HEALTH DATA HUB

Création d’un générateur de données de synthèse

11.03.2021

Sommaire

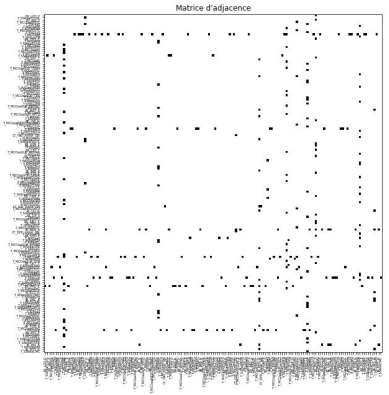
1. Approche théorique (1/2)


```
{
    "fields": [
        {
            "name": "BEN_NIR_PSA",
            "description": "Identifiant anonyme du patient dans le SNIIRAM",
            "type": "string",
            "nomenclature": "-",
            "length": "17",
            "format": "default",
            "constraints": {
                "minLength": 17,
```


Un point de départ: le schéma formel

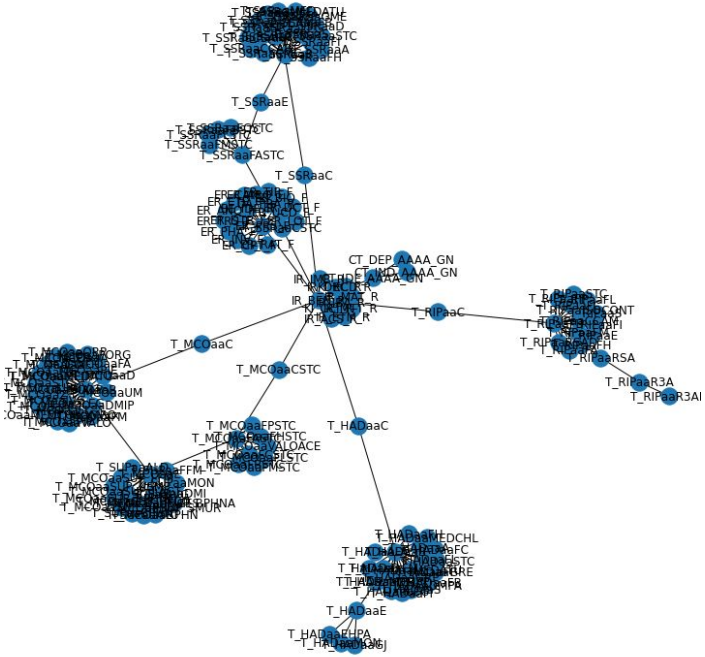
source	target	joint_var
T_HADaaFI	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFL	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFM	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFP	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaGJ	T_HADaaE	ETA_NUM_EPMSI => ETA_NUM

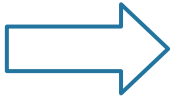
Du schéma formel aux liens entre les tables

