  " "Vlaams Gewest  " "Brussels Hoofdstedelijk Gewest  " "Waals Gewe
 $ CD_GEO_LVL_2          : chr  "BE22" "BE21" "BE10" "BE33" ...
 $ TX_GEO_DESCR_FR_LVL_2: chr  "Province du Limbourg" "Province d'Anvers" "Région de Bruxelles-Capitale" "Provinc
 $ TX_GEO_DESCR_NL_LVL_2: chr  "Provincie Limburg" "Provincie Antwerpen" "Brussels Hoofdstedelijk Gewest" "Prov
 $ CD_SEX_LVL_0          : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

StatBel - commuting

Commuting sqlite / Microsoft Access database with commuter info, namely: Number of commutes in 2011 from statistical sector to statistical sector

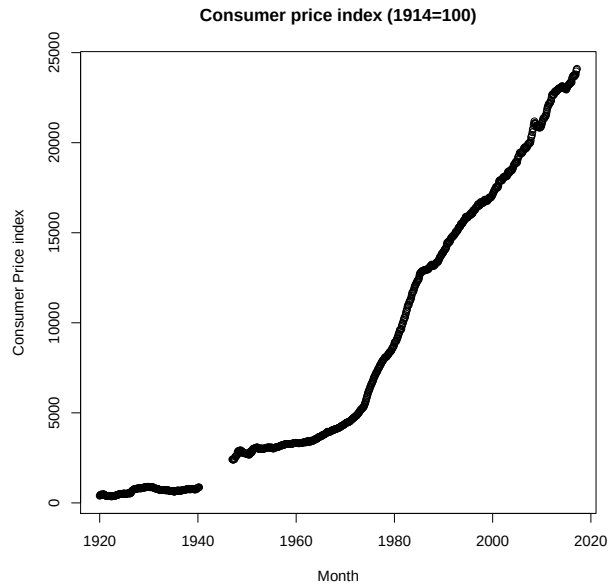
CD_SECTOR_RESIDENCE	CD_SECTOR_WORK	OBS_VALUE
BE100_21001A00-	BE100_21001A00-	121
BE100_21001A011	BE100_21001A00-	10
BE100_21001A02-	BE100_21001A00-	5
BE100_21001A031	BE100_21001A00-	7
BE100_21001A041	BE100_21001A00-	3
BE100_21001A051	BE100_21001A00-	3
BE100_21001A10-	BE100_21001A00-	7
BE100_21001A112	BE100_21001A00-	10

Figure: Commuting statistics

StatBel - prices

Prices inflation, consumer price index, health index
overall + by product group
Available at http://statbel.fgov.be/en/statistics/opendata/datasets/prices

```
## Example, showing the consumer price index
library(data.table)
f <- tempfile()
download.file(url = "http://statbel.fgov.be/nl/binaries/CPI All base years_tcm325-275955.zip",
              destfile = f)
x <- fread(input = sprintf("unzip -p %s", f),
           sep = "|", encoding = "UTF-8", na.strings = c("NA", "", "."))

## Reference level 1914
x <- subset(x, NM_BASE_YR == 1914)
x$timepoint <- as.Date(sprintf("%s-%s-01", x$NM_YR, x$NM_MTH))
plot(x$timepoint, x$MS_CPI_IDX, type = "p",
     xlab = "Month", ylab = "Consumer Price index",
     main = "Consumer price index (1914=100)")
```



Lacking

Lacking non-open

- Government backed (spatial) data from postal codes
- `http://www.bpost.be/site/nl/verzenden/adressering/zoek-een-postcode/` + `http://www.geopostcodes.com/Belgium`
- Government backed locations of cities

b nosac
open analytical helpers

Addresses + geopunt

Addresses - raw data

- Use cases: geocoding / mapping / database cleaning

Openaddresses: `https://openaddresses.io/`

- Worldwide dataset at `http://data.openaddresses.io`
- For Belgium contains all addresses in Flanders + Brussels
- Format provider-dependent (csv / shape file, data fiels different, sometimes lat/lon)
- Only require attribution
- For Flanders: based on Central Reference Address Database (CRAB = Centraal Referentieadressenbestand)

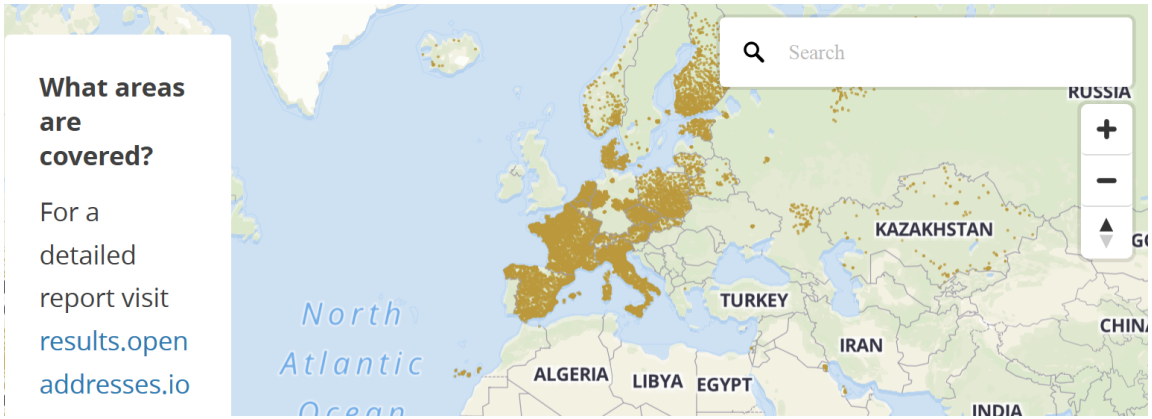


Figure: openaddresses.io

Addresses - raw data

Example of Brussels data

```
library(openadds)
x <- oa_get("http://data.openaddresses.io/runs/170853/be/bru/brussels-nl.zip")
head(x)
```

```
[[1]]
# A tibble: 219,860 × 11
        LON      LAT NUMBER                  STREET  UNIT
      <dbl>    <dbl>  <chr>                   <chr> <chr>
1  4.349066 50.87599     18             Ketelsstraat  <NA>
2  4.349304 50.87567      2             Ketelsstraat  <NA>
3  4.348664 50.87623    207     Emile Bockstaellaan  <NA>
4  4.327375 50.85250     21  Henri De Saegherstraat  <NA>
5  4.340312 50.82704     44               Wipstraat  <NA>
6  4.372233 50.89238    115             Pagodenlaan  <NA>
7  4.339682 50.82520     74 Alfred Cluysenaarstraat  <NA>
8  4.409547 50.83173     26    Edouard Lacomblélaan  <NA>
9  4.402897 50.83799     81            Tervurenlaan  <NA>
10 4.335133 50.82824     71   Fernand Bernierstraat  <NA>
# ... with 219,850 more rows, and 6 more variables:
#   CITY <chr>, DISTRICT <chr>, REGION <chr>,
#   POSTCODE <int>, ID <chr>, HASH <chr>
```

Addresses - raw data

Geonames Database with more than 11 Mio locations - worldwide dataset

Belgium: http://www.geonames.org/statistics/belgium.html

Web services e.g. wikipedia search in spatial neighbourhood, nearbystreets, postal codes

R package geonames

```
library(geonames)
options(geonamesUsername="bnosac")
GNfindNearbyWikipedia(lat=50.86619, lng=4.366354, lang = 'nl')
```

```
1                                                                                          St
2                                               De Aarschotstraat (Frans: Rue d'Aerschot) is een straat i
3                                                         Campus Sint-Lukas Brussel is een campus van de
4 Het Colignonplein (Frans: Place Colignon) is een plein in de gemeente Schaarbeek in het noordoosten van het Bru
5                                                    De Vooruitgangstraat (Frans: rue du Progrès) is een st
  elevation              lng distance rank lang
1        18          4.36529    0.321   47   nl
2        21 4.36366111111111   0.4161   40   nl
3        37 4.36677222222222   0.5229   53   nl
4        34         4.373615   0.5315   78   nl
5        23         4.361452   0.5359   91   nl
                     title              lat
1       Station Paleisstraat       50.8689981
2             Aarschotstraat 50.8628555555556
3   Campus Sint-Lukas Brussel 50.8614944444444
4              Colignonplein        50.867549
5 Vooruitgangstraat (Brussel)        50.862495
                                             wikipediaUrl
1            nl.wikipedia.org/wiki/Station_Paleisstraat
2                  nl.wikipedia.org/wiki/Aarschotstraat
3       nl.wikipedia.org/wiki/Campus_Sint-Lukas_Brussel
4                   nl.wikipedia.org/wiki/Colignonplein
5 nl.wikipedia.org/wiki/Vooruitgangstraat_%28Brussel%29
  feature countryCode
1    <NA>        <NA>
2 landmark       <NA>
```

Addresses - raw data

OpenStreetMap `http://www.openstreetmap.org`

- Worldwide coverage
- Easily imported with PostGIS or use exports in Spatial format by www.geofabrik.de (`http://download.geofabrik.de/europe.html`)
- Already structured: `http://download.geofabrik.de/osm-data-in-gis-formats-free.pdf`
- R package tmap or R package BelgiumMaps.OpenStreetMap
- Open Database License 1.0

