UMR 8199

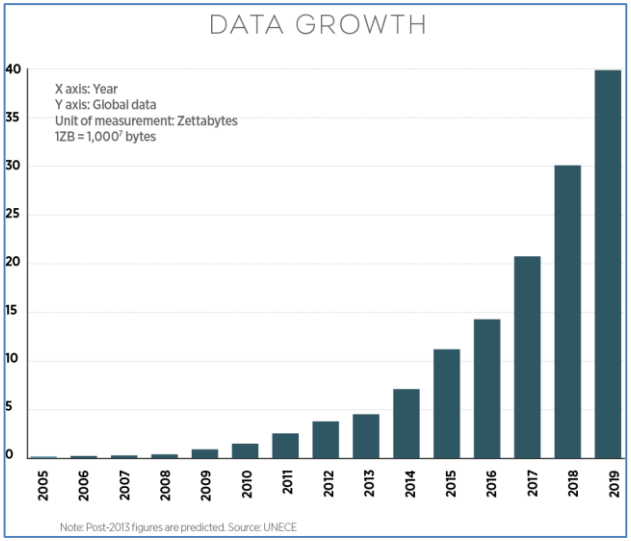
GÉNOMIQUE INTÉGRATIVE ET MODÉLISATION DES MALADIES MÉTABOLIQUES

Data cleaning avec OpenRefine : nettoyer et enrichir ses données

Stefan Gaget, stefan.gaget@cnrs.fr

Etat des lieux (1)

- Des données dans tous les sens
  - Explosion des données générées et collectées



  - Standards et formats multiples
    - structurés ou non
    - écritures différentes

Etat des lieux (2)

- Données sur supports divers :
  - Fichiers plats (.csv, .txt, .xls, .log)
  - Papiers

- Données nécessitant un traitement
  - Erreurs de saisie
  - Uniformisation
  - Contrôle de qualité
  - Données manquantes

- Données brutes difficilement exploitables :
  - Liaisons entre fichiers

Etat des lieux (3)

- (La base de données)
  - Stocker des informations
  - Facilité d'utilisation (extraction selon différents critères)
  - Vitesse de traitement
  - Typage des données
  - Relations entre données différentes
  - …
- (Quel SGBD choisir ?)
- (Comment modéliser votre base ?)
- Comment y intégrer vos données ?

Objectifs :

- Diagnostiquer et nettoyer ses données
- Améliorer les flux d'alimentation de sa base de données
- Produire des règles de gestion d'intégration des données
- Prendre en main OpenRefine

Programme :

- Les données :
  - codification
  - contrôle de qualité
  - erreurs courantes dans la manipulation
- Alimenter une base de données :
  - vue d'ensemble
  - préparation des données
- Formater des données avec OpenRefine :
  - introduction
  - installation
  - processus général par l'étude d'un cas pratique
  - fonctionnalités
- Un projet de A à Z par la pratique

Programme :

- Les données :
  - codification
  - contrôle de qualité
  - erreurs courantes dans la manipulation
- Alimenter une base de données :
  - vue d'ensemble
  - préparation des données
- Formater des données avec OpenRefine :
  - introduction
  - installation
  - processus général par l'étude d'un cas pratique
  - fonctionnalités
- Un projet de A à Z par la pratique

Codification :

- Données sont extraites pour être traitées
- Souvent par des méthodes statistiques

- Traitement automatique => codification
  - Pays = code international, Région ?
  - Localisation = données GPS
  - Personne = Initiales, autre code
  - Espèce = n° taxon, nom binomial (genre espèce), autre code
  - Publication = DOI, PMID

- Codification des données manquantes

Méta-données :

- Exemples :
  - Date d'acquisition de la donnée
  - Auteur de la donnée
  - Unité de mesure / donnée numérique (mètres ou miles ?)
  - Méthode ou protocole utilisé pour l'acquisition de la donnée
  - Niveau de confiance ou de qualité, statut de l'information
    - certain/incertain, direct/indirect, 1 mesure/moyenne,
    - local/importé d'une base de données publique
    - Ex : observation d'un animal : Incertain sur l'espèce ou sur le stade de développement, quantité exacte ou estimée, estimation indirecte (trace, cris).
- Utilité :
  - Gérer des données de provenance et de qualité hétérogènes
    - Plusieurs équipes, plusieurs protocoles
- Filtrer ou convertir les données avant traitement

Contrôle de qualité des données 1/4

- Gestion des données manquantes
  - Évaluer leur fréquence, leur provenance
    - Ex : concernent un lieu, un type de lieu, une période de temps ...
  - Filtrer les données manquantes
    - Ex : calculer le nombre de données «complètes»
  - Calculer des estimations pour les données manquantes
    - Ex si localisation GPS manquante, prendre la localisation de la commune et mettre « estimateur=commune » comme niveau de confiance

Contrôle de qualité des données 2/4

- Contrôle des valeurs
  - Données qualitatives
    - Listes de valeurs possibles (cf codification)
  - Données quantitatives
    - Plages de valeurs

- Données corrélées
  - Vérifications de certaines relations
    - Ex : heure début < heure fin

Contrôle de qualité des données 3/4

- **Contrainte d'intégrité** = expression logique qui doit être vérifiée sur le jeu de données (ie qui doit toujours être vrai)

- Ex1 : toute observation doit avoir une date (ie pas de date manquante)
  - *Contrainte de non nullité*
- Ex2 : chaque observateur a un identifiant unique
  - *Contrainte d'unicité*
- Ex3 : il ne peut y avoir 2 cours dans la même salle à la même heure
  - *Contrainte d'unicité : clé = numéro de salle + heure*
- Ex4 : un numéro de salle est un entier
  - *Contraintes de domaine*
- Ex5 : un numéro de salle est compris entre 0 et 23
  - *Contraintes de vérification*

Contrôle de qualité des données 4/4

- **Contrainte d'intégrité** = expression logique qui doit être vérifiée sur le jeu de données (ie qui doit toujours être vrai)

- Ex6 : une zone est de type PNR, PNT ou ENT
  - *Contrainte d'intégrité de référence*
- Ex7 : heure de fin > heure de début
  - *Contrainte de relation* portant sur plusieurs attributs du jeu de données

- Spécifier les contraintes
  - Permet la vérification automatique par le SGBD
  - Fait partie du cahier des charges et de la modélisation

Exemple de contraintes d'intégrité

	A	B	C	D	E	F	G	H
1	ID	date_naissance	sexe	age_exam	taille	poids	IMC	PP_grossesse
2	20715	09/04/1931	Feminin	67	157	56	22,72	Non
3	20633	28/09/1954	Masculin	44	163	86	32,37	Oui
4	20713	24/01/1955	Masculin	43	176	140	45,20	
5	20673	19/05/1956	Feminin	42	165	133	48,85	Oui
6	20630	01/08/1948	Feminin	50	257	78	11,81	Oui
7	20618	13/06/1948	Masculin	50	124	278	180,80	
8	20591	22/12/1953	Feminin	44	164	102	37,92	Oui
9	20590	21/04/1953	Masculin	45	177	98	31,28	
10	20752	10/08/1922	Masculin	76	168	84	29,76	
11	20753	01/09/1923	Feminin	74	162	66	25,15	Oui
12	20719	04/10/1925	Masculin	73	168	80	28,34	
13	20720	08/02/1927	Feminin	71	158	72	28,84	Oui
14	20629	12/07/1980	Feminin	18	169	91	31,86	Non
15	20085	31/01/1958	Masculin	39	178	141	44,5	
16	20371	18/09/1959	Masculin	39	175	80	26,12	

Programme :

- Les données :
  - codification
  - contrôle de qualité
  - erreurs courantes dans la manipulation
- Alimenter une base de données :
  - vue d'ensemble
  - préparation des données
- Formater des données avec OpenRefine :
  - introduction
  - installation
  - processus général par l'étude d'un cas pratique
  - fonctionnalités
- Un projet de A à Z par la pratique

Objectif : Présenter des solutions différentes pour alimenter une base de données; en comprendre leurs possibilités et leurs limites

➔Des solutions permettant de répondre aux besoins simples

➔Quelle solution pour ne pas perdre (trop) de temps et traiter ces données ?

Les solutions

- « A la main »
- En ligne de commande
- Des scripts
- Avec un outil d'administration de base de données

A garder en tête

- Au cas par cas
  - Ta bonne méthode, tu adopteras !
- Inconvénients
  - Perte potentielle de traçabilité entre le support d'origine et les informations saisies dans un fichier
  - Toutes les informations connexes ne sont pas nécessairement dans les fichiers
- Conseils
  - Etablir une documentation détaillée sur les transformations effectuées
  - Gardez précieusement les supports d'origine

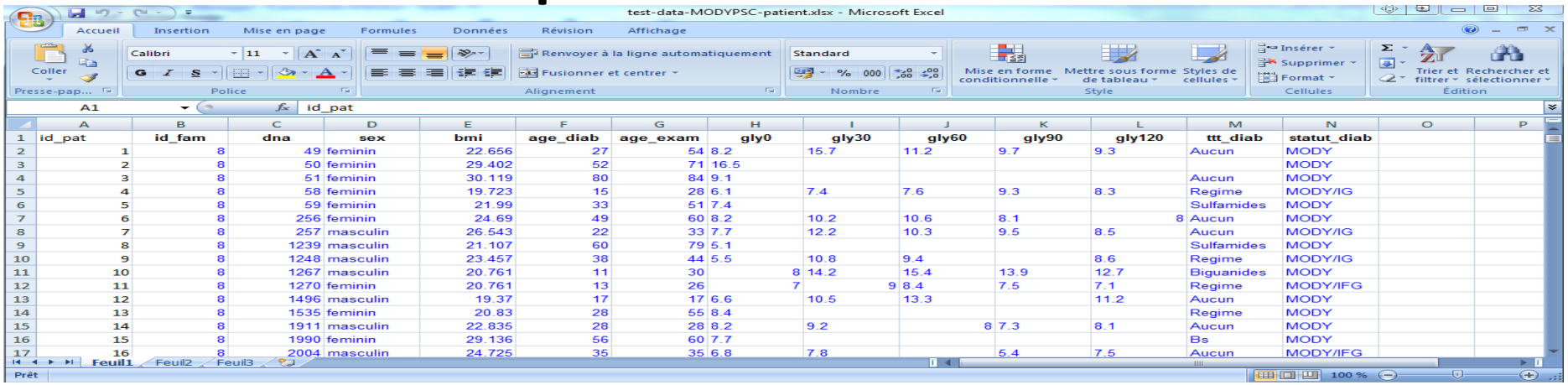
Les solutions

- **« A la main »**
- En ligne de commande
- Des scripts
- Avec un outil d'administration de base de données

« A la main »

- Importation brutale des données
  - 1 feuille = 1 table
- One shoot
- Un petit côté bricolage

Exemple fichier tableur



test-data-MODYPSC-patient.xlsx - Microsoft Excel

id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	gly30	gly60	gly90	gly120	ttt_diab	statut_diab
1	8	49	feminin	22.656	27	54	8.2	15.7	11.2	9.7	9.3	Aucun	MODY
2	8	50	feminin	29.402	52	71	16.5						MODY
3	8	51	feminin	30.119	80	84	9.1					Aucun	MODY
4	8	58	feminin	19.723	15	28	6.1	7.4	7.6	9.3	8.3	Regime	MODY/IG
5	8	59	feminin	21.99	33	51	7.4					Sulfamides	MODY
6	8	256	feminin	24.69	49	60	8.2	10.2	10.6	8.1	8	Aucun	MODY
7	8	257	masculin	26.543	22	33	7.7	12.2	10.3	9.5	8.5	Aucun	MODY/IG
8	8	1239	masculin	21.107	60	79	5.1					Sulfamides	MODY
9	8	1248	masculin	23.457	38	44	5.5	10.8	9.4		8.6	Regime	MODY/IG
10	8	1267	masculin	20.761	11	30	8	14.2	15.4	13.9	12.7	Biguanides	MODY
11	8	1270	feminin	20.761	13	26	7	9	8.4	7.5	7.1	Regime	MODY/IFG
12	8	1496	masculin	19.37	17	17	6.6	10.5	13.3		11.2	Aucun	MODY
13	8	1535	feminin	20.83	28	55	8.4					Regime	MODY
14	8	1911	masculin	22.835	28	28	8.2	9.2	8	7.3	8.1	Aucun	MODY
15	8	1990	feminin	29.136	56	60	7.7					Bs	MODY
16	8	2004	masculin	24.725	35	35	6.8	7.8		5.4	7.5	Aucun	MODY/IFG

- Préalable
  - Regrouper l'information par thème
  - Recouper l'information pour trouver des erreurs
- Conseil
  - Trouver un format facilement utilisable par la personne en charge de la saisie