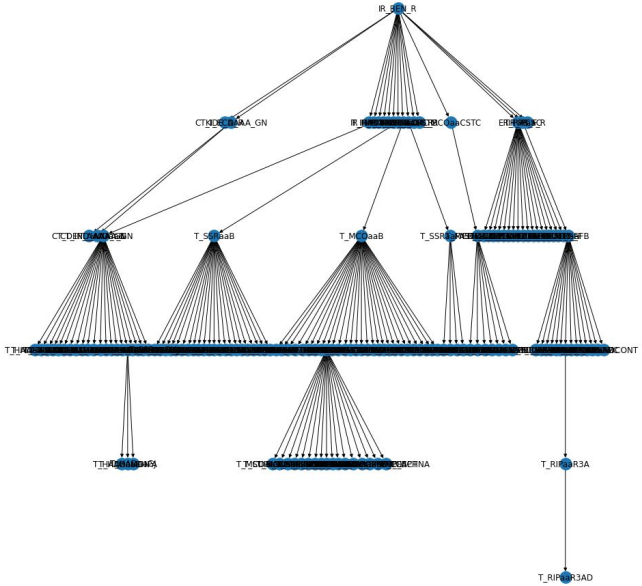


Des liens à la matrice d'adjacence

1. Approche théorique (2/2)







De la matrice d'adjacence, un graphe

Du graphe, un arbre

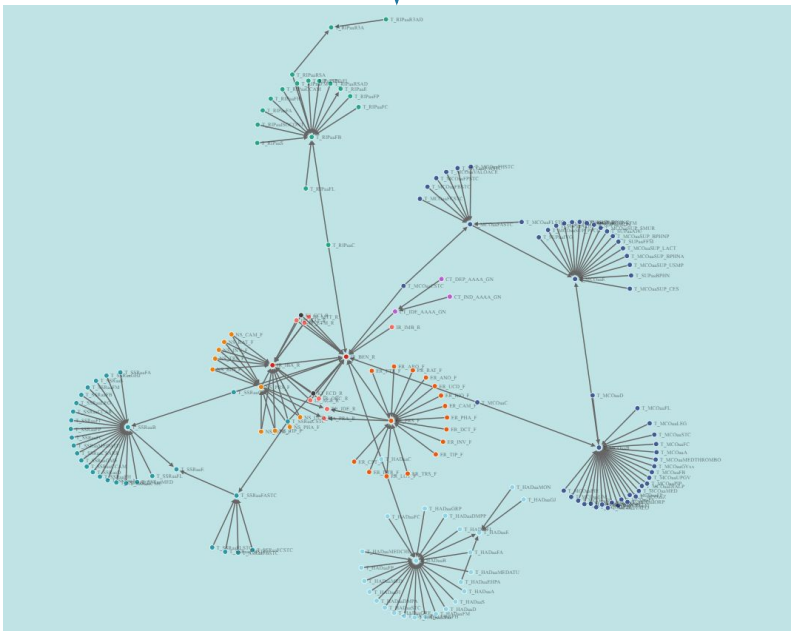
2. Implémentation (configuration du générateur)

snds.config 636 Bytes

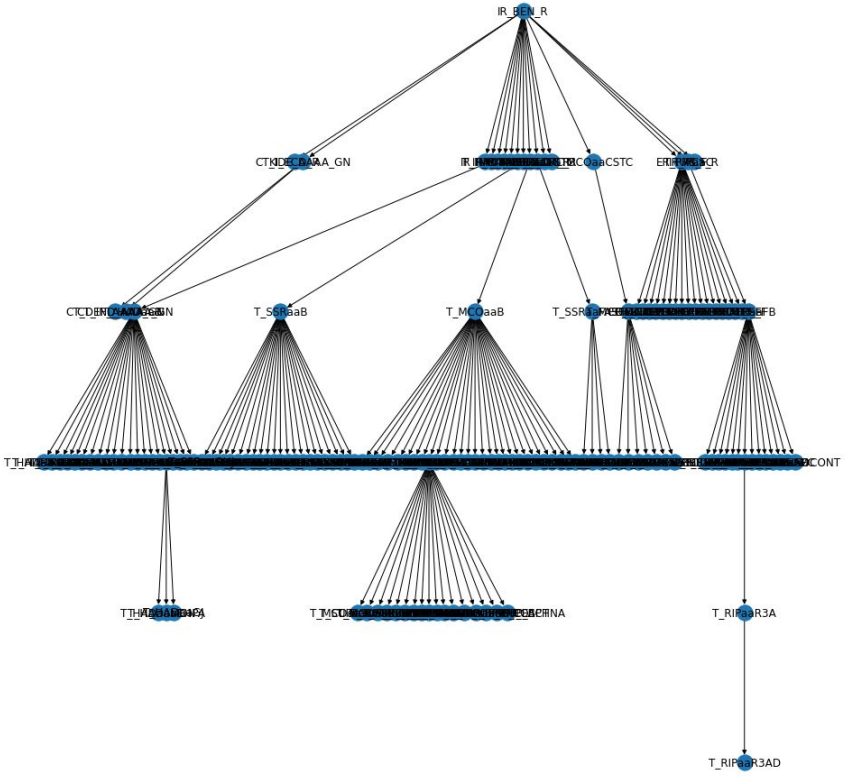
```
[BASE]
base_name = SNDS
#choose a root for every connected component of the data base which contains more than 2 tables
roots = IR_BEN_R
#fill only one parameter of n_beneficiaires, volume_beneficiaires (which is indicated in Mo)
n_beneficiaires = 10
#volume_beneficiaires = 8000
export_path = test_snds
#path2resources = src/resources

[SCHEMA MODIFIER]
#the following format is expected: table:variable:property:new value. Note that table, variable and property can be given as globstrings
# eg *:MY_DATE:type:datetime will convert the variable MY_DATE's type to "datetime" in ALL the tables
modifier1 = IR_BEN_R:BEN_IDT_ANO:length:4