Figure: geofabrik extract

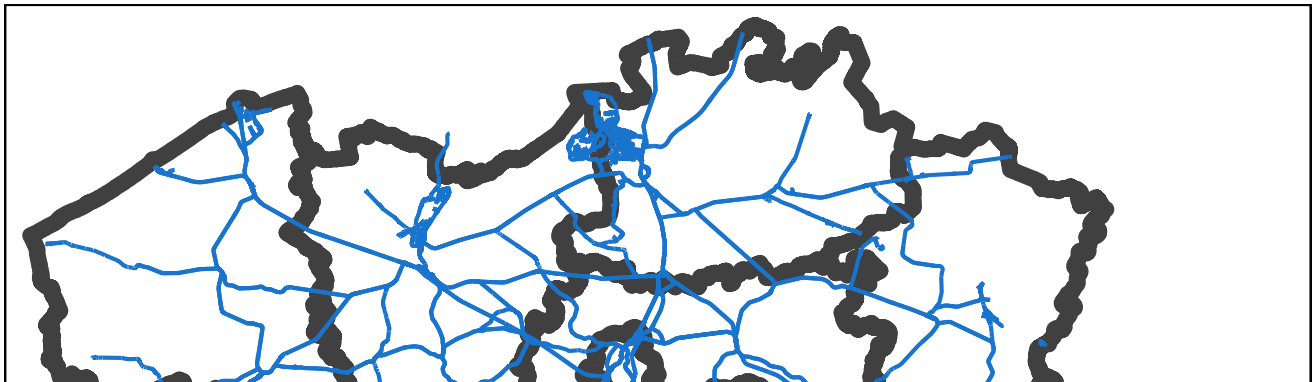
Addresses - raw data

- Example of Points of Interests
- BelgiumMaps.OpenStreetMap R package (Geofabrik extract)

```
library(BelgiumMaps.OpenStreetMap)
library(sp)
data(BE_OSM_POINTS)
tail(sort(table(BE_OSM_POINTS$type)), 10)
```

```
     restaurant     street_lamp traffic_signals
           4128            4604            5076
 level_crossing   waste_basket  turning_circle
           5298            6133            6354
      give_way           bench        crossing
           6527            7788           37684
      bus_stop
          69598
```

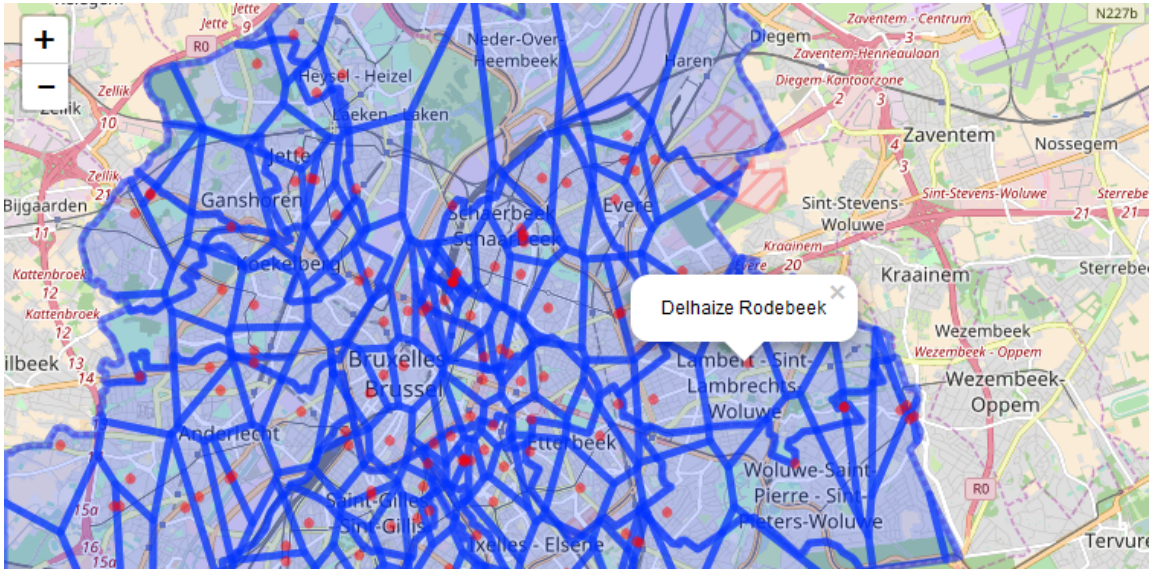
```
library(BelgiumMaps.Admin)
library(tmap)
data(BE_OSM_ADMIN)
data(BE_OSM_RAILWAYS)
provinces <- subset(BE_OSM_ADMIN, admin.level %in% 6)
rails <- subset(BE_OSM_RAILWAYS, type %in% "rail")
tm_shape(provinces) + tm_borders("grey25", lwd = 4) + tm_shape(rails) + tm_lines("dodgerblue3")
```



Addresses - raw data

- Voronoi tesselation of supermarkets in Brussels (what is the closest supermarket)

```
library(BelgiumMaps.StatBel)
library(rgeos)
library(dismo)
library(magrittr)
library(leaflet)
library(maptools)
data(BE_ADMIN_MUNTY)
bxl <- subset(BE_ADMIN_MUNTY, TX_RGN_DESCR_NL %in% "Brussels Hoofdstedelijk Gewest")

supermarkets <- subset(BE_OSM_POINTS, type %in% c("supermarket"))
supermarkets <- spCbind(supermarkets, over(x = supermarkets, y = bxl))
supermarkets_bxl <- subset(supermarkets, TX_RGN_DESCR_NL %in% "Brussels Hoofdstedelijk Gewest")
supermarketarea_bxl <- voronoi(supermarkets_bxl)
supermarketarea_bxl <- gIntersection(supermarketarea_bxl, bxl, byid=TRUE)
m <- leaflet(data = bxl) %>%
  addTiles() %>%
  addPolygons(data = supermarketarea_bxl) %>%
  addCircles(data = supermarkets_bxl, popup = supermarkets_bxl$name, col = "red")
```



Addresses - geocoding

Gisgraphy Combination of Openstreetmap, geonames + quattroshapes
Very good accuracy
Less support

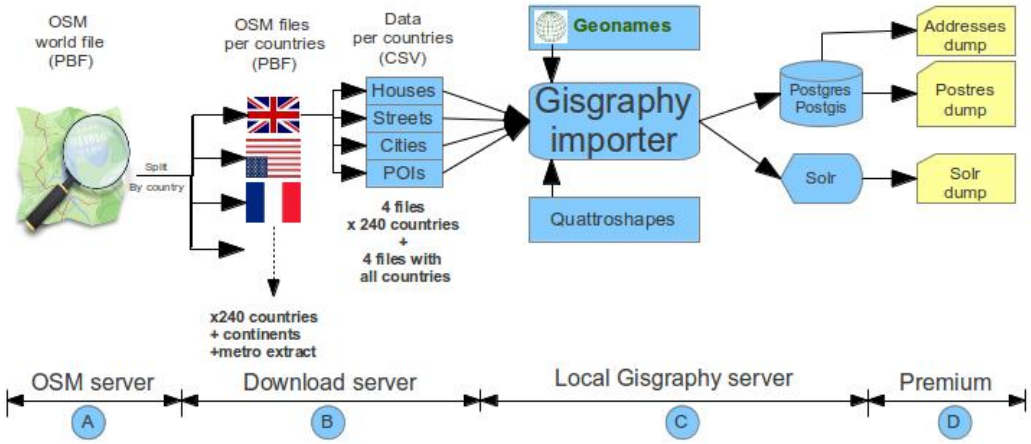


Figure: Gisgraphy

Nominatim Based on Openstreetmap
Easily setup as a webservice
Large community

`http://nominatim.openstreetmap.org/`

Addresses - kadaster

Kadaster Belgium: in CadGIS

- private parcel / parcel public domain or lost domain / parcel error
- request download at `https://www.geo.be/#/catalog/details/7797?l=en&content=full`

Figure: Cadastral parcels

Direct downloads non-trivial

- Open for Brussels: `http://opendatastore.brussels/dataset/58d00d48-daec-41dd-b9bf-14eaf8f552cb` or `http://data.gov.be/nl/node/16839` in GeoJSON /CSV format
- Flanders 2013 data: `http://www.geopunt.be` + `http://www.geopunt.be/catalogus/datasetfolder/7/70769db8-5416-479b-858b-de5f833dc767/74941e03-747f-4922-bb9d-b517b5d7f697`
- Wallonie: `http://geoservices.wallonie.be/cadmap` + `http://geoportail.wallonie.be/catalogue/c05542aa-be9b-4989-93a7-2c496c048d9a.html`

Flanders specific

- AGIV products: `https://download.agiv.be` & www.agiv.be/international/en/products
- Digital Heights data used for flood modelling, 3D city models, noise studies or archaeological research
- Bodemkaart morfogenetisch voor binnen-Vlaanderen (substraat, textuurklasse, drainageklasse, profielontwikkelingsgroep, fase en variante) en geomorfologisch voor de Kuststreek (substraat, serie, subserie, type, subtype en variante). Shapefile.

```
bodemkaart <- read_shape("data/20141212_Bodemkaart_download/20141212shape/20141212_bodemkaart_download_shp.shp",
                         stringsAsFactors = FALSE)
View(bodemkaart@data)
```

gid	id_kaartvl	Bodemtype	Bodemser_c	Bodemserie	Substr_leg	Grove_leg	Substr_V_c	SubstraatV	Textuu_c	Textuur	Drainage_c
26513	26513	o.A4	o.A4	Kreekruggronden (Oudlandpolders)	NA	Kreekrug	NA	NA	NA	NA	NA
34854	34854	Zdh	Zdh	Matig natte zandbodem met verbrokkelde ijzer en/of...	NA	Vochtig zand	NA	NA	Z	zand	d
35560	35560	Zagd	Zag	Zeer droge zandbodem met duidelijke ijzer en/of hu...	NA	Droog zand	NA	NA	Z	zand	a
35016	35016	Zbh	Zbh	Droge zandbodem met verbrokkelde ijzer en/of hum...	NA	Droog zand	NA	NA	Z	zand	b
35381	35381	Zdh	Zdh	Matig natte zandbodem met verbrokkelde ijzer en/of...	NA	Vochtig zand	NA	NA	Z	zand	d
40494	40494	Sep	Sep	Natte lemig zandbodem zonder profiel	NA	Nat zand	NA	NA	S	lemig zand	e