Rappel SQL

- **Création de la table : CREATE TABLE**

  CREATE TABLE nom_table (
  colonne1 type1(taille1),
  colonne2 type2(taille2),
  …);

  http://www.w3schools.com/sql/sql_create_table.asp

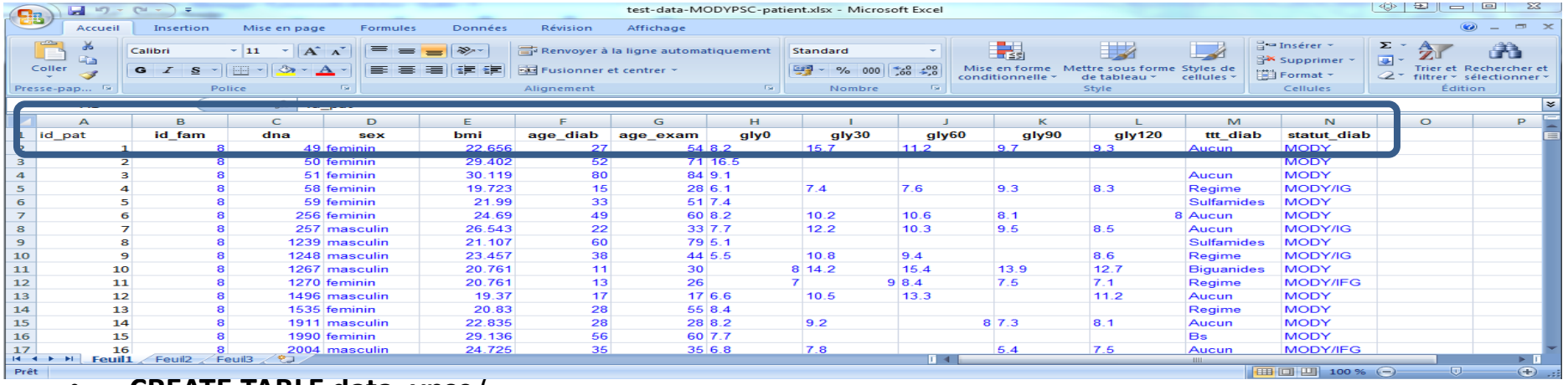
- **Insertion d'une ligne de donnée : INSERT INTO**

  INSERT INTO nom_table (colonne1, colonne2, …)
  VALUES (valeur1, valeur2 , …);

  INSERT INTO nom_table **VALUES** (valeur1, valeur2, …, valeurN);
  (respecter le nombre et l'ordre des colonnes)

  http://www.w3schools.com/sql/sql_insert.asp

Création de la table



id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	gly30	gly60	gly90	gly120	ttt_diab	statut_diab
1	8	49	feminin	22.656	27	54	8.2	15.7	11.2	9.7	9.3	Aucun	MODY
2	8	50	feminin	29.402	52	71	16.5						MODY
3	8	51	feminin	30.119	80	84	9.1					Aucun	MODY
4	8	58	feminin	19.723	15	28	6.1	7.4	7.6	9.3	8.3	Regime	MODY/IG
5	8	59	feminin	21.99	33	51	7.4					Sulfamides	MODY
6	8	256	feminin	24.69	49	60	8.2	10.2	10.6	8.1	8	Aucun	MODY
7	8	257	masculin	26.543	22	33	7.7	12.2	10.3	9.5	8.5	Aucun	MODY/IG
8	8	1239	masculin	21.107	60	79	5.1					Sulfamides	MODY
9	8	1248	masculin	23.457	38	44	5.5	10.8	9.4		8.6	Regime	MODY/IG
10	8	1267	masculin	20.761	11	30	8	14.2	15.4	13.9	12.7	Biguanides	MODY
11	8	1270	feminin	20.761	13	26	7	9	8.4	7.5	7.1	Regime	MODY/IFG
12	8	1496	masculin	19.37	17	17	6.6	10.5	13.3		11.2	Aucun	MODY
13	8	1535	feminin	20.83	28	55	8.4					Regime	MODY
14	8	1911	masculin	22.835	28	28	8.2	9.2	8	7.3	8.1	Aucun	MODY
15	8	1990	feminin	29.136	56	60	7.7					Bs	MODY
16	8	2004	masculin	24.725	35	35	6.8	7.8		5.4	7.5	Aucun	MODY/IFG

- **CREATE TABLE data_ypsc** (
  id_pat INT,
  id_fam INT,
  dna INT,
  sex CHAR,
  bmi FLOAT,
  age_diab INT,
  age_exam INT,
  gly0 FLOAT
  ttt_diab VARCHAR,
  statut_diab CHAR
  );

!!! Les types varient selon les SGDB !!!

MySQL	PostgreSQL
FLOAT	REAL
SMALLINT	TINYINT
…	…

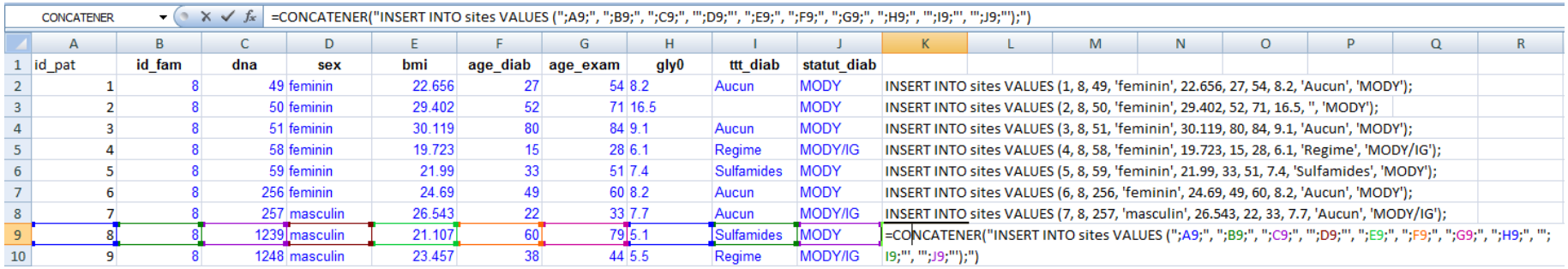
!!! La syntaxe légèrement aussi !!!

Générer le code SQL d'insertion

- Utiliser les fonctions du tableur

- Avec la formule Excel :

```
=CONCATENER("INSERT INTO data_ypsc
        VALUES (";A9;", ";B9;", ";C9;", '";D9;"', ";E9;", ";F9;", ";G9;", ";H9;", '";I9;"',
                '";J9;"');")
```

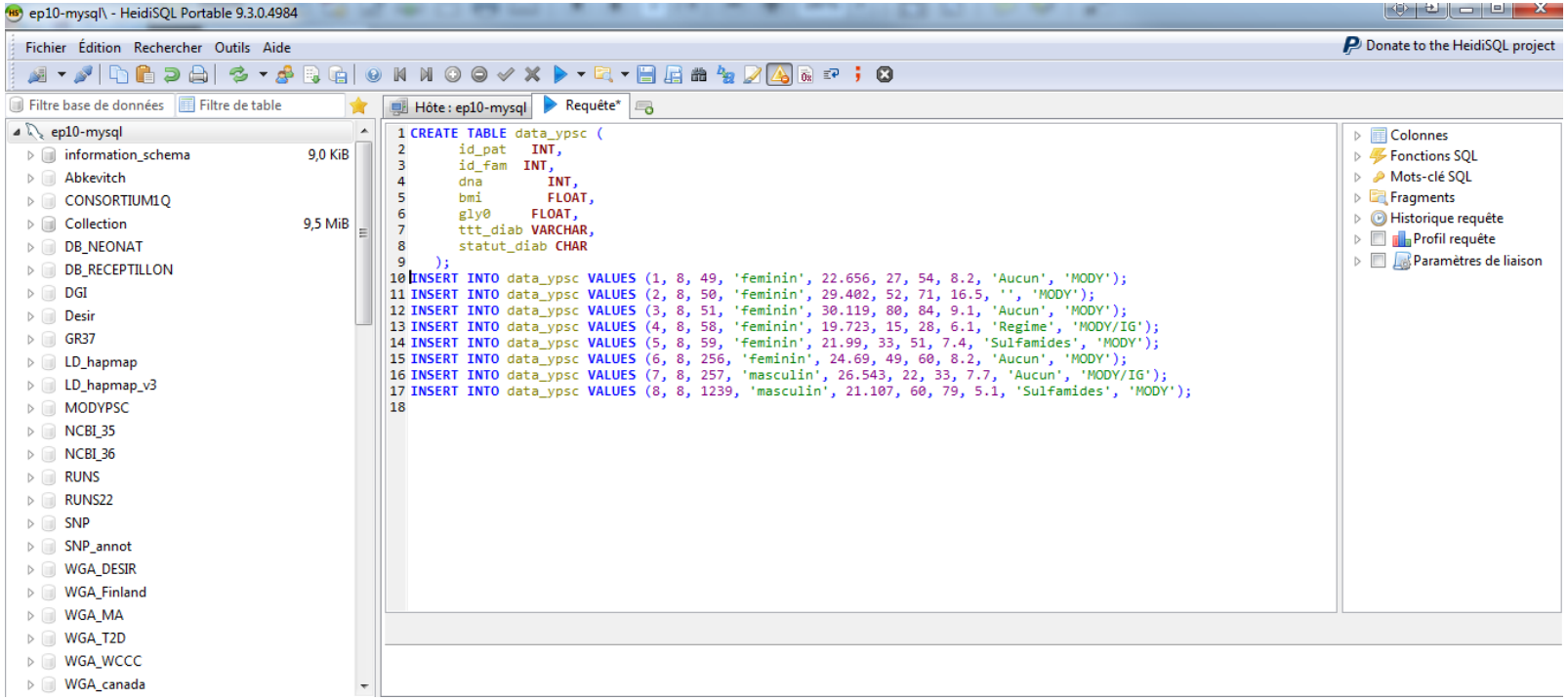


	A	B	C	D	E	F	G	H	I	J	K
1	id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab	
2	1	8	49	feminin	22.656	27	54	8.2	Aucun	MODY	INSERT INTO sites VALUES (1, 8, 49, 'feminin', 22.656, 27, 54, 8.2, 'Aucun', 'MODY');
3	2	8	50	feminin	29.402	52	71	16.5		MODY	INSERT INTO sites VALUES (2, 8, 50, 'feminin', 29.402, 52, 71, 16.5, '', 'MODY');
4	3	8	51	feminin	30.119	80	84	9.1	Aucun	MODY	INSERT INTO sites VALUES (3, 8, 51, 'feminin', 30.119, 80, 84, 9.1, 'Aucun', 'MODY');
5	4	8	58	feminin	19.723	15	28	6.1	Regime	MODY/IG	INSERT INTO sites VALUES (4, 8, 58, 'feminin', 19.723, 15, 28, 6.1, 'Regime', 'MODY/IG');
6	5	8	59	feminin	21.99	33	51	7.4	Sulfamides	MODY	INSERT INTO sites VALUES (5, 8, 59, 'feminin', 21.99, 33, 51, 7.4, 'Sulfamides', 'MODY');
7	6	8	256	feminin	24.69	49	60	8.2	Aucun	MODY	INSERT INTO sites VALUES (6, 8, 256, 'feminin', 24.69, 49, 60, 8.2, 'Aucun', 'MODY');
8	7	8	257	masculin	26.543	22	33	7.7	Aucun	MODY/IG	INSERT INTO sites VALUES (7, 8, 257, 'masculin', 26.543, 22, 33, 7.7, 'Aucun', 'MODY/IG');
9	8	8	1239	masculin	21.107	60	79	5.1	Sulfamides	MODY	=CONCATENER("INSERT INTO sites VALUES (";A9;", ";B9;", ";C9;", '";D9;"', ";E9;", ";F9;", ";G9;", ";H9;", '";
10	9	8	1248	masculin	23.457	38	44	5.5	Regime	MODY/IG	I9;"', '";J9;"');")

```
INSERT INTO data_ypsc VALUES (1, 8, 49, 'feminin', 22.656, 27, 54, 8.2, 'Aucun', 'MODY');
INSERT INTO data_ypsc VALUES (2, 8, 50, 'feminin', 29.402, 52, 71, 16.5, '', 'MODY');
INSERT INTO data_ypsc VALUES (3, 8, 51, 'feminin', 30.119, 80, 84, 9.1, 'Aucun', 'MODY');
INSERT INTO data_ypsc VALUES (4, 8, 58, 'feminin', 19.723, 15, 28, 6.1, 'Regime', 'MODY/IG');
INSERT INTO data_ypsc VALUES (5, 8, 59, 'feminin', 21.99, 33, 51, 7.4, 'Sulfamides', 'MODY');
INSERT INTO data_ypsc VALUES (6, 8, 256, 'feminin', 24.69, 49, 60, 8.2, 'Aucun', 'MODY');
INSERT INTO data_ypsc VALUES (7, 8, 257, 'masculin', 26.543, 22, 33, 7.7, 'Aucun', 'MODY/IG');
INSERT INTO data_ypsc VALUES (8, 8, 1239, 'masculin', 21.107, 60, 79, 5.1, 'Sulfamides', 'MODY');
```

Exécution du code SQL

- En ligne de commande
- Avec un outil d'administration de base de données (PhpMyAdmin, pgAdmin, HeidiSql…)



