Data on Cancer vs Environment in France – archeology of Epidemium season 1 and 2

Edouard Debonneuil, 2021-11-05

Abstract

I did some potholing ("spéléologie") in the wikis (Wiki Season 1 and Wiki Season 2) and githubs (https://github.com/Epidemium and https://github.com/EpidemiumOpenCancer) from Epidemium season 1 and season 2, in search for data used for projects similar to the Open Data & Environment challenge of Epidemium season 3. After quite bad news (must urls are not functional anymore), teaching us to be careful now, the few cases with data should save time for next Epidemium users. So leveraging from the work of others still seems efficient even when most of it is lost. In addition I searched through my personal computer from Epidemium season 1 and bring complementary data. Now, this peace of data sources is just what it is – it needs **to be investigated** as a complement of what will be found in data.gouv.fr

First gross analysis

Apart from Cancer Baseline and ELSE, no project **provides data or urls for data that are still available**!!

- ⇨ First conclusion: store urls on GitHub as well as associated data and code

Finer analysis

From ALL available info from ALL past Epidemium projects, I searched for links that still work for **France**, (with a tolerance for larger scopes when the data seems appropriate for the Challenge) and **apart from data.gouv.fr** links (such as https://www.data.gouv.fr/fr/datasets/qualite-de-l-air-nd/): the goal here is to find data sources that we wouldn't have seen otherwise, being immerged in data.gouv.fr

based on ELSE (mentioning it to reuse code, but they didn't provide code!):

- https://side.developpement-durable.gouv.fr/ et https://ree.developpement-durable.gouv.fr/ : à explorer !! nouveau lien de liens à présent non fonctionnel
- http://www.data.eaufrance.fr/ : absolument super
- https://www.eea.europa.eu/data-and-maps/data/air-pollutant-concentrations-at-station/ https://www.eea.europa.eu/data-and-maps/data/european-past-floods/ https://www.eea.europa.eu/data-and-maps/data/waterbase-uwwtd-urban-waste-water-treatment-directive-7 : kept in my list because it is aboslutely great, but at European level
- http://mobile-users.net/_antennes_mobiles.zip : where does this come from?!!
- https://ades.eaufrance.fr/ : données sur les eaux souterraines : super
- https://public.opendatasoft.com/explore/dataset/registre-francais-des-emission-polluantes-traitement-dechets/table/ wow.. strange
- https://data.oecd.org/fr/agroutput/consommation-de-viande.htm world level but very interesting

- https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-[…]lts-2012-or-nearest-year_health_glance_eur-2014-graph48-en : at a European level and in 2014, but it shows that similar reports should exist on that website
- Rerences without linked (I didn't take the time to search them): INCA 2 : Afssa (2009). Etude Individuelle Nationale des Consommations Alimentaires 2 (INCA 2) (2006-2007). Rapport Afssa. Consommation de légumes par sexe, âge et niveau d'éducation (%) - vague d'enquête 2008. Ce jeu de données provient d'un service public certifié Tableau 1 : Données interprétées Tableau 2 : Données brutes
- https://fr.openfoodfacts.org/ : **question**: is anyone sufficiently familiar with this website to know if it contains estimations of consumption by geographical area? (or 'just' facts on the specific food)

based on Cancer Baseline :

- https://wonder.cdc.gov/ : cancer, demographics, environment.. by age, county etc in the USA. Wow !
- http://eco.iarc.fr/ now leads to https://ecis.jrc.ec.europa.eu/ : this it the Y in Europe and various other countries but all country level - it used to be so much more granular :/. The entry page has links to websites that may be more granular - to investigate
- #20160221_Ci5_IncidenceAndPop_registry_level Contains cancer incidence and underlying population size, gy gender, year and age tranches for 300 registries worldwide Downloaded from http://ci5.iarc.fr/CI5plus/Pages/download.aspx . Great+ http://ci5.iarc.fr to be investigated further (in 2015-2016 it was indicating that new sections would open soon)
- https://www.gapminder.org/data/ : data at country level
- a great dataset https://github.com/Epidemium/Baseline/tree/master/DATA
- notably this list of putative risk factors with names and units, according to the LOINC referential: https://github.com/Epidemium/Baseline/blob/master/DATA/manual_assembly/Names%26Units.xlsx

Going through my computer directories of Epidemium Season 1
Joseph Lam and Beoit Choffin sent me their memorandum, with code

Notebook_C4C_France.html

- http://www.ecosante.fr/index2.php?base=DEPA&langs=FRA&langh=FRA&source=800190 : a well-known ressource for French open data, available at department, but it is not maintained anymore (data from 1968 to 2013)

- https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=129&id_rubrique=52 : data from MeteoFrance,, still available !! The website provides an access to one region and month at a time ==> need for a script... that is found in the above memorandum. It starts with this:

```
def download_file(date):
    import urllib

urllib.urlretrieve('https://donneespubliques.meteofrance.fr/donnees_libres/bulletins/BCM/{0}.pdf'.format(date), 'meteo/BCM{0}.pdf'.format(date))
    print("Completed")


#Nous bouclons sur les dates de publication des bulletins météo
date = 199901
while date < 201511:
    download_file(date)
    if date%100==12:
        date+=89
    else:
        date+=1
```

while searching for cancer in my computer

- I have the main dataset provided to Epidemium season 1 participants. → May I put it on GitHub ? (question for Epidemium) + it would be great if Epidemium provides the dataset that used to be in Epidemium season 2: http://qa.epidemium.cc/data/epidemiology_dataset/ (the directory structure is still there, but not documents).
- SEER in the USA provides a lot of cohort data on cancer, using a software : very granular knowledge, difficult to use however. One needs to ask for it indicating the reason for its use, in a few lines
- in France, causes of death are available at a very fine granularity here: http://cepidc-data.inserm.fr/inserm/html/index2.htm One needs to use ICD-10 codes to be precise on the cancer : https://ci5.iarc.fr/CI5I-X/Pages/cancer.aspx

Overall conclusion

Stepping on the shoulders of a variety of past Epidemium participants seems to be of high value, even here when most knowledge was lost. So for future persons to step on your work i) store code and references of data in github ii) as well as your retrieved data because online data easily disappears.

Also, it would be interesting to put the main datasets from seasons 1 and 2 on github.