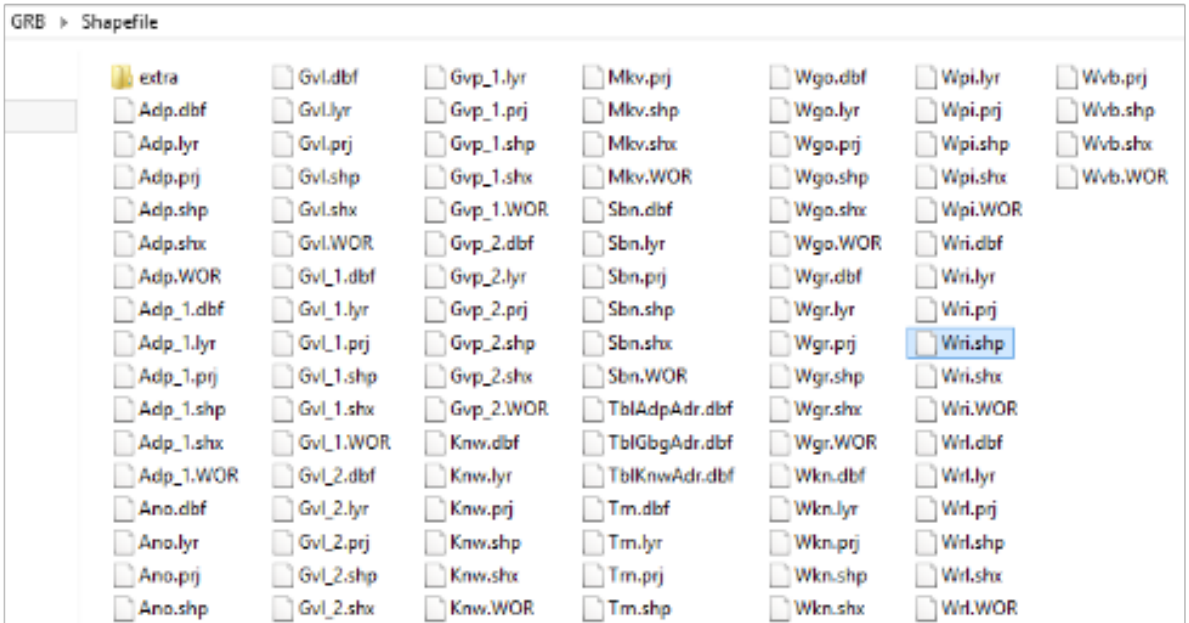
Figure: Bodemkaart

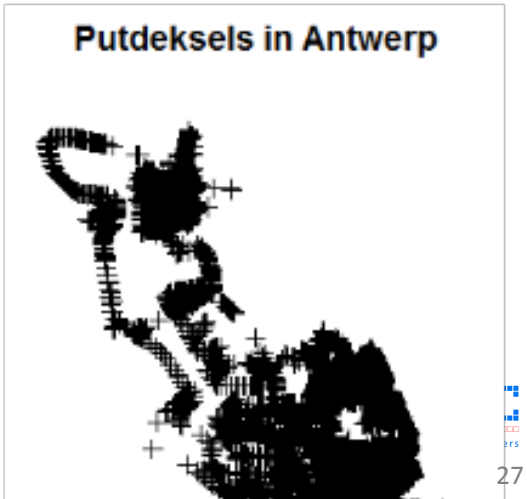
- INSPIRE directive: `http://www.geopunt.be/nl/voor-experts/inspire` + `http://geoportail.wallonie.be/INSPIRE` + `http://geobru.irisnet.be/en/about-inspire-and-geobru`

Flanders specific

- Grootschalig Referentiebestand (GRB), from 28-03-2017 met gebouwen, kavels, wegen en hun inrichting en waterlopen
- Containing: spatial location + inrichtingselementen (address, type, surface, length, niscode) wegbaan, spoorbaan, terrein, wegopdeling, wegbaaninrichting, oppervlaktewater, kadastrale indeling, ondersteunende entiteit, wegennetwerk

```
## Example: all putdeksels in Antwerp
library(BelgiumMaps.StatBel)
library(sp)
data(BE_ADMIN_MUNTY)
antwerp <- subset(BE_ADMIN_MUNTY, TX_MUNTY_DESCR_NL == "Antwerpen")
putdeksels <- read_shape("C:/Users/Jan/Desktop/GRB/Shapefile/Wri.shp", stringsAsFactors=FALSE)
put_locations <- spTransform(putdeksels, antwerp@proj4string) %over% antwerp
plot(putdeksels[which(put_locations$TX_MUNTY_DESCR_NL %in% "Antwerpen"), ], main = "Putdeksels in Antwerp")
```





b nosac

open analytical helpers

Belgium Business Register

www.bnosac.be

KBO

- http://economie.fgov.be/nl/ondernemingen/bce/pub/opendata
- Active companies, company locations and legal info in KBO (Banque-Carrefour des Entreprises / Kruispuntdatabank)
- Use cases: data cleaning, contacting (e.g. get me all dentists), benchmarking, sales. example: https://www.finactum.be
- https://kbopub.economie.fgov.be/kbo-open-data/login

Maand	Full	Update
november 2016	KboOpenData_0034_2016_11_Full.zip	KboOpenData_0034_2016_11_Update.zip
december 2016	KboOpenData_0035_2016_12_Full.zip	KboOpenData_0035_2016_12_Update.zip
januari 2017	KboOpenData_0036_2017_01_Full.zip	KboOpenData_0036_2017_01_Update.zip
februari 2017	KboOpenData_0037_2017_02_Full.zip	KboOpenData_0037_2017_02_Update.zip
maart 2017	KboOpenData_0038_2017_03_Full.zip	KboOpenData_0038_2017_03_Update.zip

Figure: KBO download

KBO

Companies Only active companies
Registration required
Monthly updated (free)
Dayly updated web API (paying)

CSV format Current status + Monthly changes

- Name, language, Legal form, Ondernemingsnummer, Startdate
- Status of the company
- Address + contact information (siège/zetel + other locations) like tel/website/e-mail
- Name, address
- Activities + NACE codes

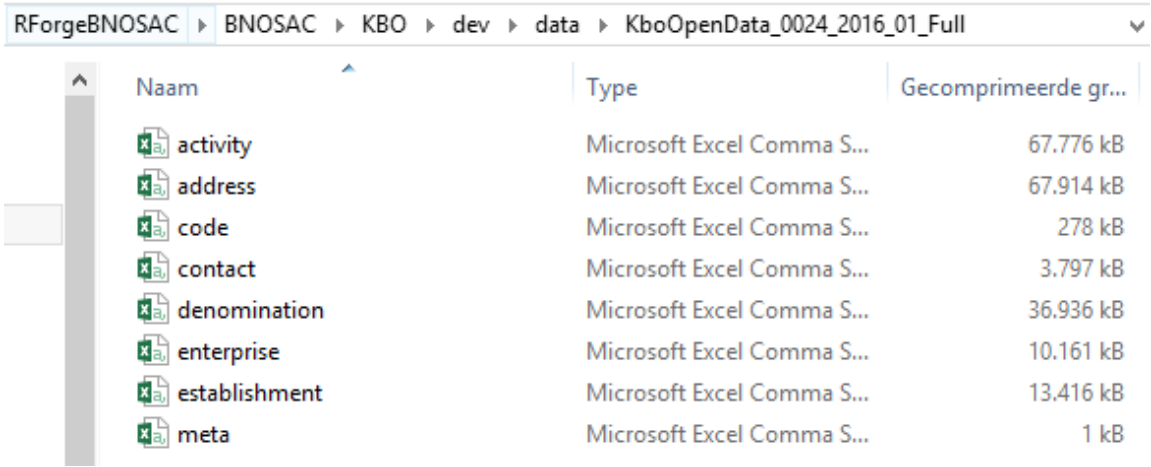


Figure: KBO files

KBO - example

Example of entity 2.154.581.321

```
entity <- "2.154.581.321"
load("data/kbo_example.RData")
##
## Name
##
kbo_example$denomination

   entitynumber language typeofdenomination
1 2.154.581.321        2                 003
        denomination
1 Python Predictions

##
## Enterprises
##
kbo_example$establishment

   entitynumber  startdate enterprisenumber
1 2.154.581.321 2006-07-01     0881.095.342

kbo_example$enterprise

  enterprisenumber status juridicalsituation
1     0881.095.342 Actief   Normale toestand
  typeofenterprise juridicalform  startdate
1    Rechtspersoon           015 2006-05-04

##
## Contacts
##
kbo_example$contact[, c("entitynumber", "contacttype", "value", "entitycontact.nl")]

   entitynumber contacttype                    value
1 2.154.581.321         TEL             0474/39.85.45
2 2.154.581.321         TEL             0486/72.09.68
3 2.154.581.321       EMAIL info@pythonpredictions.com
   entitycontact.nl
1 Vestigingseenheid
2 Vestigingseenheid
3 Vestigingseenheid
```