Les solutions

- « A la main »
- **En ligne de commande**
- Des scripts
- Avec un outil d'administration de base de données

En ligne de commande

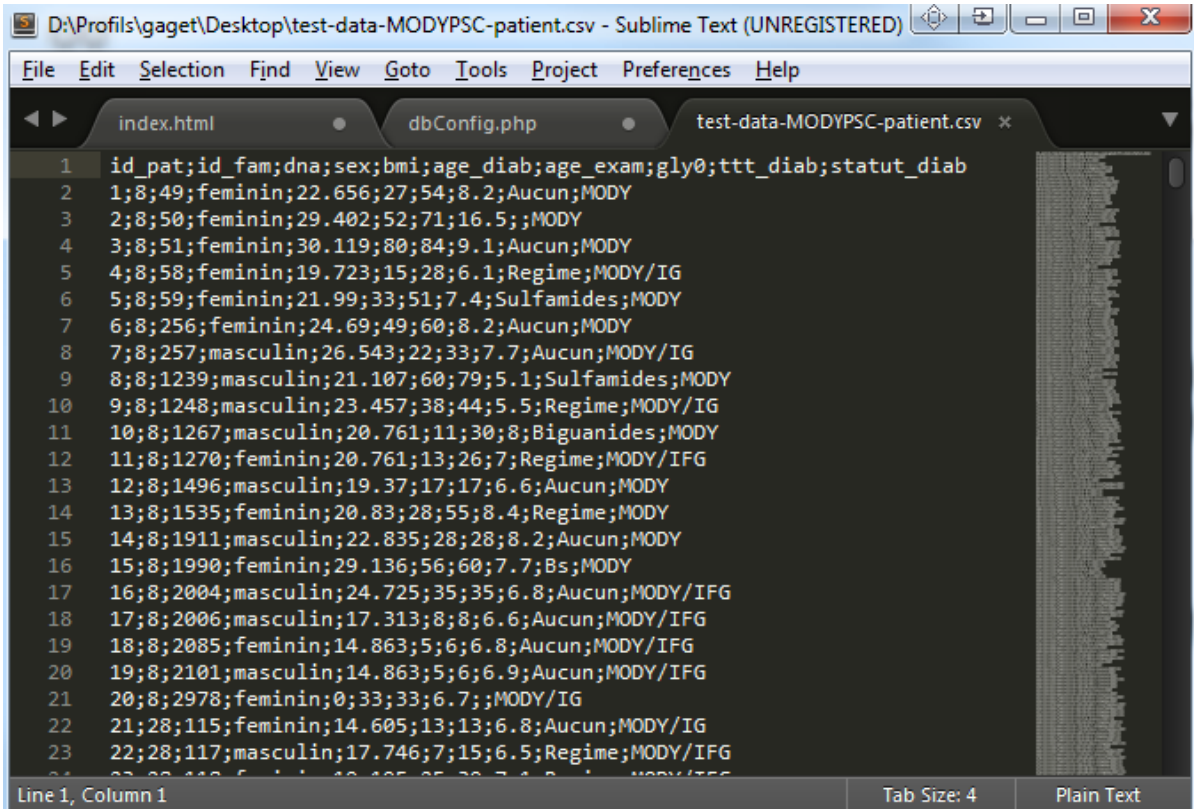
- **(CREATE TABLE** … pour créer la table vide)

- **COPY** pour charger les valeurs à partir du fichier .csv ou .txt (; ou TAB)

  – COPY data_ypsc(
    colonne1,
    colonne2, …)
    FROM nom_fichier
    WITH options;

- **COPY n'est pas un mot-clé standard de SQL …**

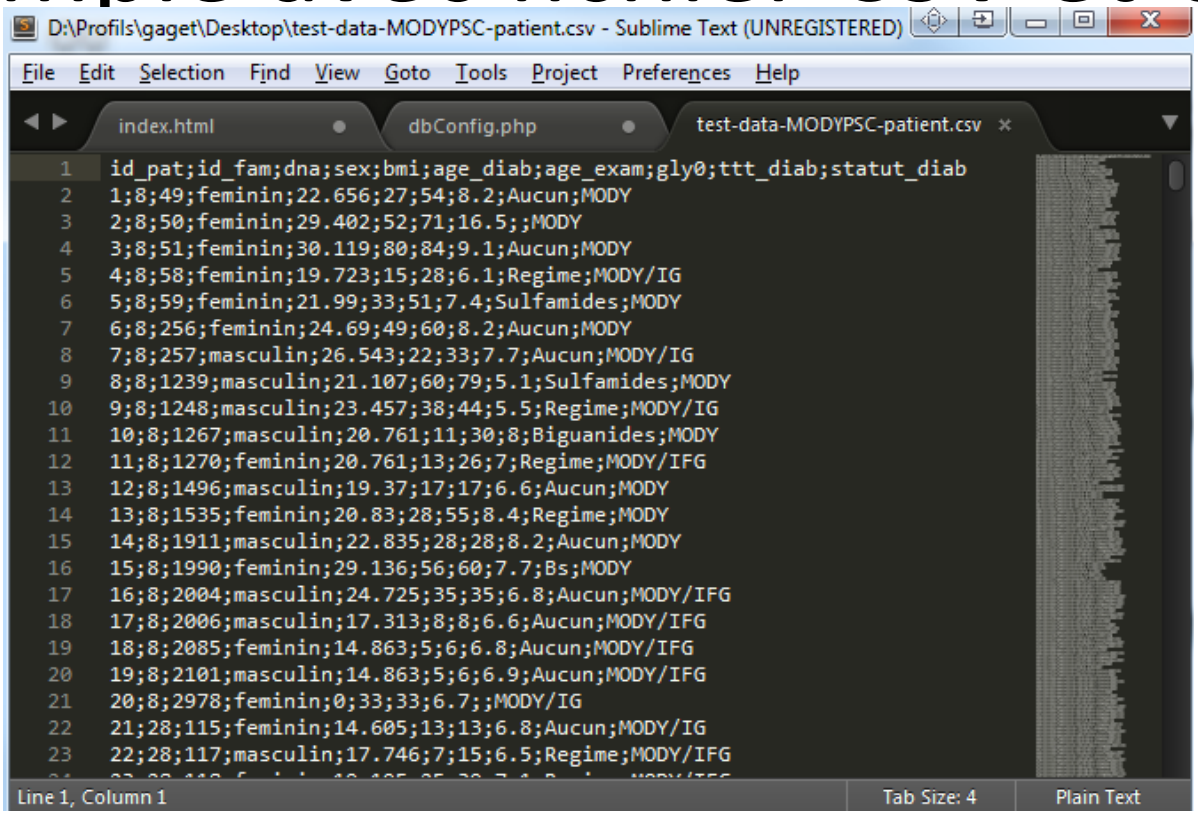
Exemple fichier texte avec délimiteurs



Les options de **COPY**

- **CSV HEADER** permet de sauter la 1ère ligne (contenant les noms de colonne)
- **DELIMITER ';' ou DELIMITER E'\t'** pour définir le délimiteur (; ou tab)
- **NULL 'NA'** pour définir une chaîne équivalente à NULL (par défaut : '')
- **ENCODING 'UTF8' ou ENCODING 'WIN1252';** pour définir l'encodage (UTF8 ou ANSI)
  - **Notepad ++ permet de vérifier l'encodage**
  - **Dans Excel : « Texte (séparateur:tabulation)(*.txt) », « CSV (séparateur : pointvirgule)(*.**csv) » donnent des fichiers ANSI.
  - **Dans Excel : « Texte Unicode (*.txt) » donne des fichiers UTF8**

Exemple avec fichier CSV et UTF8



- **COPY data_ypsc (id_pat, id_fam, dna, bmi, gly0, ttt_diab, statut_diab) FROM ‘D:\test-data-MODYPSC-patient.csv‘ WITH CSV HEADER DELIMITER ‘;' ENCODING 'UTF8';**

- **COPY data_ypsc FROM ‘D:\test-data-MODYPSC-patient.csv‘ WITH CSV HEADER DELIMITER ‘;' ENCODING 'UTF8';**

Les solutions

- « A la main »
- En ligne de commande
- **Des scripts**
- Avec un outil d'administration de base de données

Avec des scripts (R)

- Utilisation du **package RPostgreSQL** pour importer des données
- RPostgreSQL permet de se connecter à une base PostgreSQL pour consulter / mettre à jour les données
  - Documentation : https://code.google.com/p/rpostgresql/
- Les principales fonctions sont héritées du package DBI :
  - **dbSendQuery** exécute une requête (résultat accessible ligne après ligne via un ResultSet)
  - **dbGetQuery** envoie une requête et récupère le résultat dans un data.frame
  - **dbReadTable** charge une table de la BD dans un data.frame
  - **dbWriteTable** crée une table dans la BD à partir d'un data.frame

La connexion à une base de données

- Définir le driver puis la connexion
  - Paramètres : host, port, dbname, user, password

```
Library(RPostgreSQL)
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, host="193.49.134.163",
        port="5433", dbname="bd_formation",
        user="user_fp0", password="user_fp0")
```

- Terminer proprement un programme : libérer la connexion et le driver

```
dbDisconnect(con)
dbUnloadDriver(drv)
```

RPostgreSQL : importer une table

- **dbWriteTable**

```
# exemple RPostgreSQL

library(RPostgreSQL)
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, host="193.49.134.163", port="5433", dbname="bd_formation",
    user="user_fp0", password="user_fp0")

# lire fichier txt (enc=ANSI)
setwd("D:\\Formations\\RequeteSQL2")
df=read.table("ClasseurtotalWet1964-2013VF_sites_ANSI.txt", header=TRUE, sep="\t",
    quote="")

# test requête
dbSendQuery(con,"SET client_encoding = WIN1252")
dbSendQuery(con,"SET search_path = entree_donnees_0")

# créer table
dbWriteTable(con,"sites",df)

# terminer connection, libérer mémoire
dbDisconnect(con)
dbUnloadDriver(drv)
```

Les solutions

- « A la main »
- En ligne de commande
- Des scripts
- **Avec un outil d'administration de base de données**

Avec un outil d'administration de base de données

- **Interface client**
- **Multi-plateformes, multi-installations**
- **Génération de code automatique**
- **Proscrire l’écriture d’instructions SQL (« click and go »)**

Les fonctionnalités

- **Création, éditions de bases, tables, …**
- **Edition, exécution d'instructions SQL**
- **Constructeur de requête**
- **Importer/Exporter données et structures**
- **PhpMyAdmin, pgAdmin, MySQL Workbench, SqlWorkbench, DBeaver, HeidiSql, Adminer…**

- http://www.heidisql.com/
- projet libre, licence GPL

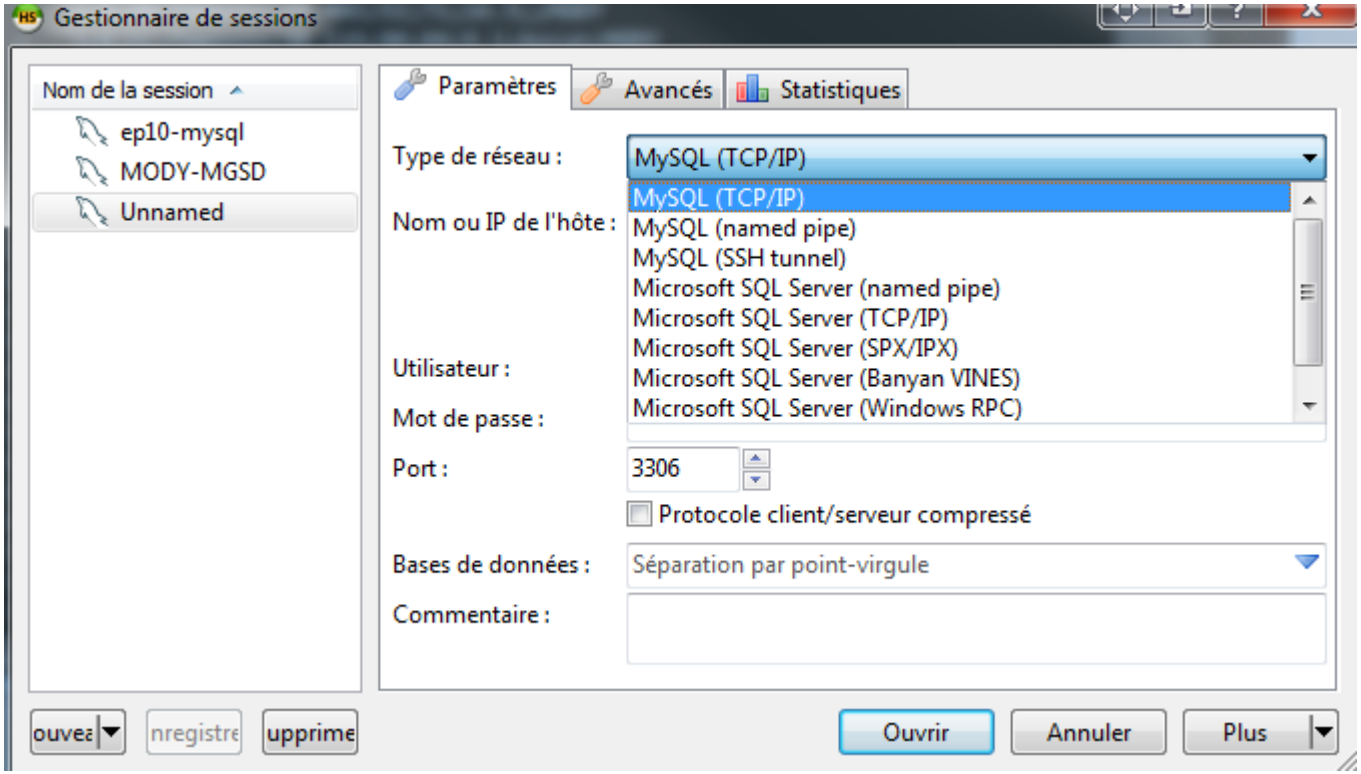
- Ansgar Becker, version 9.4 (octobre 2016)