```

4. Implémentation (ressources: schéma, nomenclatures)

Name
..
nomenclatures
schemas



2. Implémentation



Génération colonne par colonne

2. Implémentation - exemple: génération de T_MCOaaB

ETA_NUM	RSA_NUM
9000325451	151242
9000646567	135868

Récupérés depuis T_MCOaaC

ETA_NUM	RSA_NUM	AGE_ANN
9000325451	151242	65
9000646567	135868	42

Générés aléatoirement

ETA_NUM	RSA_NUM	AGE_ANN	...	DGN_PAL
9000325451	151242	65	...	X34018
9000646567	135868	42	...	S2200

Générés aléatoirement

Premier levier: La cohérence temporelle

On peut s'assurer que l'**ordre** des dates est cohérent, par exemple qu'une sortie d'hospitalisation a bien lieu après l'entrée correspondante.

→ comment trouver les couples de variables qui se correspondent?

Notre proposition: une recherche du **plus proche voisin**, avec une distance calculée à partir du nom de la variable et de sa description

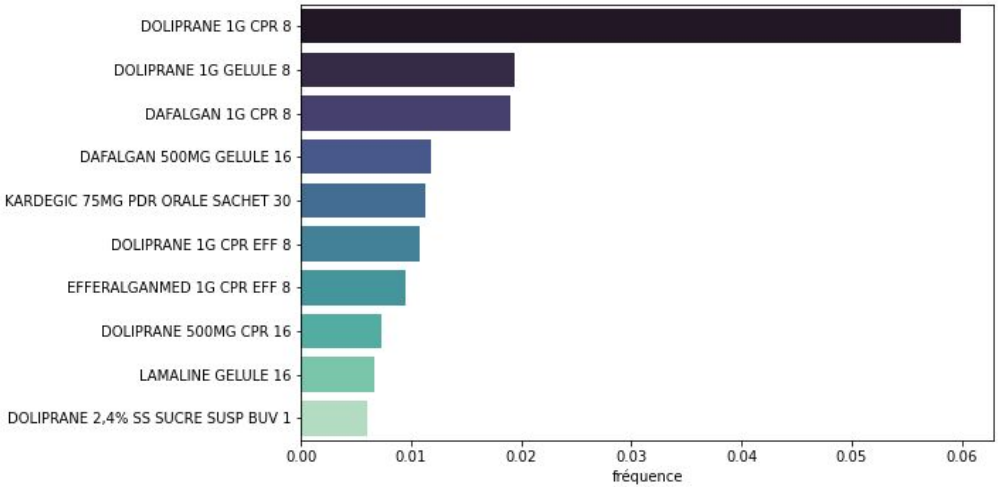
Pour $x_1 = (var_1, desc_1)$ et $x_2 = (var_2, desc_2)$:

$$d(x_1, x_2) = \frac{1}{Z} Levenshtein(var_1, var_2) + 1 - \frac{<bow(desc_1), bow(desc_2)>}{||bow(desc_1)||_2 . ||bow(desc_2)||_2}$$

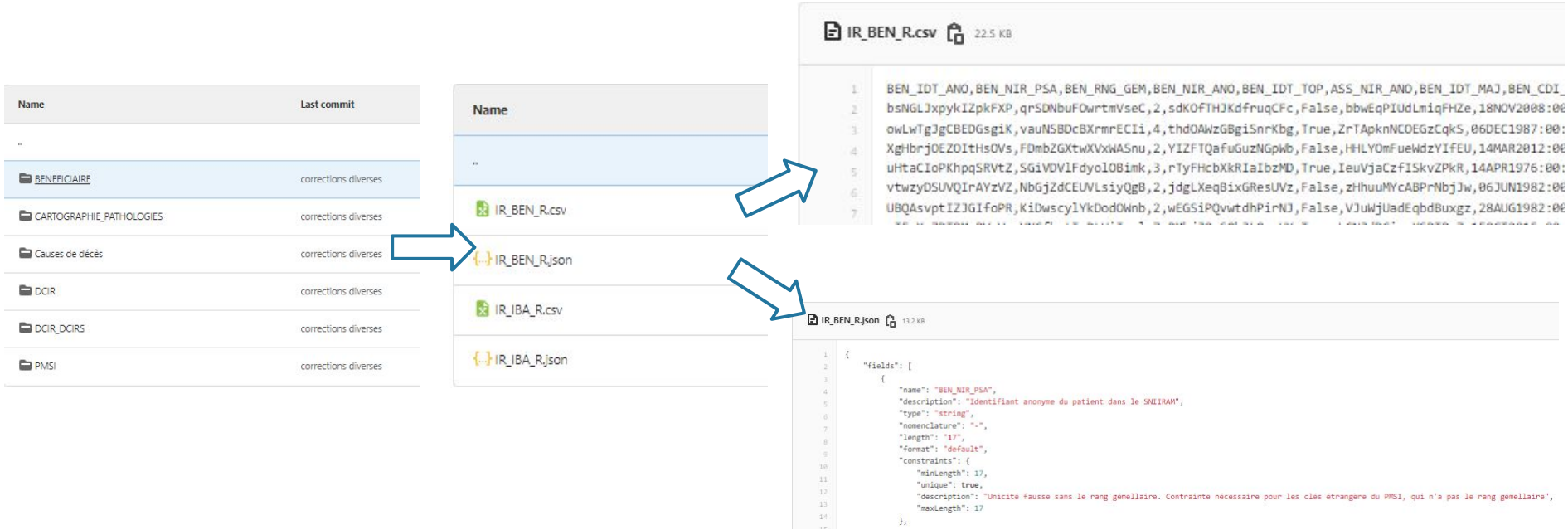
3. Comment introduire du réalisme? (2/2)