KBO - example

```
##
## Address
##
str(kbo_example$address)
```

```
'data.frame':   1 obs. of  16 variables:
 $ entitynumber    : chr "2.154.581.321"
 $ typeofaddress   : chr "BAET"
 $ countrynl       : chr NA
 $ countryfr       : chr NA
 $ zipcode         : chr "1150"
 $ municipalitynl  : chr "Sint-Pieters-Woluwe"
 $ municipalityfr  : chr "Woluwe-Saint-Pierre"
 $ streetnl        : chr "Roger Vandendriesschelaan"
 $ streetfr        : chr "Avenue Roger Vandendriessche"
 $ housenumber     : chr "9"
 $ box             : chr NA
 $ extraaddressinfo: chr NA
 $ datestrikingoff : Date, format: NA
 $ typeofaddress.de: chr NA
 $ typeofaddress.fr: chr "Unité d'établissement"
 $ typeofaddress.nl: chr "Vestigingseenheid"
```

```
##
## NACE codes
##
cat(unique(kbo_example$activity$nace.nl[kbo_example$activity$naceversion == 2008]), sep = "\n")
```

```
Overige adviesbureaus op het gebied van bedrijfsbeheer; adviesbureaus  op het gebied van bedrijfsvoering
Markt- en opinieonderzoekbureaus
Onderzoek naar de omvang van de markt voor bepaalde producten, de ontvangst en de bekendheid van producten en koo
Statistische analyses van de resultaten van het onderzoek
Overige zakelijke dienstverlening, n.e.g.
```

bnosac

open analytical helpers

Government other

The parlement

- De Kamer / La chambre open data.
- Questions / Answers of people in parlement
- Webservice at `http://data.dekamer.be`

DEPOTDAT	AUT	TITF	TEXTQF
20170227	Jean-Marc Nollet, Ecolo-Groen	Le prix des médicaments innovants.	En France, le Conseil économique, social et environ…
20170224	Roel Deseyn, CD&V	Les services par contournement.	Les opérateurs ont plusieurs droits et obligations et i…
20170224	Marco Van Hees, PTB-GO!	Le financement extérieur et les élections (QO 15590).	Je me permets de reposer cette question que mon c…
20170222	Gautier Calomne, MR	Start-up. - Commande publique.	L'innovation dans la commande publique ouvre aux s…
20170222	Jean-Marc Nollet, Ecolo-Groen	Les ondes des téléphones portables.	En Europe, pour les téléphones portables, une régle…
20170222	Gautier Calomne, MR	Start-up. - Commande publique.	L'innovation dans la commande publique ouvre aux s…
20170222	Françoise Schepmans, MR	Les contrôles en matière de délivrance de passeports.	La lutte contre le terrorisme passe notamment par le …
20170221	Benoît Friart, MR	Épanouissement professionnel.	C'est un constat indéniable, le monde du travail est e…
20170221	Françoise Schepmans, MR	Les contrôles en matière de délivrance de passeports.	La lutte contre le terrorisme passe notamment par le …
20170221	Barbara Pas, VB	Les condamnations de la Belgique par la Cour de just…	1. Pourriez-vous fournir un récapitulatif de toutes les…
20170221	Barbara Pas, VB	Les condamnations de la Belgique par la Cour de just…	1. Pourriez-vous fournir un récapitulatif de toutes les…
20170220	Jean-Marc Nollet, Ecolo-Groen	Sites de vente en ligne de billets d'avion.	En France, la Direction générale de la concurrence, …
20170220	Barbara Pas, VB	IBPT. - Infraction aux lois linguistiques.	Je souhaite revenir sur votre réponse à ma question …

Figure: De Kamer/La chambre

The parlement - example PTB data

```
load("data/dekamer.RData")
x <- subset(dekamer$antwoorden_data, PARTY == "PTB-GO!")
library(RDRPOSTagger)
library(tokenizers)
library(data.table)
m <- rdr_model(annotation = "UniversalPOS", language = "UD_French")
words <- lapply(x$TEXTQF, FUN=function(txt){
  rdr_pos(m, tokenize_sentences(txt, strip_punctuation = TRUE)[[1]])
})
words <- rbindlist(words)
```

Example, visualising words of the PTB

```
library(wordcloud)
x <- subset(words, word.type %in% "NOUN")
x <- table(tolower(x$word))
wordcloud(names(x), freq = x, scale=c(3, .75), min.freq = 10, max.words = 75,
          random.order = FALSE, colors = brewer.pal(6,"Dark2"))
```

Open data specific from NGI

`http://inspire.ngi.be/download-free`

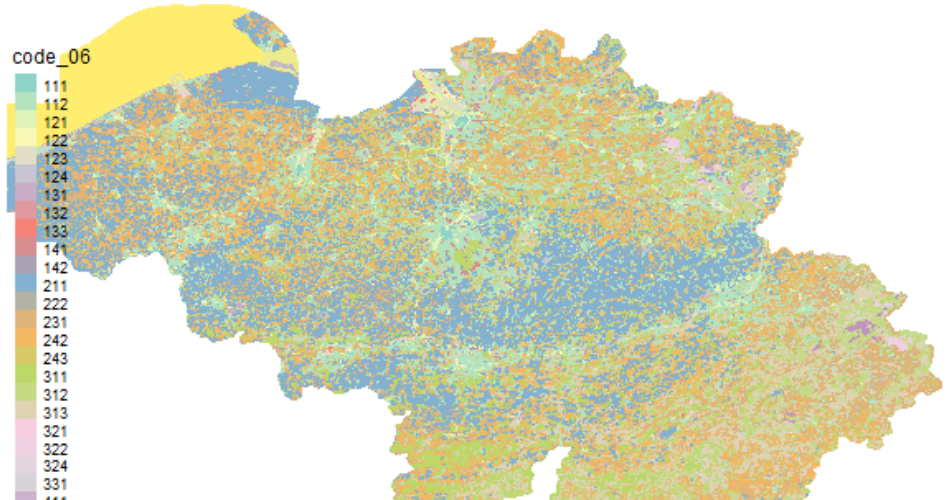
Corine Land Cover From 2006 and 2012

- 32 classes from the CORINE Land Cover legend which are present un Belgium
- shapefile format, minimum map unit 25 hectares

High Resolution Layers From 2012

- Imperviousness, Tree cover density, Forest type, Permanent grasslands, Wetlands and Permanent waterbodies
- tiff format

```
library(tmap)
corine <- read_shape("data/CLC06_BE/CLC06_BE.shp")
str(corine@data)
tm_shape(corine) + tm_fill("code_06")
```



EuroGlobalMap

- http://www.eurogeographics.org/products-and-services/euroglobalmap
- 45 countries and territories in the European region.
- Themes: Administrative boundaries, The water network, Tranport networks, Settlements, Named locations
- opendata as ESRI database, scale 1:1000000

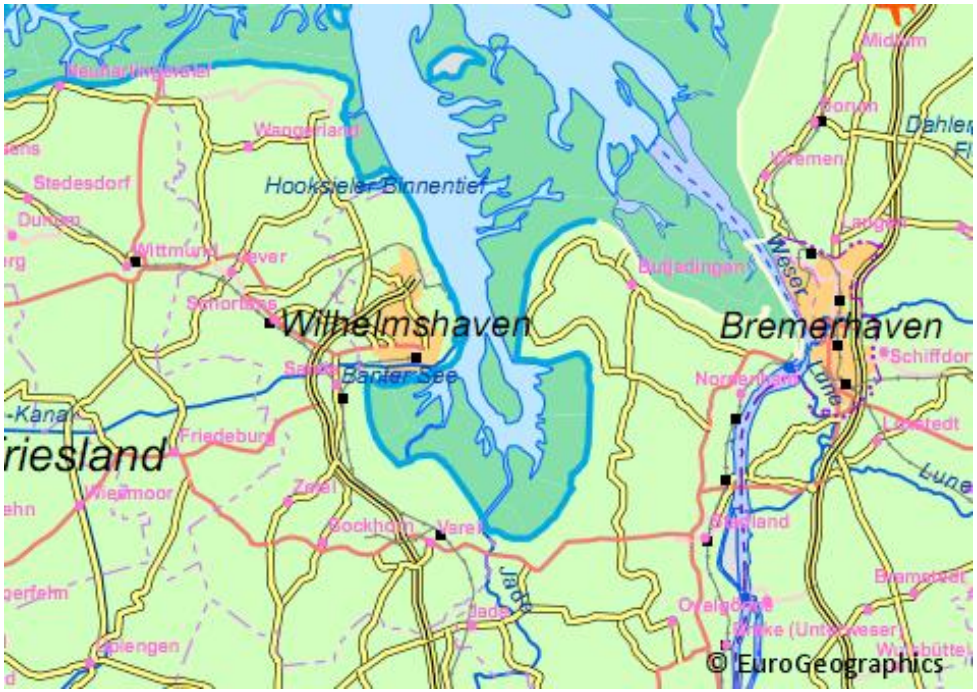


Figure: EuroGlobalMap

LEX data - http://belgielex.be

Belgisch Staatsblad

http://justitie.belgium.be/nl/overheidsdienst_justitie/organisatie/belgisch_staatsblad

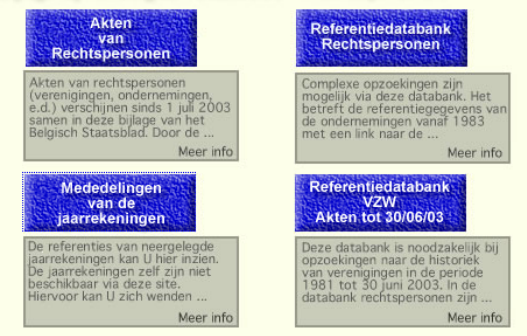


Figure: Rechtspersonen

- Staatsblad: http://www.ejustice.just.fgov.be/cgi/welcome.pl
- Rechtspersonen: http://www.ejustice.just.fgov.be/tsv_pub/index_n.htm
- Referentiebank rechtspersonen: http://data.gov.be/en/node/14728
- Akten van rechtspersonen: http://data.gov.be/en/node/14725
- Referentiebank VZW: http://data.gov.be/en/node/14727
- Jaarrekeningen: http://data.gov.be/en/node/14726