- installation locale sur poste de travail

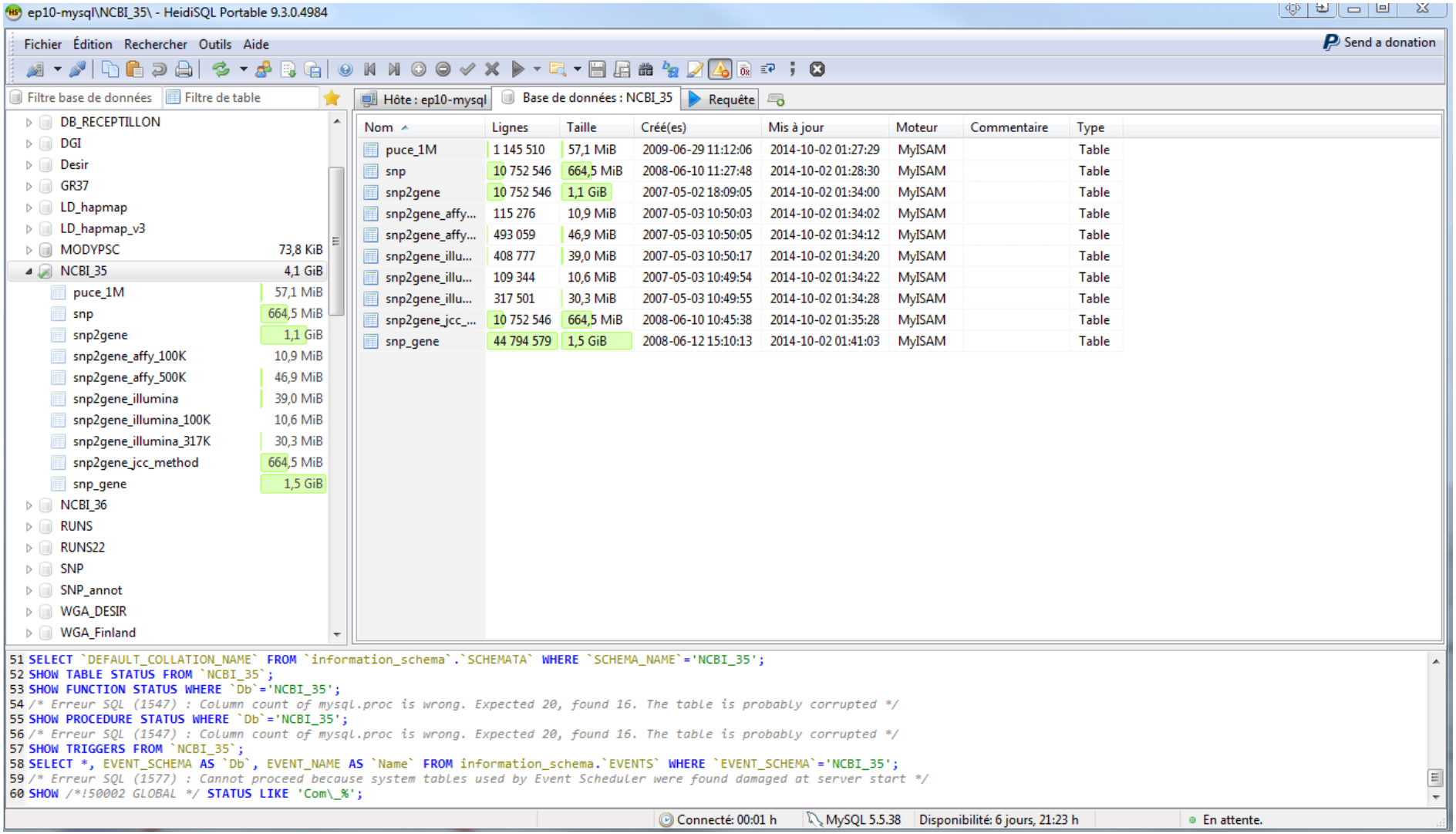
- Windows (installeur ou portable)
- (Linux et Mac OS X avec Wine)

- développé et optimisé pour MySQL
- connexion avec Mysql, Microsoft SQL et PostgreSQL

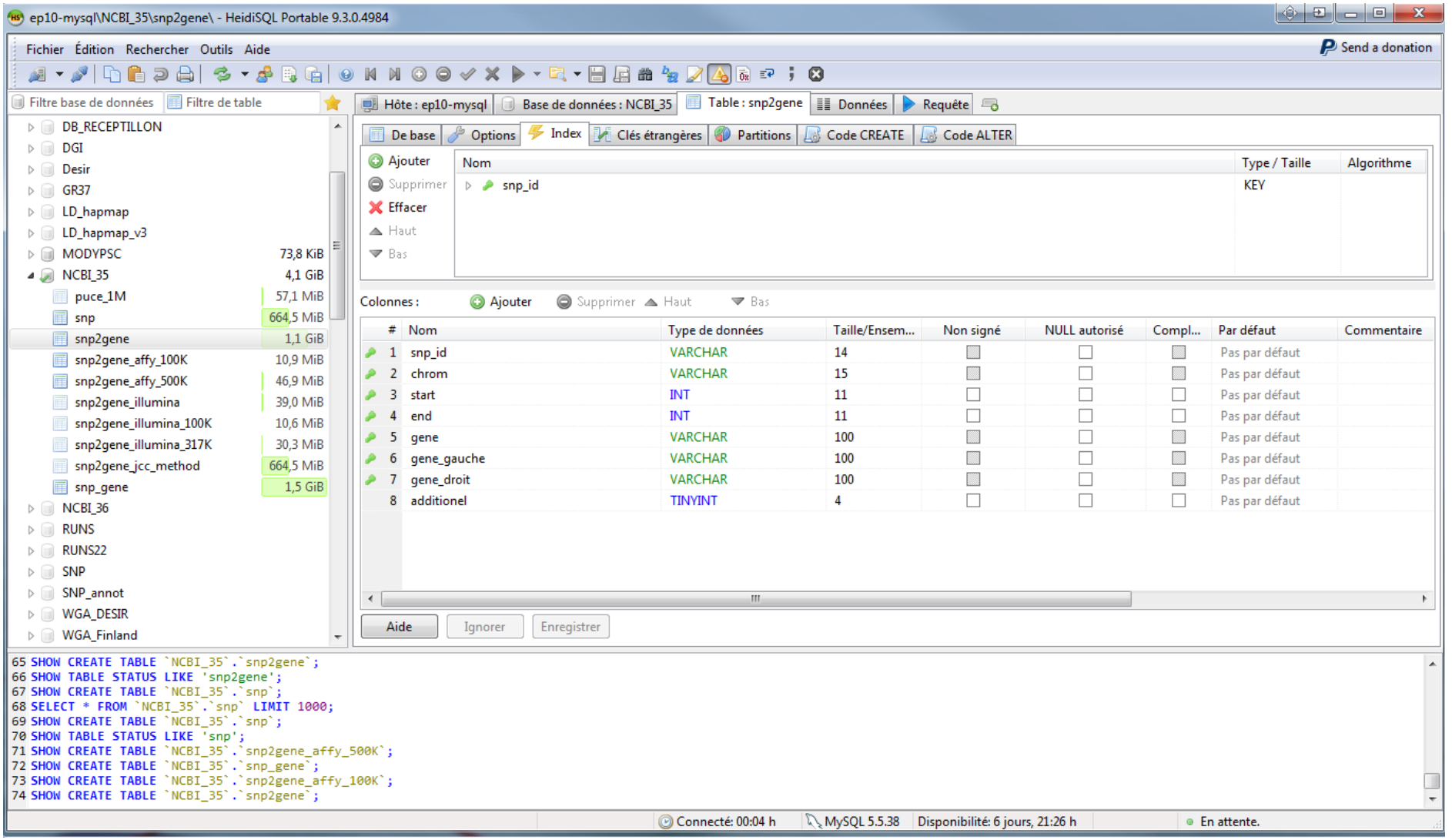
Connexion au serveur



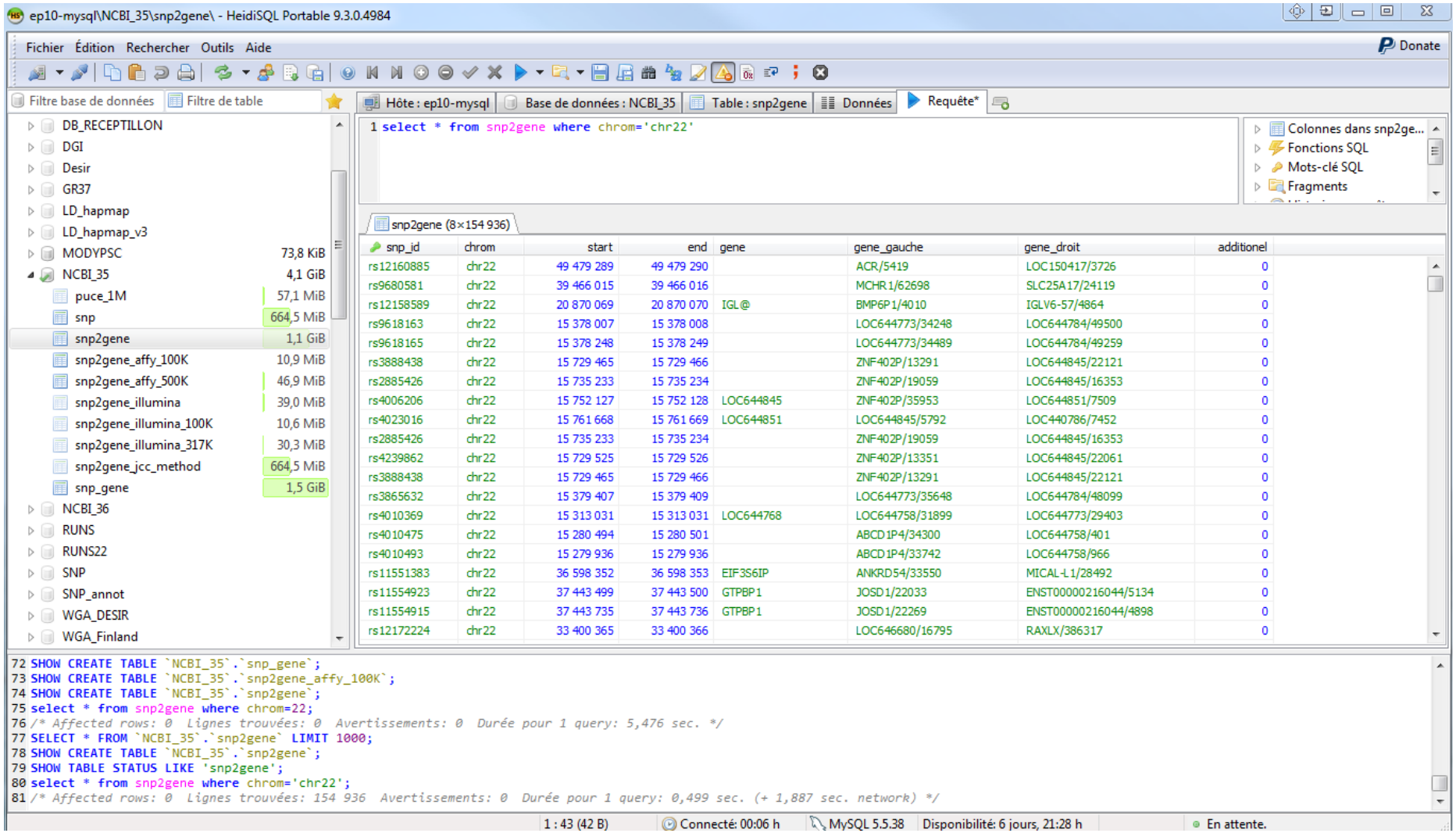
Vue générale (1)