Deuxième levier: statistiques descriptives
Plutôt que de générer nos colonnes aléatoirement, on peut utiliser des données en open data pour se rapprocher des distributions réelles, par exemple des médicaments.

⚠️ En revanche, introduire des corrélations est beaucoup plus difficile:
- les données de corrélation ne sont pas toujours en open data
- les corrélations sont souvent au niveau du parcours du patient, et pas au niveau tabulaire



4. Aperçu


Name
Last commit
..
BENEFICIAIRE
corrections diverses
CARTOGRAPHIE_PATHOLOGIES
corrections diverses
Causes de décès
corrections diverses
DCIR
corrections diverses
DCIR_DCIRS
corrections diverses
PMSI
corrections diverses
Name
..
IR_BEN_R.csv
IR_BEN_R.json
IR_IBA_R.csv
IR_IBA_R.json
IR_BEN_R.csv
22.5 KB
BEN_IDT_ANO,BEN_NIR_PSA,BEN_RNG_GEM,BEN_NIR_ANO,BEN_IDT_TOP,ASS_NIR_ANO,BEN_IDT_MAJ,BEN_CDI_
bsNGLJxpykIZpkFXP,qrSDNbuFOwrtmVseC,2,sdKOfTHJKdfruqCFc,False,bbwEqPIUdLmiqFHZe,18NOV2008:00
owLwTgJgCBEDGsgiK,vauNSBDcBXrmrECIi,4,thdOAWzGBgiSnrKbg,True,ZrTApknNCOEGzCqkS,06DEC1987:00:
XgHbrjOEZOItHsOVs,FDmbZGXtwXVxWASnu,2,YIZFTQafuGuzNGpWb,False,HHLYOmFueWdzYIfEU,14MAR2012:00
uHtaCIoPKhpqSRVtZ,SGiVDVlFdyolOBimk,3,rTyFHcbXkRIaIbzMD,True,IeuVjaCzfISkvZPkR,14APR1976:00:
vtwzyDSUVQIrAYzVZ,NbGjZdCEUVLsiyQgB,2,jdgLXeqBixGResUVz,False,zHhuuMYcABPrNbjJw,06JUN1982:00
UBQAsvptIZJGIfoPR,KiDwscylYkDodOWnb,2,wEGSiPQvwtdhPirNJ,False,VJuWjUadEqbdBuxgz,28AUG1982:00
IR_BEN_R.json
13.2 KB
{
"fields": [
{
"name": "BEN_NIR_PSA",
"description": "Identifiant anonyme du patient dans le SNIIRAM",
"type": "string",
"nomenclature": "-",
"length": "17",
"format": "default",
"constraints": {
"minLength": 17,
"unique": true,
"description": "Unicité fausse sans le rang gémellaire. Contrainte nécessaire pour les clés étrangère du PMSI, qui n'a pas le rang gémellaire",
"maxLength": 17
},

4. Aperçu (Remarques)

Repo Gitlab: https://gitlab.com/healthdatahub/synthetic-generator

Limites:

- Pas de cohérence médicale
- Temps de traitement assez conséquent sur de grands volumes de bénéficiaires

Prochaines étapes:

- Prise en compte de plus de bases
- API
- Amélioration des performances (notamment en termes de vitesse)

MERCI POUR VOTRE ATTENTION