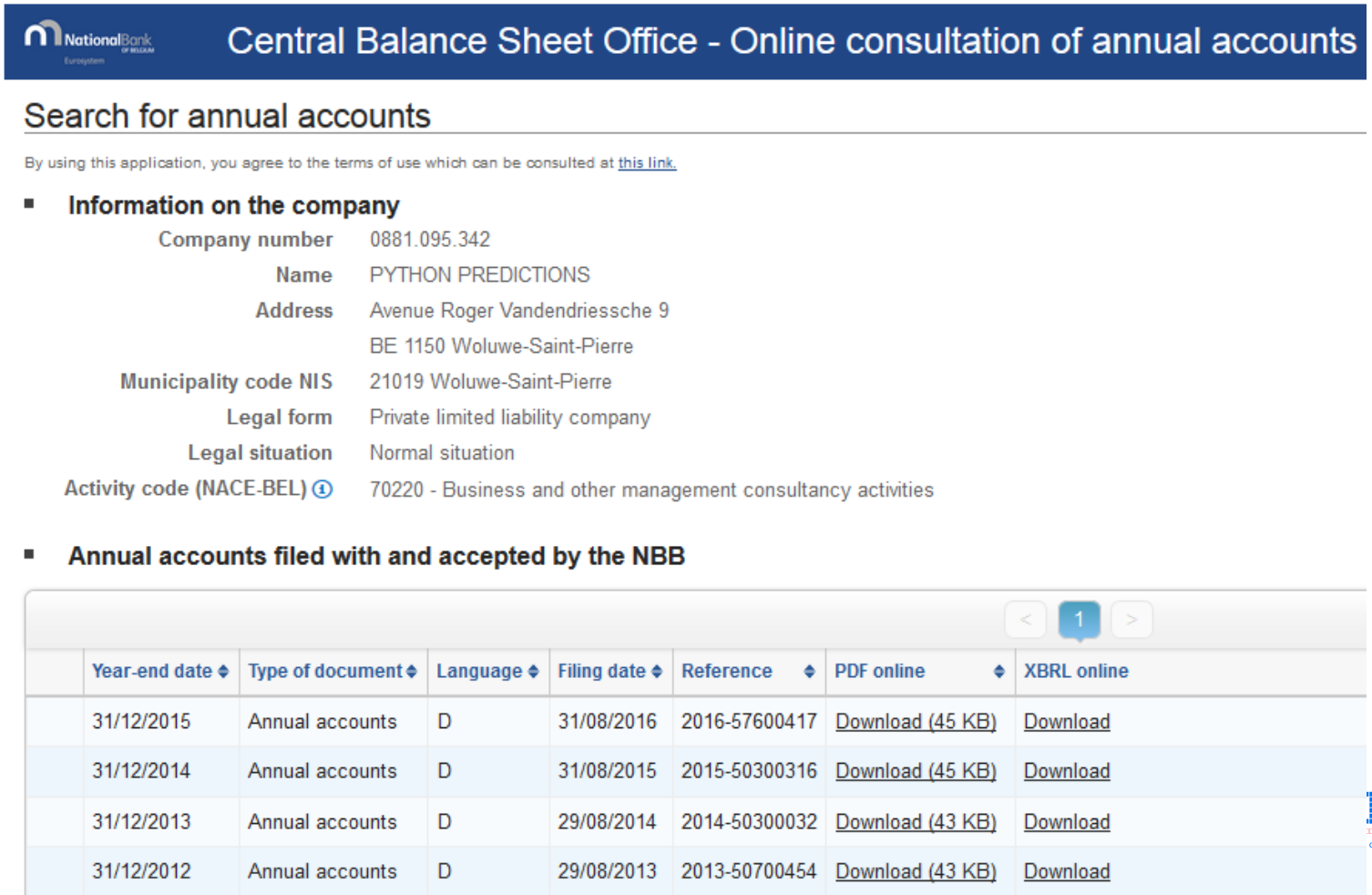
Jurisdiction Arresten & rechterlijke beslissingen

http://data.gov.be/en/node/14881

Financial statements

- Nationale Bank van Belgie
- Public search `http://cri.nbb.be/bc9/web/catalog?execution=e1s1`
- XBRL format
- Under following terms:

`https://www.nbb.be/en/central-balance-sheet-office/consultation-annual-accounts/terms-use?l=en`



National Bank of Belgium — Central Balance Sheet Office - Online consultation of annual accounts

Search for annual accounts

By using this application, you agree to the terms of use which can be consulted at this link.

- **Information on the company**

Company number	0881.095.342
Name	PYTHON PREDICTIONS
Address	Avenue Roger Vandendriessche 9
	BE 1150 Woluwe-Saint-Pierre
Municipality code NIS	21019 Woluwe-Saint-Pierre
Legal form	Private limited liability company
Legal situation	Normal situation
Activity code (NACE-BEL)	70220 - Business and other management consultancy activities

- **Annual accounts filed with and accepted by the NBB**

< 1 >

Year-end date	Type of document	Language	Filing date	Reference	PDF online	XBRL online
31/12/2015	Annual accounts	D	31/08/2016	2016-57600417	Download (45 KB)	Download
31/12/2014	Annual accounts	D	31/08/2015	2015-50300316	Download (45 KB)	Download
31/12/2013	Annual accounts	D	29/08/2014	2014-50300032	Download (43 KB)	Download
31/12/2012	Annual accounts	D	29/08/2013	2013-50700454	Download (43 KB)	Download

RSZ data

- Rijksdienst voor Sociale Zekerheid
- Datawarehouse: `https://www.ksz-bcss.fgov.be/nl/dwh/homepage/index.html`
- No Open Data, only for research purposes: `https://www.ksz-bcss.fgov.be/nl/dwh/dwh_page/content/websites/datawarehouse/data/ad_hoc_demands/description_general_framework.html`

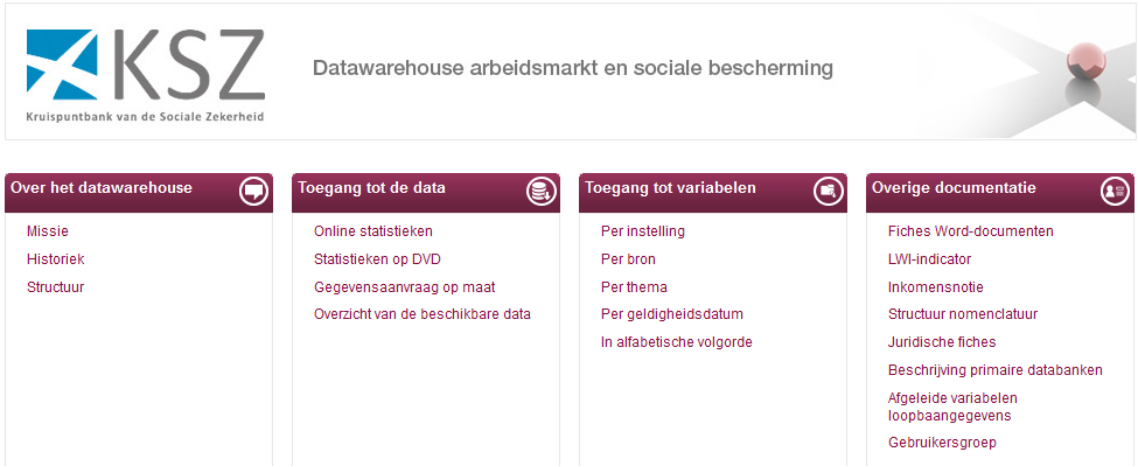


Figure: RSZ DWH

Belgium - WAL/FLA/BXL + Belgium

WAL/FLA/BXL + Belgium No central place

- Belgium: `http://data.gov.be` + `https://www.geo.be` (since 2017)
- Brussels: `http://opendatastore.brussels`
- Flanders: `http://opendata.vlaanderen.be` + `http://www.geopunt.be/catalogus`
- Wallonie: `http://data.digitalwallonia.be` + `http://geoportail.wallonie.be`

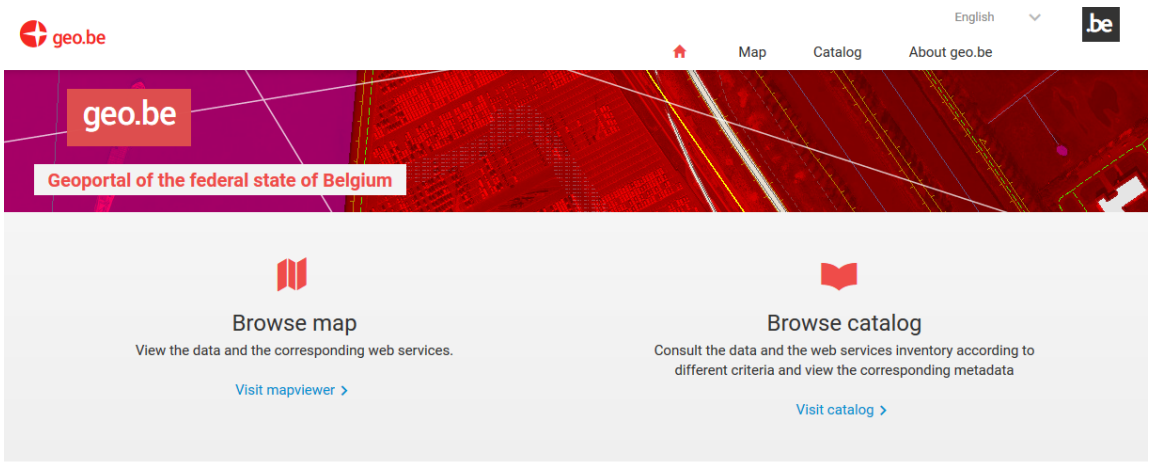


Figure: geo.be

b nosac
open analytical helpers

Text Analytics

DBPedia

- DBpedia `http://wiki.dbpedia.org`
- Large, Multilingual, Semantic Knowledge Graph, structuring Wikipedia
- DBpedia Spotlight: annotate mentions of DBpedia resources in text
- License: CC BY-SA. Data avaiable in RDF / CSV / JSON