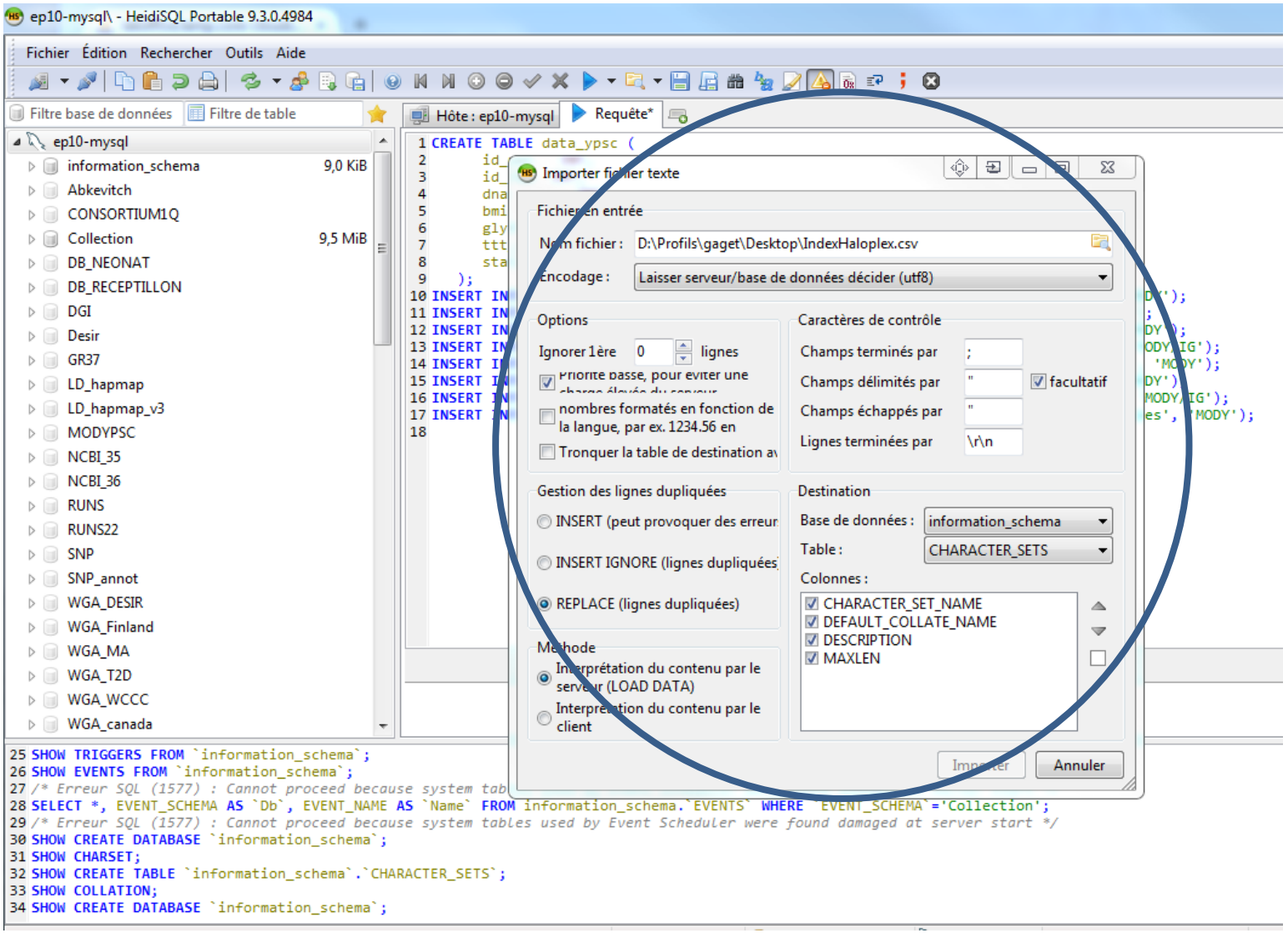
Vue générale (2)



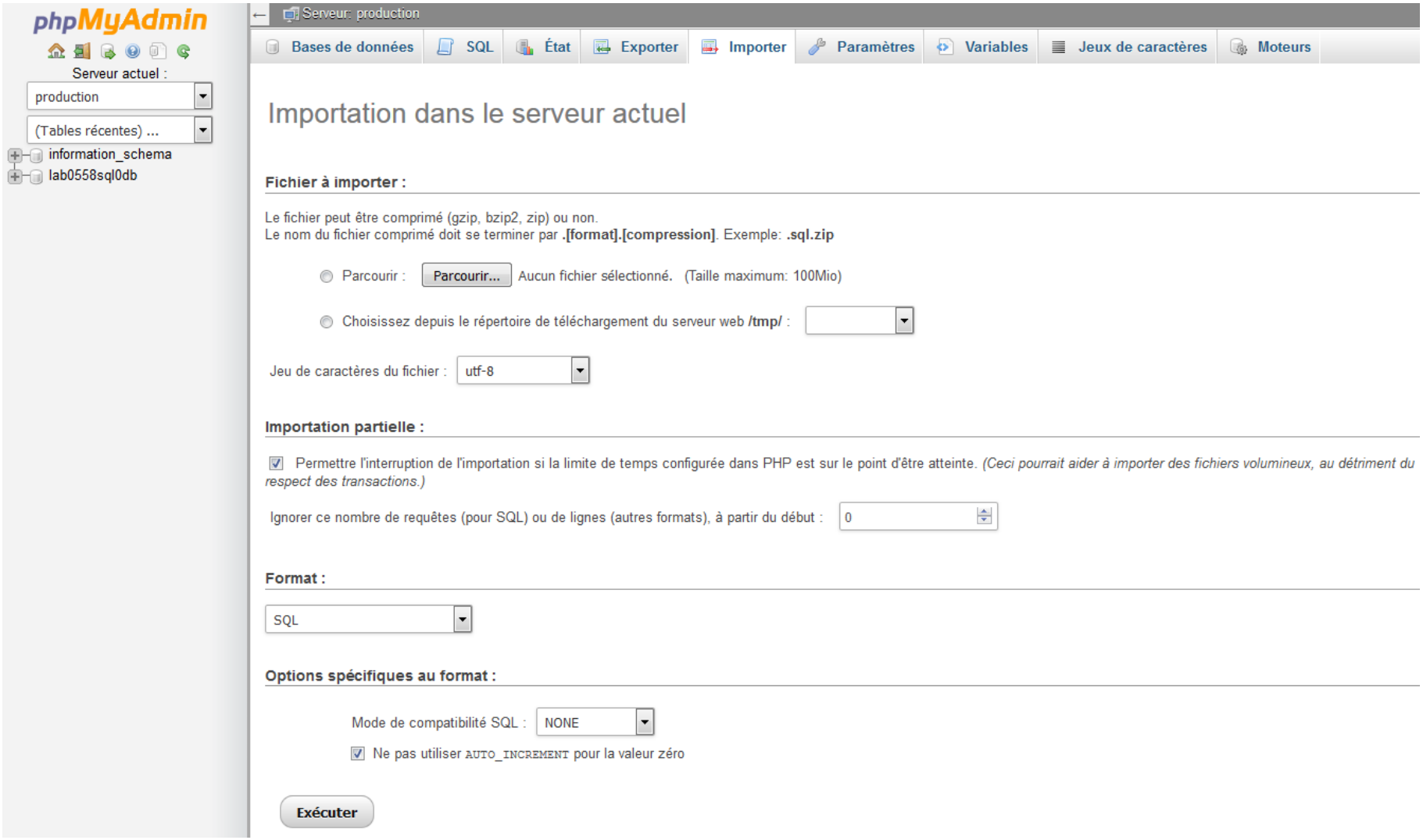
Vue générale (3)



L'outil **Importer** dans HeidiSQL

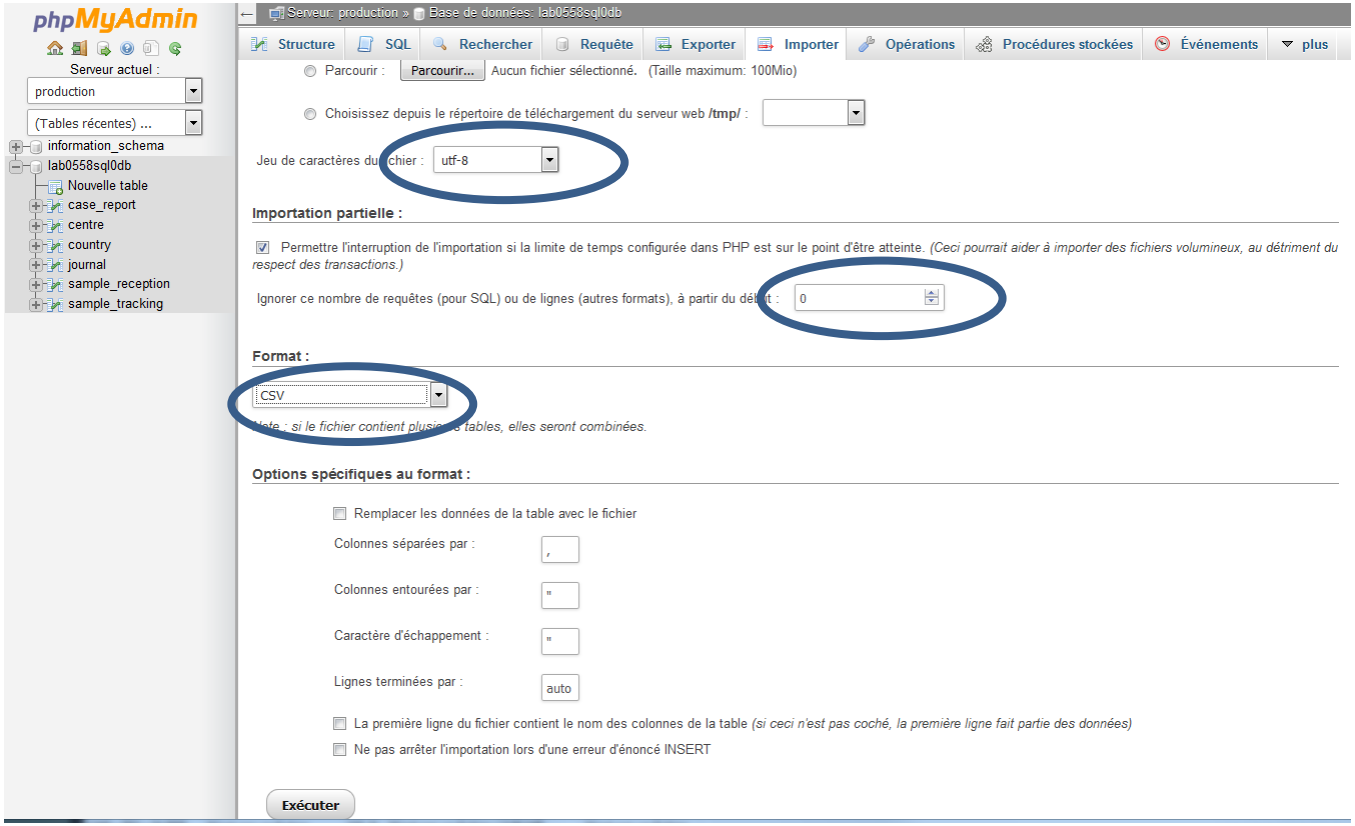


L’outil **Importer** dans phpMyAdmin



L’outil **Importer** dans phpMyAdmin

- Permet d’importer les données à partir d’un fichier
- Options semblables à celles de Copy



Les solutions

- « A la main »
- En ligne de commande
- Des scripts
- Avec un outil d'administration de base de données
- **BONUS : Le couteau suisse pour manipuler les données …**

Programme :

- Les données :
  - codification
  - contrôle de qualité
  - erreurs courantes dans la manipulation
- Alimenter une base de données :
  - vue d'ensemble
  - préparation des données
- Formater des données avec OpenRefine :
  - introduction
  - installation
  - processus général par l'étude d'un cas pratique
  - fonctionnalités
- Un projet de A à Z par la pratique

OpenRefine : un couteau suisse pour manipuler les données

A free, open source, powerful tool for working with messy data

- Ex Google-Refine
- http://www.openrefine.org

« A free, open source, powerful tool for working with messy data »

- Outil de manipulation de données

- Application multi-plateforme fonctionnant localement
- Interface web (nécessite une connexion internet)