Figure: DBPedia

- Use cases: Entity extraction, entity detection, name resolution / disambiguation. Annotation with > 320 DBpedia classes.
- Easily launched as docker application: `https://hub.docker.com/u/dbpedia`

```
library(httr)
library(rjson)
library(magrittr)
txt <- "Paris est la capitale du monde et Marlene Dietrich n'est pas né a ZiZi. J'aime la boisson Cointreau"
response <- POST(url = "http://datatailor.be:2224/rest/annotate", body = list(text = txt), encode = "form")
txt_annotation <- response$content %>% rawToChar %>% fromJSON
View(txt_annotation$Resources)
```

@URI	@support	@types	@surfaceForm	@offset	@similarityScore
http://fr.dbpedia.org/resource/Paris	150209	DBpedia:Location,Schema:Place,DBpedia:Place,DBpe...	Paris	0	0.9999966152235388
http://fr.dbpedia.org/resource/Marlene_Dietrich	625	DUL:Agent,DBpedia:Agent,DUL:NaturalPerson,Wikidat...	Marlene Dietrich	34	1.0
http://fr.dbpedia.org/resource/Cointreau	63		Cointreau	90	0.9999999915838487

Figure: DBPedia Spotlight Annotation

OpenNER lexicons

- OpenNER `http://www.opener-project.eu`
- Detecting sentiments, opinions and named entities in texts
- Dutch, English, German, French, Spanish and Italian
- Available as webservices: `http://www.opener-project.eu/webservices`
- Part of this is open sentiment lexicons:

`https://github.com/opener-project/VU-sentiment-lexicon`

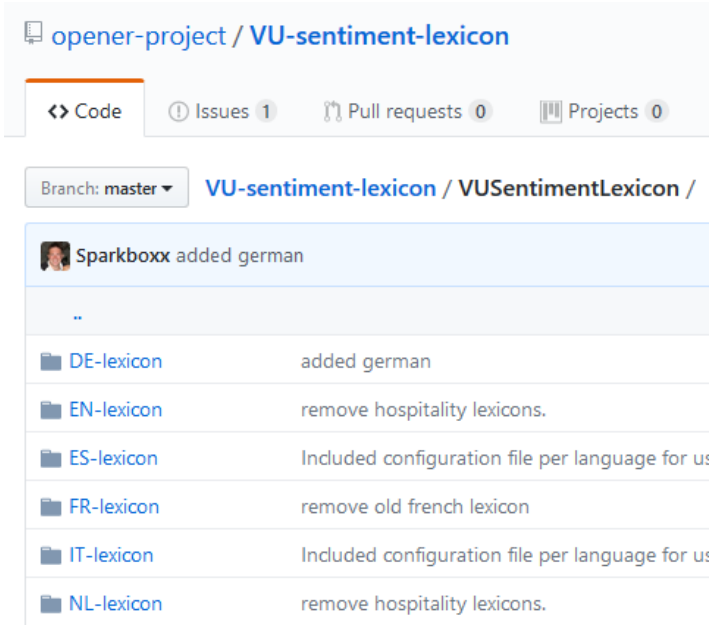


Figure: OpenNER lexicons

Word embeddings trained on Wikipedia

Word vectors are mainly used for analogies, machine translations and as input to predictive models or to explain word meaning.

fastText Based on Facebook's fastText

- Trained on Wikipedia
- https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
- 90 languages
- word vectors of dimension 300

Glove Based on Glove (extension of Google word2vec)

- Trained on Wikipedia 2014 + Gigaword 5, Common Crawl, Twitter corpus
- https://nlp.stanford.edu/projects/glove
- All languages in the data
- word vectors of dimension 50 up to 300

b nosac
open analytical helpers

Weather

Weather - Belgium

KMI open data `http://opendata.meteo.be`

- Clouds based on Lidar data of 4 stations (archives of the last two months)
- Weather forecasts based on Alaro (4km grid)
- Extreme climate Belgium data (e.g. for insurance)

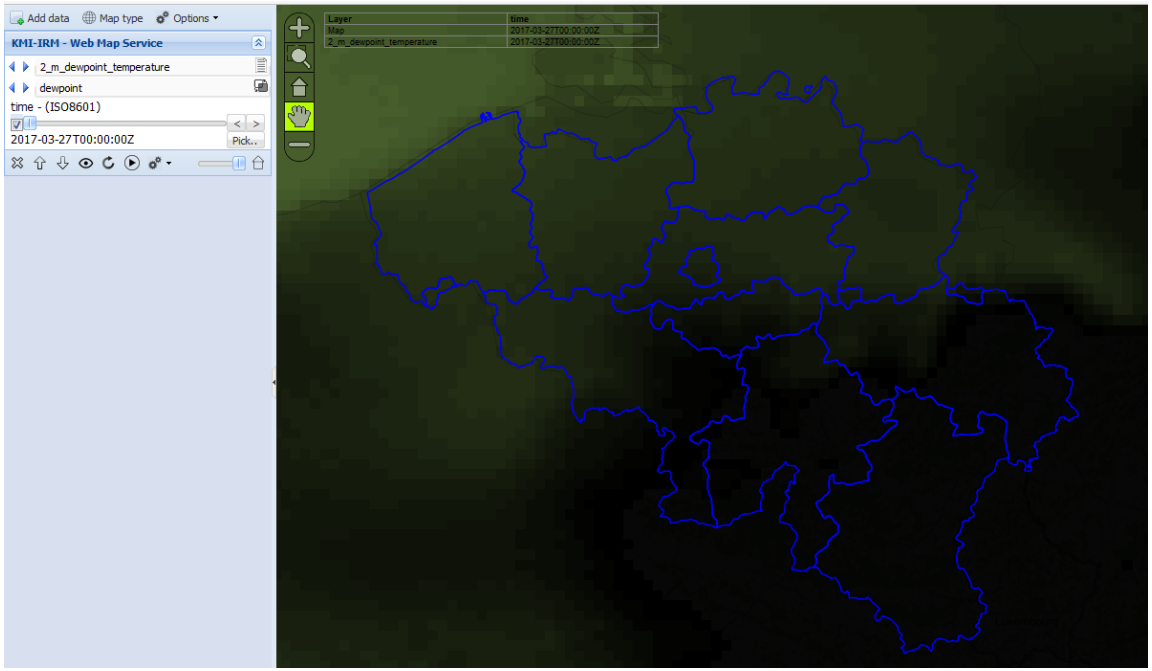


Figure: Weather forecast KMI

Weather - Belgium

METAR reports Automatic weather stations worldwide (at airports)

- Used in aggregated format by metereologists
- Contains temperature, dew point, wind direction and speed, precipitation, cloud cover and heights, visibility, and barometric pressure, precipitation amounts, lightning, a pilot report, colour states, runway visual range (RVR)
- E.g. Brussels: `ftp://tgftp.nws.noaa.gov/data/observations/metar/stations/EBBR.TXT`
- Looks like this: `EBBR 271420Z 13007KT 080V150 CAVOK...`
- Parse with RMETAR R package

```
library(RMETAR)
x <- decodemetar("EBBR 271420Z 13007KT 080V150 CAVOK 17/00 Q1021 NOSIG")
cat(x$printout.mdsplib)
```

```
/*****************************************/
/*    THE DECODED METAR REPORT FOLLOWS     */
/*****************************************/

STATION ID          : EBBR
OBSERVATION DAY     : 27
OBSERVATION HOUR    : 14
OBSERVATION MINUTE  : 20
WIND DIRECTION      : 130
WIND SPEED          : 7
WIND UNITS          : KT
MIN WIND DIRECTION  : 80
MAX WIND DIRECTION  : 150
TEMP. (CELSIUS)     : 17
D.P. TEMP. (CELSIUS): 0
ALTIMETER (PASCALS) : 1021
CAVOK               : TRUE
```

Weather - Worldwide

- The number of automatic weather stations with METAR reports is large.
- For Belgium: Antwerp International Airport (Deurne), Beauvechain Air Base, Kleine Brogel Air Base, Brussels Airport, Brussels South Charleroi Airport, Chièvres Air Base, Schaffen Airport, Koksijde Air Base, Florennes Air Base, Elsenborn Air Base, Liège Airport, Ostend-Bruges International Airport