- Données restent en local

- … Quelque chose entre un tableur et SQL

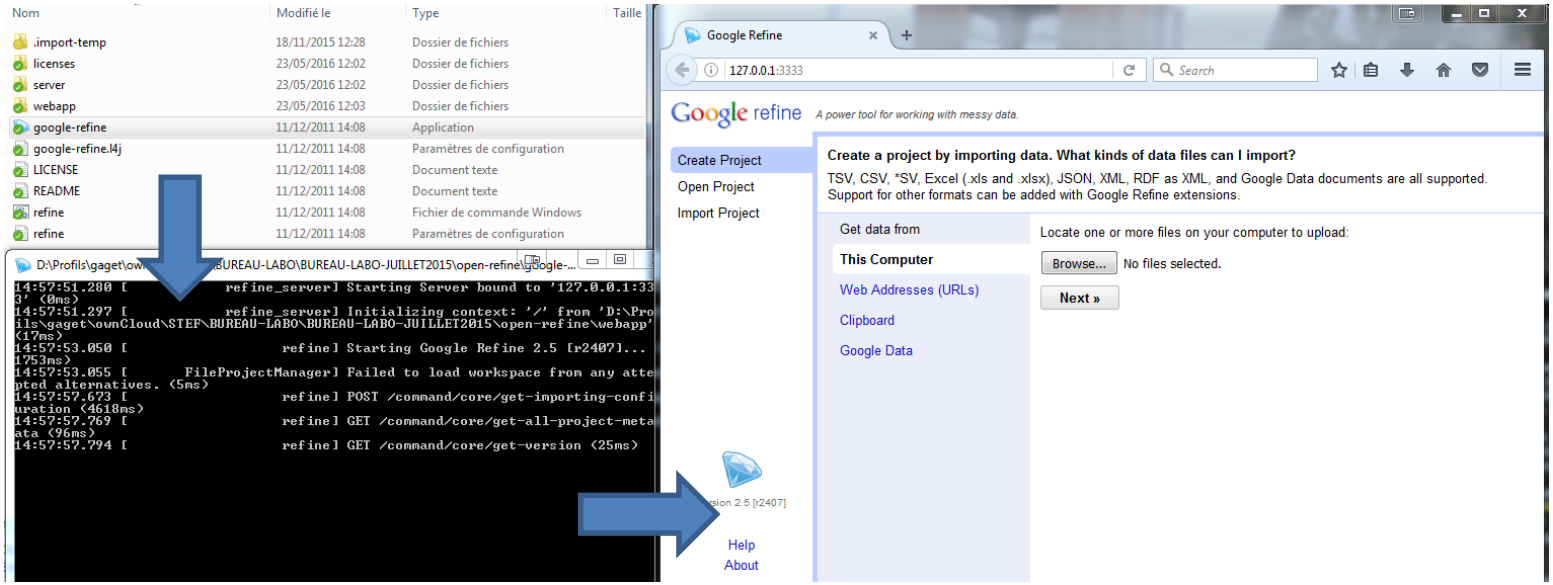
Weborama

- http://www.openrefine.org
- https://github.com/OpenRefine/OpenRefine/wiki

- tutoriels, docs, videos …
  - https://github.com/hpiedcoq/Documentation
  - http://wiki.inra.fr/wiki/traitementsdocumentaires/Main/OpenRefine
  - http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial
  - https://endormitoire.wordpress.com/2016/08/12/openrefine/
  - http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/
  - http://freeyourmetadata.org/
  - http://kb.refinepro.com/
  - ….

Installation

- http://www.openrefine.org/download.html
- OpenRefine 2.6-rc2 Release Candidate 2 (13 octobre 2015, 42Mo)
- Pré-requis : Java JRE
- Installation :
  - Décompresser l'archive téléchargée
  - lancer l'exécutable contenu dans le dossier
  - http://127.0.0.1:3333/ depuis votre navigateur internet



Avec un exemple

	A	B	C	D	E	F	G	H	I	J
1	id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab
2	1	8	49	Women	22.656	27	54	8.2	Aucun	MODY
3	2	8	50	w	29.402	52	71	16.5		MODY
4	3	8	51	F	30.119	80	84	9.1	Aucun	MODY
5	4	8	58	women	19.723	15	28	6.1	Regime	MODY/IG
6	5	8	59	Women	21.99	33	51	7.4	Sulfamides	MODY
7	6	8	256	female	24.69	49	60	8.2	Aucun	MODY
8	7	8	257	Female	26.543	22	33	7.7	Aucun	MODY/IG
9	8	8	1239	m	21.107	60	79	5.1	Sulfamides	MODY
10	9	8	1248	Male	23.457	38	44	5.5	Regime	MODY/IG
11	10	8	1267	Man	20.761	11	30	8	Biguanides	MODY
12	11	8	1270	women	20.761	13	26	7	Regime	MODY/IFG
13	12	8	1496	Men	19.37	17	17	6.6	Aucun	MODY
14	13	8	1535	women	20.83	28	55	8.4	Regime	MODY
15	14	8	1911	women	22.835	28	28	8.2	Aucun	MODY
16	15	8	1990	M	29.136	56	60	7.7	Bs	MODY
17	16	8	2004	F	24.725	35	35	6.8	Aucun	MODY/IFG
18	17	8	2006	f	17.313	8	8	6.6	Aucun	MODY/IFG
19	18	8	2085	Woman	14.863	5	6	6.8	Aucun	MODY/IFG
20	19	8	2101	Man	14.863	5	6	6.9	Aucun	MODY/IFG
21	20	8	2978	f	0	33	33	6.7		MODY/IG
22	21	28	115	women	14.605	13	13	6.8	Aucun	MODY/IG
23	22	28	117	Woman	17.746	7	15	6.5	Regime	MODY/IFG
24	23	28	118	Man	19.195	25	39	7.1	Regime	MODY/IFG
25	24	28	419	F	26.73	57	65	6.4	Bs	MODY
26	25	28	1147	Male	22.275	33	37	6.8	Regime	MODY/IG
27	26	28	1304	Men	20.196	33	32	7.4	Aucun	MODY/IFG
28	27	28	1324	men	22.34	34	33	6.1	Aucun	MODY/IFG
29	28	28	1917	female	21.5	32	32	7.2	Aucun	MODY/IFG
30	29	28	2062	Man	22.863	40	60	8.4	Sulfamides	MODY
31	30	28	2750	women	16.529		5	4.7	Aucun	NonDiab

- Une donnée peut-être codée de multiples façons :
  - [Mm]ale, [Ff]emale, [Hh]omme , [Ff]emme , [Mm]asculin , [Ff]eminin, [Hh], [Ff] , [12], [Ww]om[ae]n, [Mm][ae]n, …

Et les expressions régulières ?

- Une donnée peut-être codée de multiples façons :
  - [Mm]ale, [Ff]emale, [Hh]omme , [Ff]emme , [Mm]asculin , [Ff]eminin, [Hh], [Ff] , [12], [Ww]om[ae]n, [Mm][ae]n, …

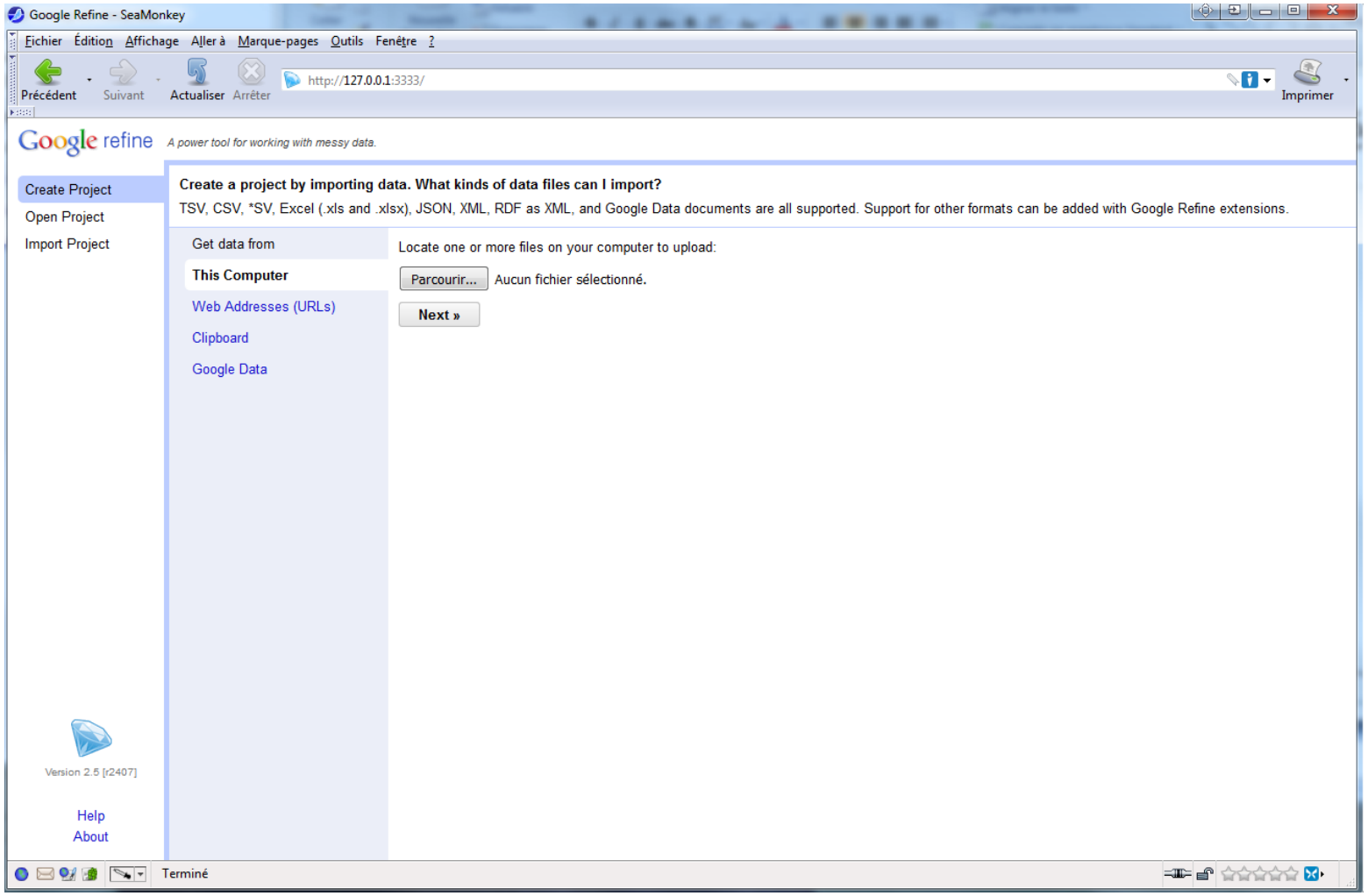
- Solution TGV (Très Geek et Vieux !) :

```
/[Ww]om[ae]n|[Ff]emme|[Ff]emal?e|[Ff]eminin/
```

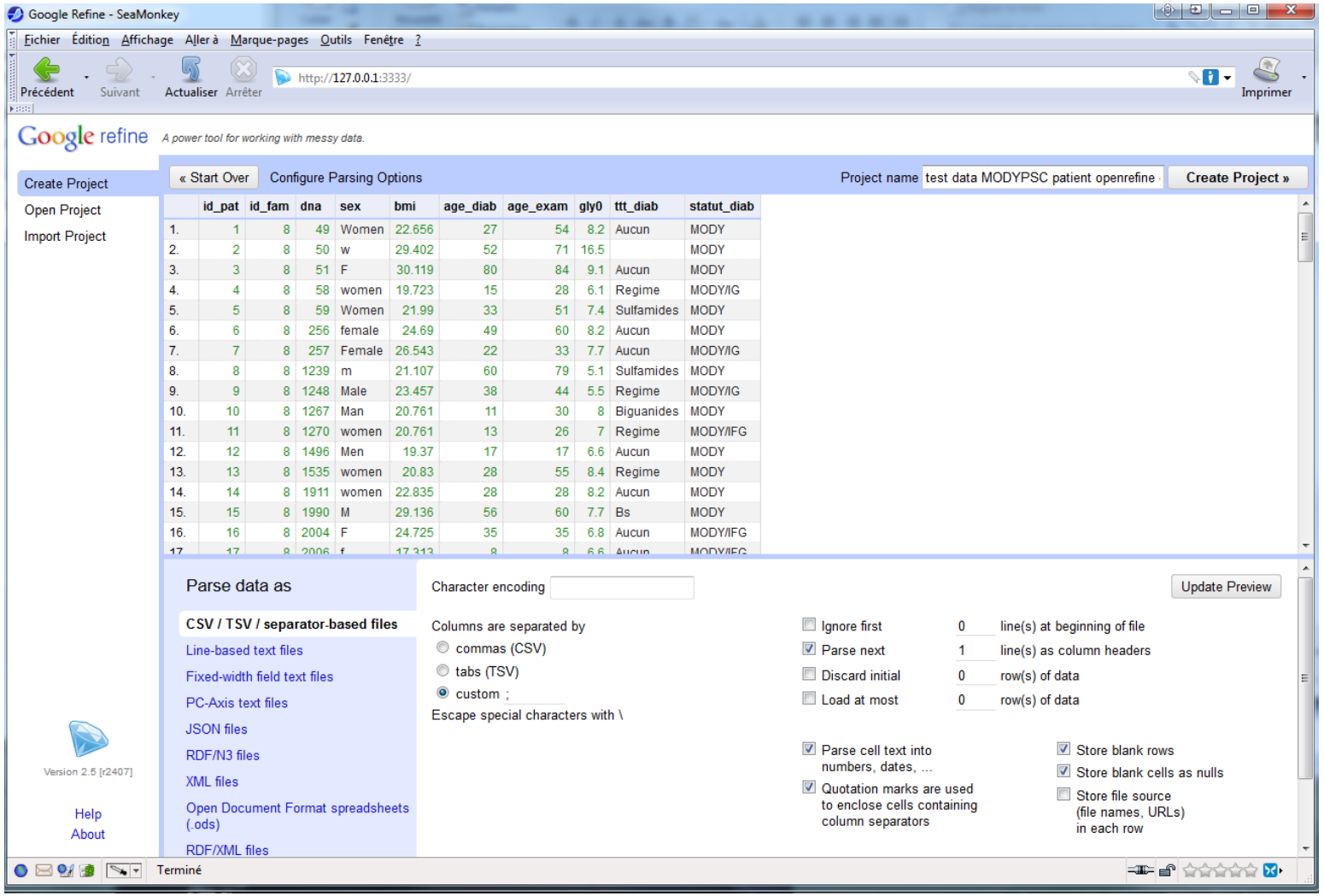
- … mais potentiellement non exhaustive

```
cut –f4 mesdonnees.txt | sort | uniq
```