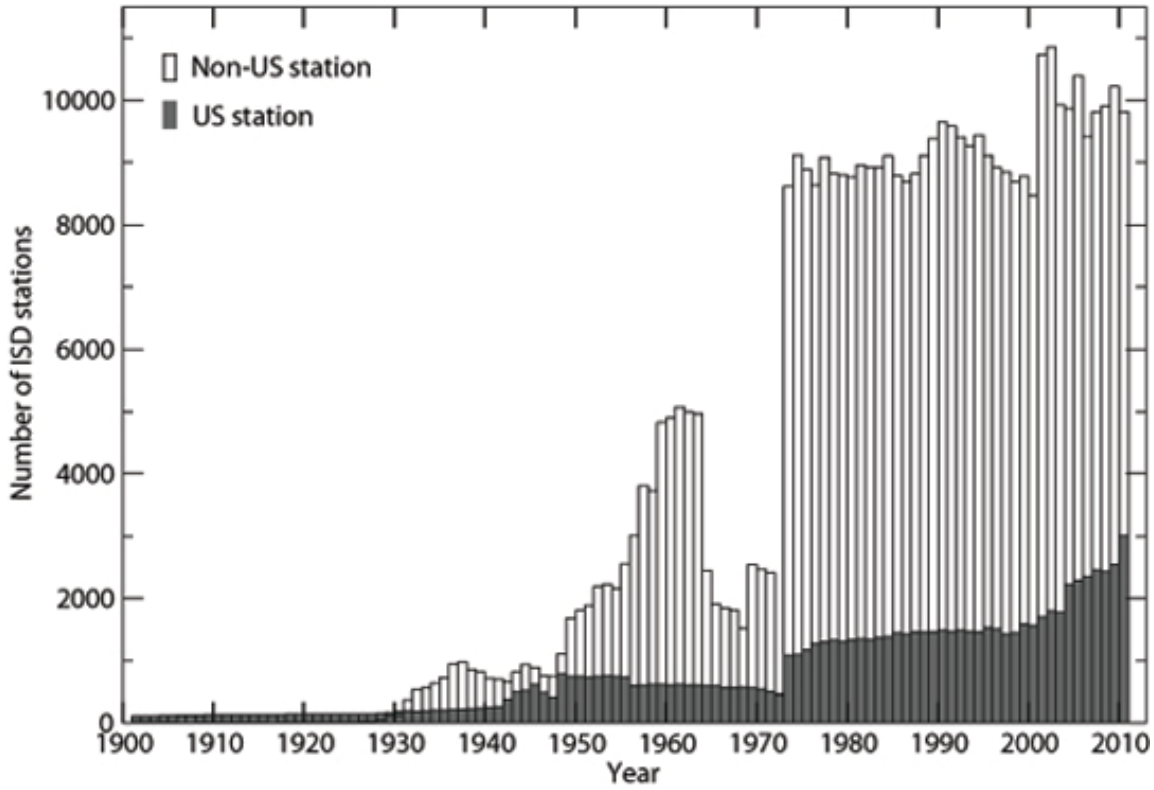


Figure: Number of stations

Weather - Worldwide

Current weather Worldwide

- Current weather of all weather stations worldwide: through **Aviation Digital Data Service** https://www.aviationweather.gov/adds
- Web service or raw txt files: example for Brussels Airport:

  https://www.aviationweather.gov/adds/dataserver_current/httpparam?dataSource=metars&requestType=retrieve&format=xml&stationString=EBBR&hoursBeforeNow=36

- Example of current weather data in Belgium

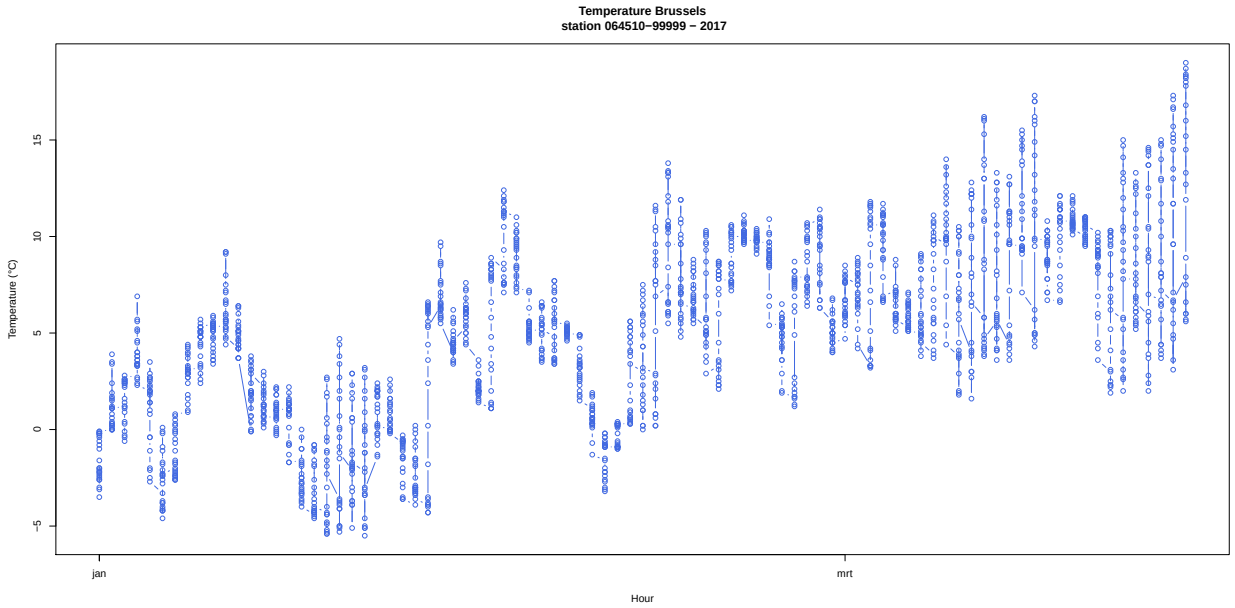
```
## Get all airports - Open Airport data:
##   http://ourairports.com/data or use https://www.aviationweather.gov/docs/metar/stations.txt
airports <- fread("http://ourairports.com/data/airports.csv")
airports <- subset(airports, iso_country %in% "BE")
airports <- setDF(airports)
## Get CURRENT weather at METAR locations (airports)
x <- fread("https://www.aviationweather.gov/adds/dataserver_current/current/metars.cache.csv", skip = 5)
x <- subset(x, station_id %in% airports$gps_code)
x <- merge.data.frame(airports[, c("gps_code", "name", "iata_code")], x,
                      by.x = "gps_code", by.y = "station_id", suffixes = c(".airport", ".weather"))
str(x)

'data.frame':   12 obs. of  46 variables:
 $ gps_code                    : chr  "EBAW" "EBBE" "EBBL" "EBBR" ...
 $ name                        : chr  "Antwerp International Airport (Deurne)" "Beauvechain Air Base" "Kleine Br
 $ iata_code                   : chr  "ANR" "" "" "BRU" ...
 $ raw_text                    : chr  "EBAW 301150Z 20011KT 170V240 CAVOK 21/08 Q1023 NOSIG" "EBBE 301203Z 1900
 $ observation_time            : chr  "2017-03-30T11:50:00Z" "2017-03-30T12:03:00Z" "2017-03-30T12:03:00Z" "201
 $ latitude                    : num  51.2 50.8 51.2 50.9 50.5 ...
 $ longitude                   : num  4.47 4.77 5.47 4.5 4.45 3.82 5.07 2.63 4.65 6.17 ...
 $ temp_c                      : num  21 20 21 21 21 20 20 21 19 17 ...
 $ dewpoint_c                  : num  8 8 7 8 8 8 8 10 7 5 ...
 $ wind_dir_degrees            : int  200 190 200 210 210 180 210 210 170 200 ...
 $ wind_speed_kt               : int  11 9 7 6 7 8 9 13 9 6 ...
 $ wind_gust_kt                : int  NA NA NA NA NA NA NA NA NA NA ...
 $ visibility_statute_mi       : num  6.21 6.21 6.21 6.21 6.21 6.21 6.21 6.21 6.21 6.21 ...
```

Weather - Worldwide

Historic weather data Worldwide

- METAR data at **hour** level for automatic weather stations
- **Since the start of recording history**:
  `ftp://ftp.ncdc.noaa.gov/pub/data/noaa`
- History period for each location: `ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-history.txt`, for Brussels from 1930 on.

```
require(readr)
x <- read_fwf("ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/2017/064510-99999-2017.gz",
  fwf_widths(widths = c(5, 3, 3, 2, 6, 6, 6, 6, 6, 6, 6, 6),
  col_names=c("year", "month", "day", "hour", "temperature.air", "dewpoint", "sea.level.pressure",
              "wind.direction", "wind.speed.rate", "sky.condition", "precip.1hour", "precip.6hour")))
x$timepoint <- as.POSIXct(sprintf("%s-%s-%s %s:00:00", x$year, x$month, x$day, x$hour))
plot(x$timepoint, x$temperature.air/10, type = "b", col = "royalblue",
     ylab = "Temperature (°C)", xlab = "Hour", main = "Temperature Brussels\n station 064510-99999 - 2017")
```



b nosac

open analytical helpers

Look beyond Belgium

GDELT

- http://www.gdeltproject.org
- GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.
- CSV format / Graph based format / Google BigQuery format
- Data since 1979, integrated with Google Vision API
- Applications vary from Location Sentiment API to mentioning names/organisations/people or trending reports

Figure: GDELT project

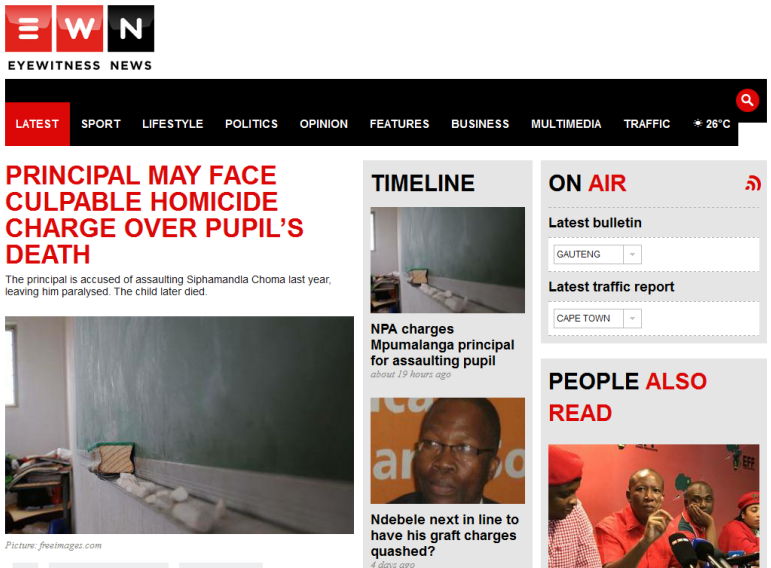
GDELT

- Ready-made analysis based on GDELT: `http://analysis.gdeltproject.org`
- R package gdeltr2: `https://github.com/abresler/gdeltr2`
  - The GDELT Events Database [EVENTS]: Global Events, 1979 to present.
  - The GDELT Global Knowledge Graph [GKG] : GDELT's Knowledge Graph, April 2013 to present.
  - The GDELT Full Text API Full Text API: Full text search for all monitored sources within a 24 hour window. Output includes raw data, sentiment, and word counts.
  - The GDELT Visual Knowledge Graph VGKG: Google Cloud Vision API output for every indexed piece of GKG media.

```
library(gdeltr2)
library(readr)
## Event data ready for download
historic_datasets <- get_urls_gdelt_event_log()
str(historic_datasets)

Classes 'tbl_df', 'tbl' and 'data.frame':   1570 obs. of  7 variables:
 $ idHash         : chr  "f6fcb7e955e35f93c9dae427c07b545d" "7fb1b6ff781aea3d441abf512dc5e0f9" "6b4341632bcc8cff1
 $ dateData       : Date, format: NA ...
 $ isZipFile      : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ isDaysData     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ urlData        : chr  "http://data.gdeltproject.org/events/GDELT.MASTERREDUCEDV2.1979-2013.zip" "http://data.g
 $ periodData     : chr  "1979-2013" "1979" "1980" "1981" ...
 $ idDatabaseGDELT: chr  "EVENTS" "EVENTS" "EVENTS" "EVENTS" ...
```