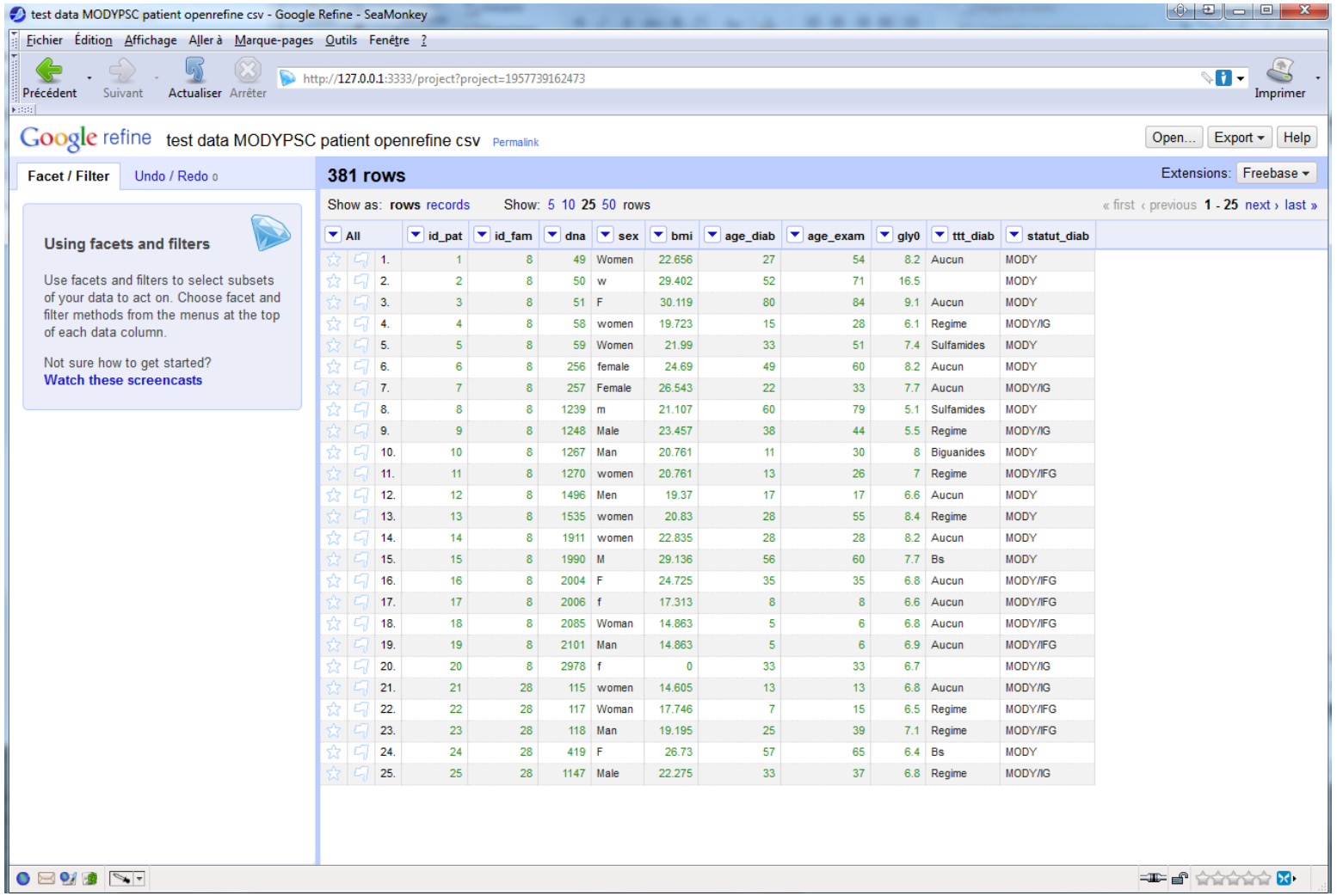
Importer des données



Créer le projet

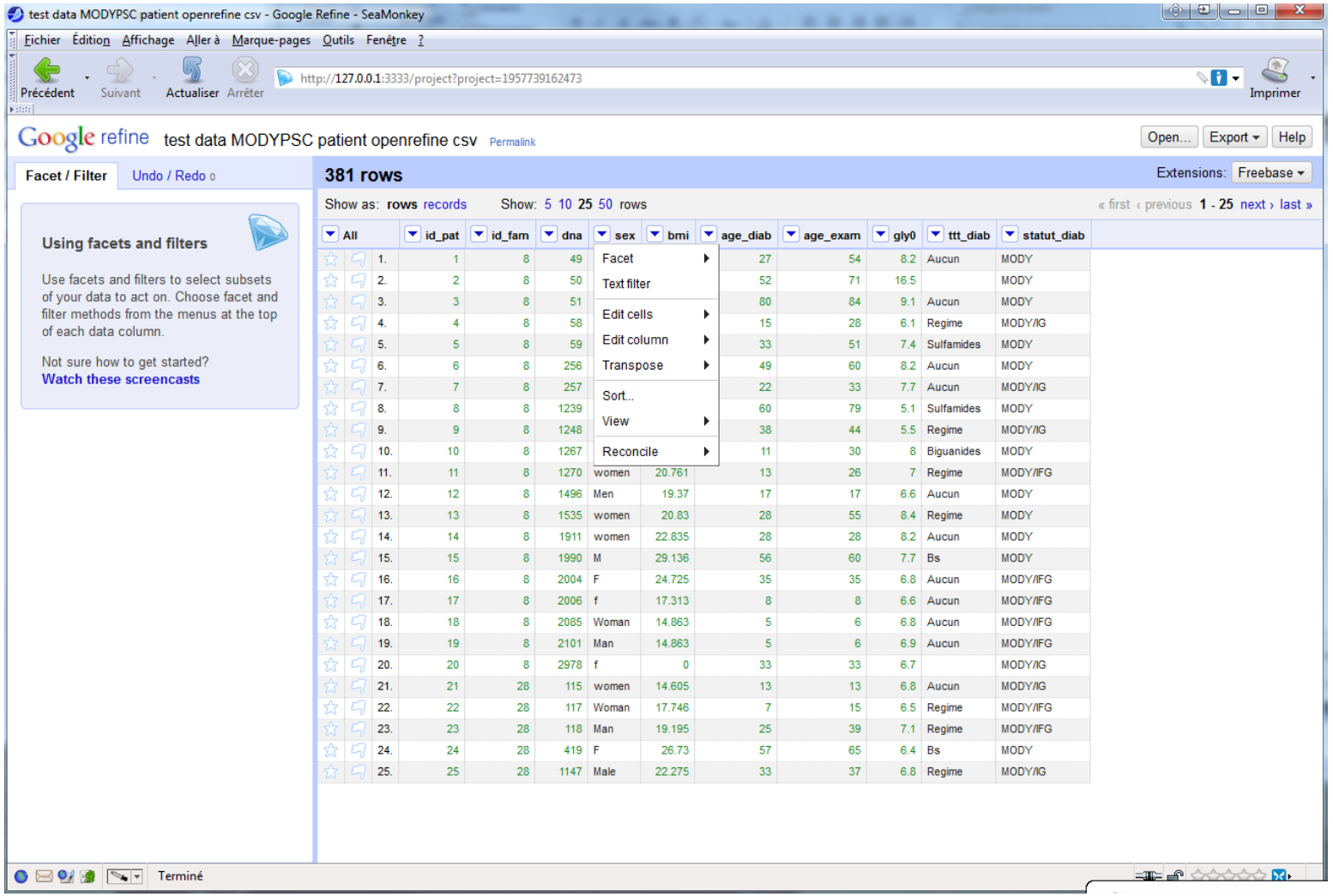


Apparence



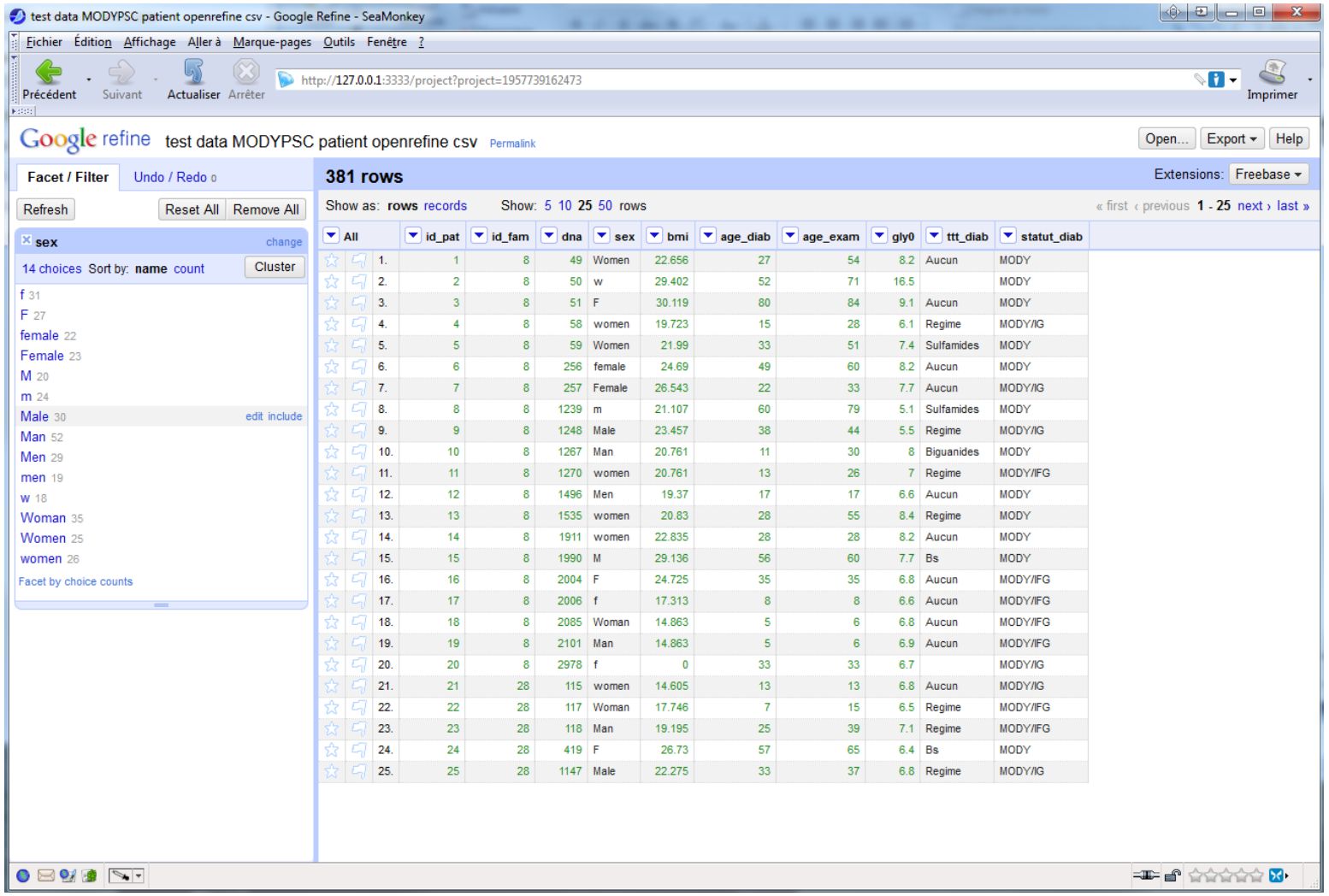
All		id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab
1.	1	8	49	Women	22.656	27	54	8.2	Aucun	MODY
2.	2	8	50	w	29.402	52	71	16.5		MODY
3.	3	8	51	F	30.119	80	84	9.1	Aucun	MODY
4.	4	8	58	women	19.723	15	28	6.1	Regime	MODY/IG
5.	5	8	59	Women	21.99	33	51	7.4	Sulfamides	MODY
6.	6	8	256	female	24.69	49	60	8.2	Aucun	MODY
7.	7	8	257	Female	26.543	22	33	7.7	Aucun	MODY/IG
8.	8	8	1239	m	21.107	60	79	5.1	Sulfamides	MODY
9.	9	8	1248	Male	23.457	38	44	5.5	Regime	MODY/IG
10.	10	8	1267	Man	20.761	11	30	8	Biguanides	MODY
11.	11	8	1270	women	20.761	13	26	7	Regime	MODY/IFG
12.	12	8	1496	Men	19.37	17	17	6.6	Aucun	MODY
13.	13	8	1535	women	20.83	28	55	8.4	Regime	MODY
14.	14	8	1911	women	22.835	28	28	8.2	Aucun	MODY
15.	15	8	1990	M	29.136	56	60	7.7	Bs	MODY
16.	16	8	2004	F	24.725	35	35	6.8	Aucun	MODY/IFG
17.	17	8	2006	f	17.313	8	8	6.6	Aucun	MODY/IFG
18.	18	8	2085	Woman	14.863	5	6	6.8	Aucun	MODY/IFG
19.	19	8	2101	Man	14.863	5	6	6.9	Aucun	MODY/IFG
20.	20	8	2978	f	0	33	33	6.7		MODY/IG
21.	21	28	115	women	14.605	13	13	6.8	Aucun	MODY/IG
22.	22	28	117	Woman	17.746	7	15	6.5	Regime	MODY/IFG
23.	23	28	118	Man	19.195	25	39	7.1	Regime	MODY/IFG
24.	24	28	419	F	26.73	57	65	6.4	Bs	MODY
25.	25	28	1147	Male	22.275	33	37	6.8	Regime	MODY/IG