GDELT

```
## Trending terms last 15 minutes
x <- get_data_ft_trending_terms()
head(x = x, 3)

# A tibble: 3 × 3
      nameTerm isGDELTTag        datetimeData
         <chr>      <lgl>              <dttm>
1  A Abu Zneid      FALSE 2017-03-30 14:15:00
2 A Act Center      FALSE 2017-03-30 14:15:00
3     A Agency      FALSE 2017-03-30 14:15:00


## Current news items
current <- read_delim("http://data.gdeltproject.org/gdeltv2/lastupdate.txt", delim = " ", col_names = FALSE)
current_events <- grep("export", current$X3, value=TRUE)
download.file(current_events, destfile = "export.CSV.zip")

x <- read_delim("export.CSV.zip", delim = "\t", col_names = FALSE)
url_news <- x$X61[1]
url_news

[1] "http://ewn.co.za/2017/03/30/school-principal-may-face-culpable-homicide-charge-over-pupil-s-death"
```



Common Crawl - all websites in the world

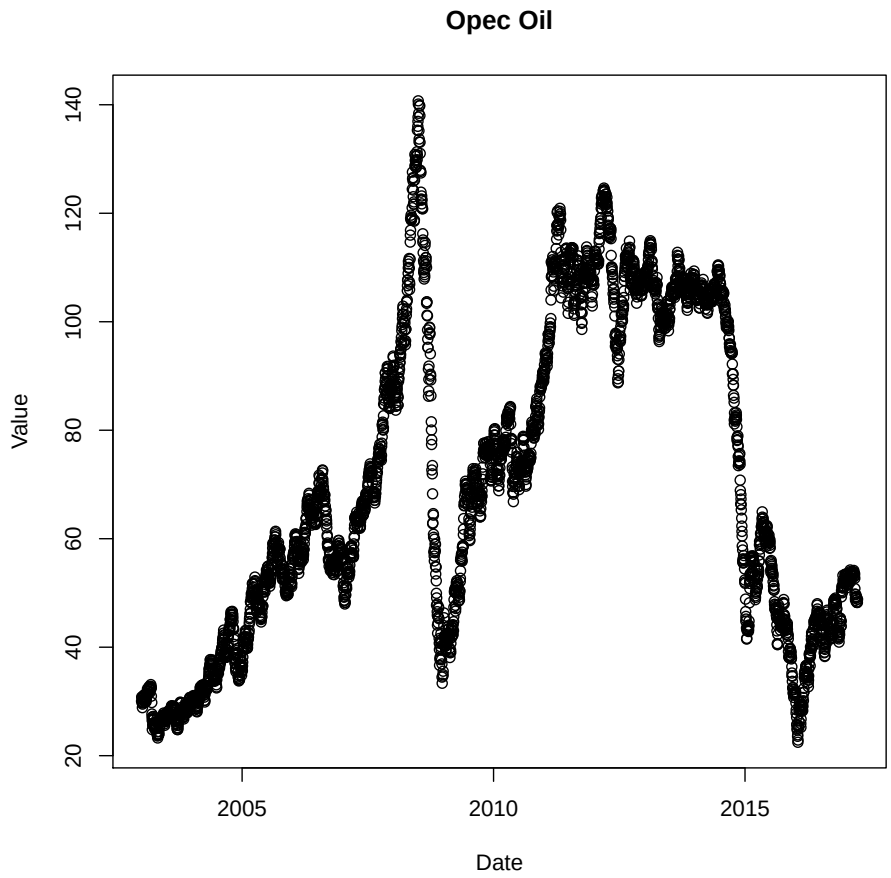
- `http://commoncrawl.org`
- open repository of web crawl data that can be accessed and analyzed by anyone.
- The Common Crawl corpus contains petabytes of data collected over the last 7 years.
- It contains raw web page data, extracted metadata and text extractions from these html Files ~ WARC/WAT/WET
- WARC files which store the raw crawl data
- WAT files which store computed metadata for the data stored in the WARC
- WET files which store extracted plaintext from the data stored in the WARC

Figure: Common Crawl

Stock market data

- Quandl (open stock data and economic data, max 50 calls a day)

```
opecoil <- Quandl::Quandl("OPEC/ORB")
plot(opecoil, main = "Opec Oil")
```



Need more inspiration

- Landsat 8 images for the whole earth observation (satellite launched in 2013) https://landsat.usgs.gov/about-landsat . In GeoTIFF format.
- Open multilingual wordnet: http://compling.hss.ntu.edu.sg/omw
- Wikipedia data: Wikidata https://www.wikidata.org
- Wikimedia commons (open images/sounds/video): https://commons.wikimedia.org
- Biodiversity data from Belgium: http://www.gbif.org
- Environmental data: https://github.com/ropensci/opendata
- Eurostat database: http://ec.europa.eu/eurostat
- Open flights data: http://openflights.org/data.html
- Multimedia Commons: auto/video/images from Flickr/Yahoo images, videos and sound http://multimediacommons.org . Open Images: openimages.org
- 1000 genomes project http://www.internationalgenome.org , Protein sequences: http://www.uniprot.org
- https://github.com/caesar0301/awesome-public-datasets

b nosac

open analytical helpers

Contact

www.bnosac.be

Need more, send a message at
www.bnosac.be/index.php/contact/get-in-touch

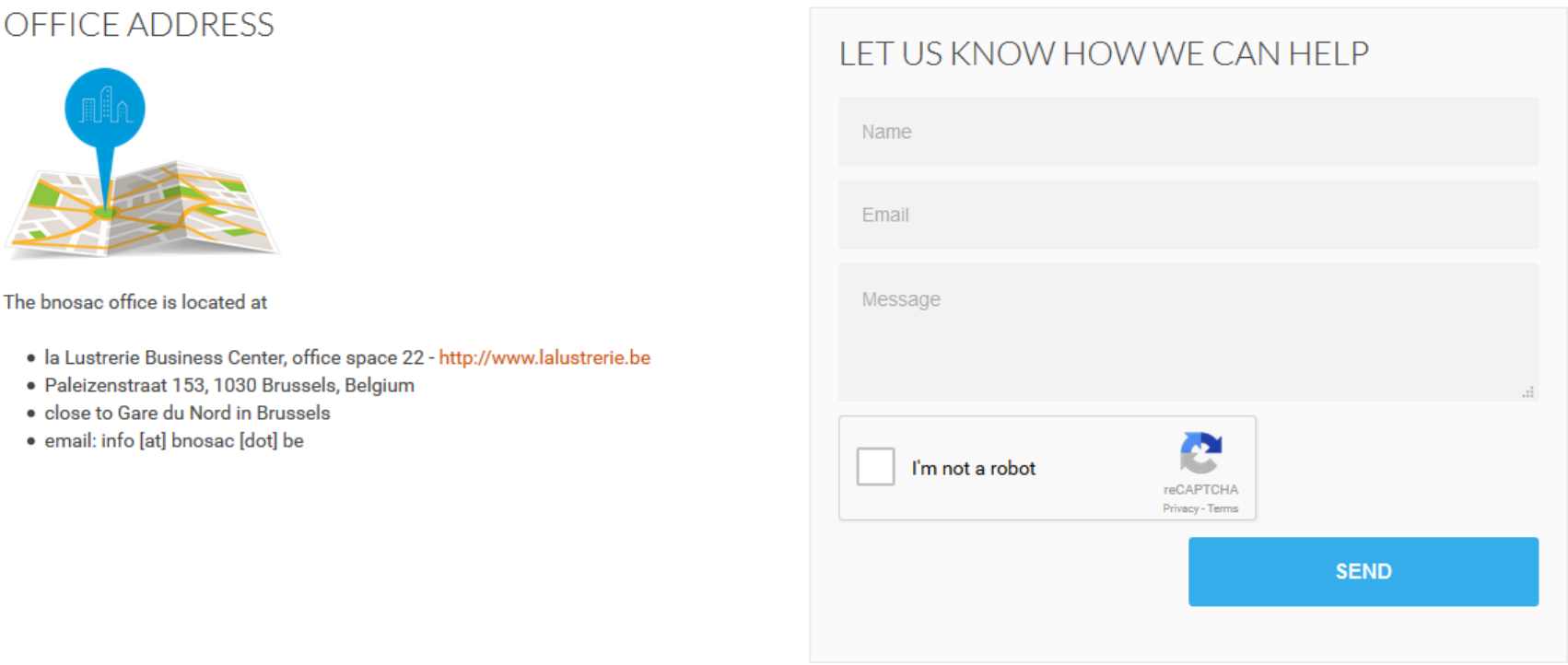


Figure: Contact www.bnosac.be