Fonctionnalités



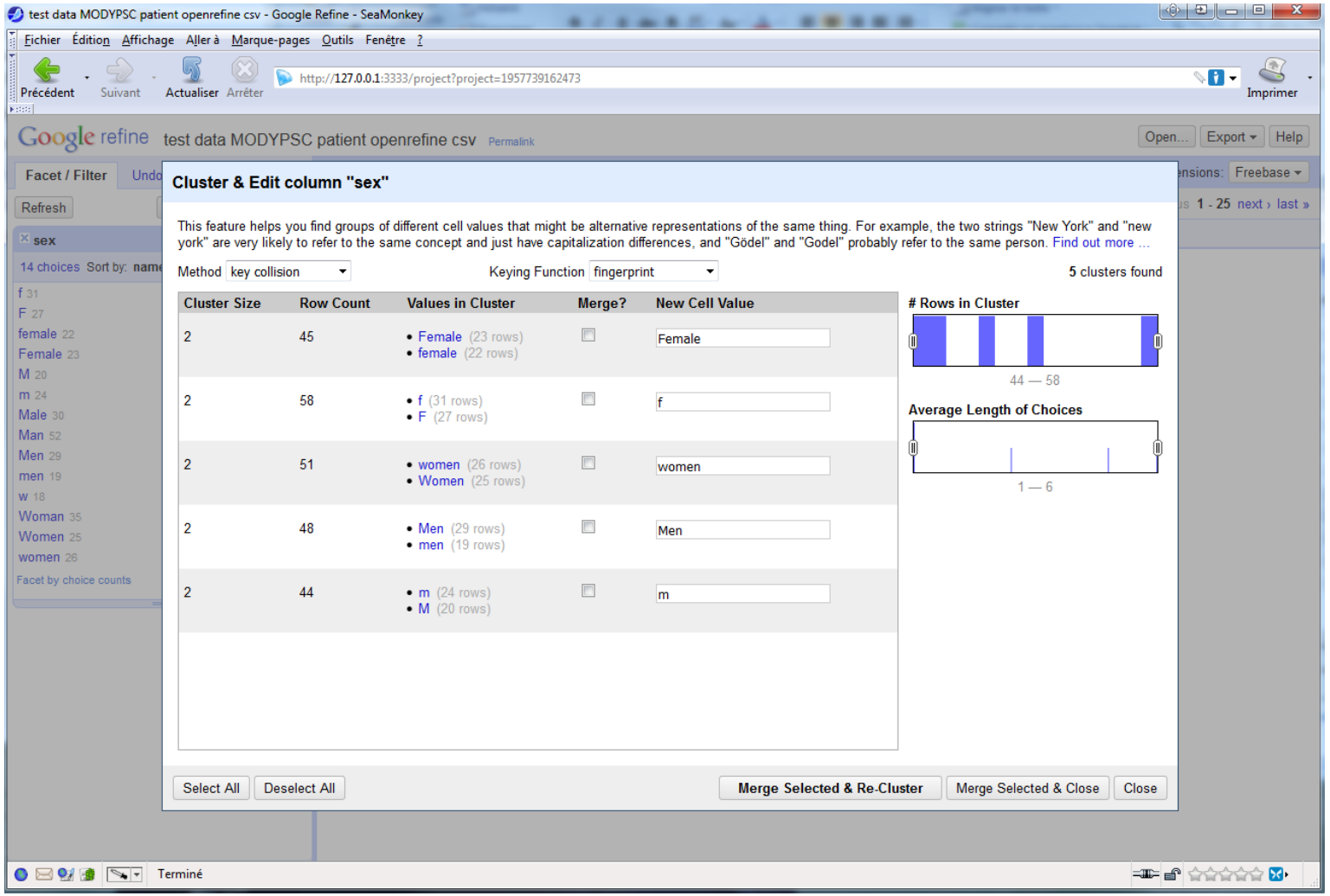
All		id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab
1.	1	8	49			27	54	8.2	Aucun	MODY
2.	2	8	50			52	71	16.5		MODY
3.	3	8	51			80	84	9.1	Aucun	MODY
4.	4	8	58			15	28	6.1	Regime	MODY/IG
5.	5	8	59			33	51	7.4	Sulfamides	MODY
6.	6	8	256			49	60	8.2	Aucun	MODY
7.	7	8	257			22	33	7.7	Aucun	MODY/IG
8.	8	8	1239			60	79	5.1	Sulfamides	MODY
9.	9	8	1248			38	44	5.5	Regime	MODY/IG
10.	10	8	1267			11	30	8	Biguanides	MODY
11.	11	8	1270	women	20.761	13	26	7	Regime	MODY/IFG
12.	12	8	1496	Men	19.37	17	17	6.6	Aucun	MODY
13.	13	8	1535	women	20.83	28	55	8.4	Regime	MODY
14.	14	8	1911	women	22.835	28	28	8.2	Aucun	MODY
15.	15	8	1990	M	29.136	56	60	7.7	Bs	MODY
16.	16	8	2004	F	24.725	35	35	6.8	Aucun	MODY/IFG
17.	17	8	2006	f	17.313	8	8	6.6	Aucun	MODY/IFG
18.	18	8	2085	Woman	14.863	5	6	6.8	Aucun	MODY/IFG
19.	19	8	2101	Man	14.863	5	6	6.9	Aucun	MODY/IFG
20.	20	8	2978	f	0	33	33	6.7		MODY/IG
21.	21	28	115	women	14.605	13	13	6.8	Aucun	MODY/IG
22.	22	28	117	Woman	17.746	7	15	6.5	Regime	MODY/IFG
23.	23	28	118	Man	19.195	25	39	7.1	Regime	MODY/IFG
24.	24	28	419	F	26.73	57	65	6.4	Bs	MODY
25.	25	28	1147	Male	22.275	33	37	6.8	Regime	MODY/IG

Les « Facet »

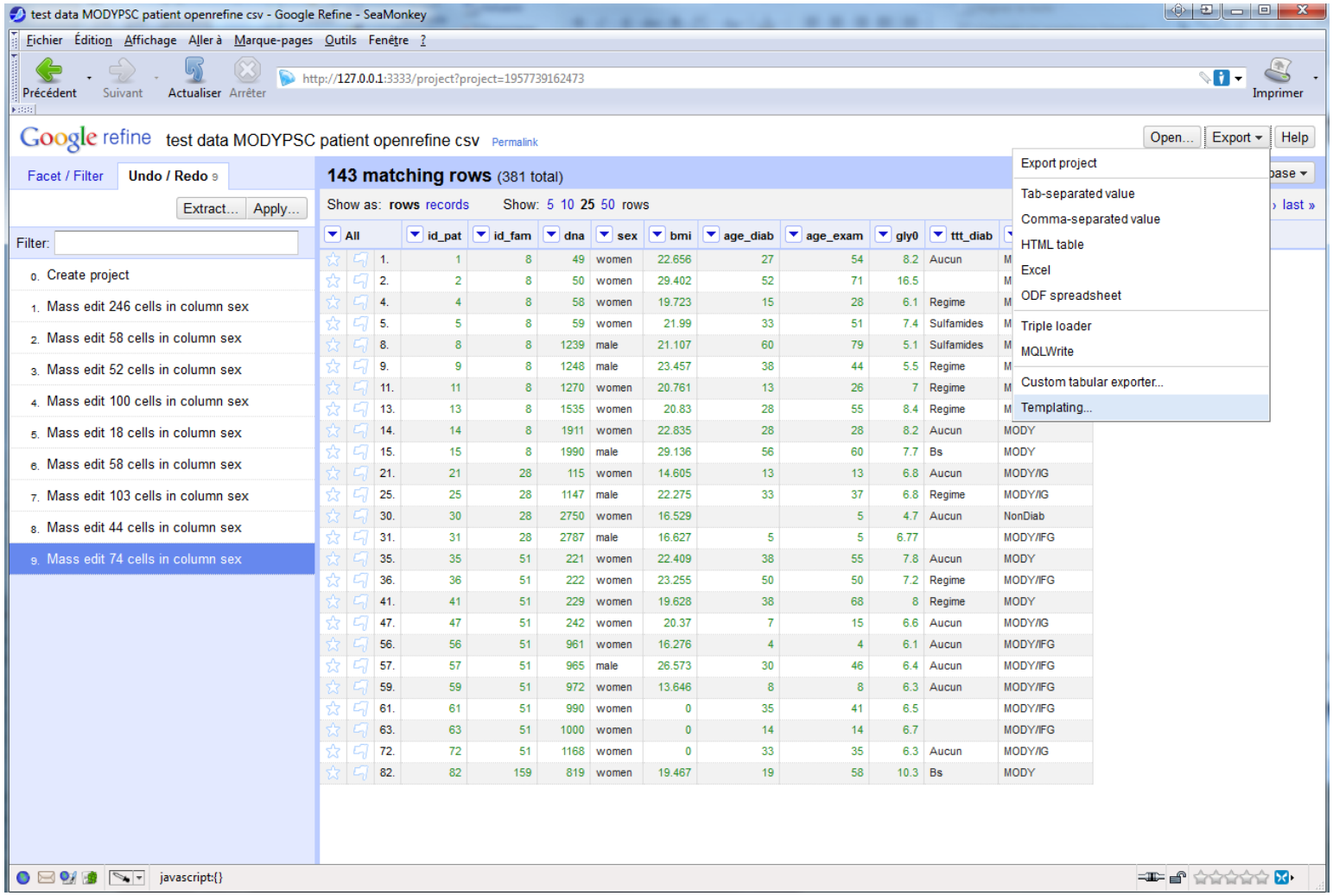


All	id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab
1.	1	8	49	Women	22.656	27	54	8.2	Aucun	MODY
2.	2	8	50	w	29.402	52	71	16.5		MODY
3.	3	8	51	F	30.119	80	84	9.1	Aucun	MODY
4.	4	8	58	women	19.723	15	28	6.1	Regime	MODY/IG
5.	5	8	59	Women	21.99	33	51	7.4	Sulfamides	MODY
6.	6	8	256	female	24.69	49	60	8.2	Aucun	MODY
7.	7	8	257	Female	26.543	22	33	7.7	Aucun	MODY/IG
8.	8	8	1239	m	21.107	60	79	5.1	Sulfamides	MODY
9.	9	8	1248	Male	23.457	38	44	5.5	Regime	MODY/IG
10.	10	8	1267	Man	20.761	11	30	8	Biguanides	MODY
11.	11	8	1270	women	20.761	13	26	7	Regime	MODY/IFG
12.	12	8	1496	Men	19.37	17	17	6.6	Aucun	MODY
13.	13	8	1535	women	20.83	28	55	8.4	Regime	MODY
14.	14	8	1911	women	22.835	28	28	8.2	Aucun	MODY
15.	15	8	1990	M	29.136	56	60	7.7	Bs	MODY
16.	16	8	2004	F	24.725	35	35	6.8	Aucun	MODY/IFG
17.	17	8	2006	f	17.313	8	8	6.6	Aucun	MODY/IFG
18.	18	8	2085	Woman	14.863	5	6	6.8	Aucun	MODY/IFG
19.	19	8	2101	Man	14.863	5	6	6.9	Aucun	MODY/IFG
20.	20	8	2978	f	0	33	33	6.7		MODY/IG
21.	21	28	115	women	14.605	13	13	6.8	Aucun	MODY/IG
22.	22	28	117	Woman	17.746	7	15	6.5	Regime	MODY/IFG
23.	23	28	118	Man	19.195	25	39	7.1	Regime	MODY/IFG
24.	24	28	419	F	26.73	57	65	6.4	Bs	MODY
25.	25	28	1147	Male	22.275	33	37	6.8	Regime	MODY/IG

Agrégation (cluster)

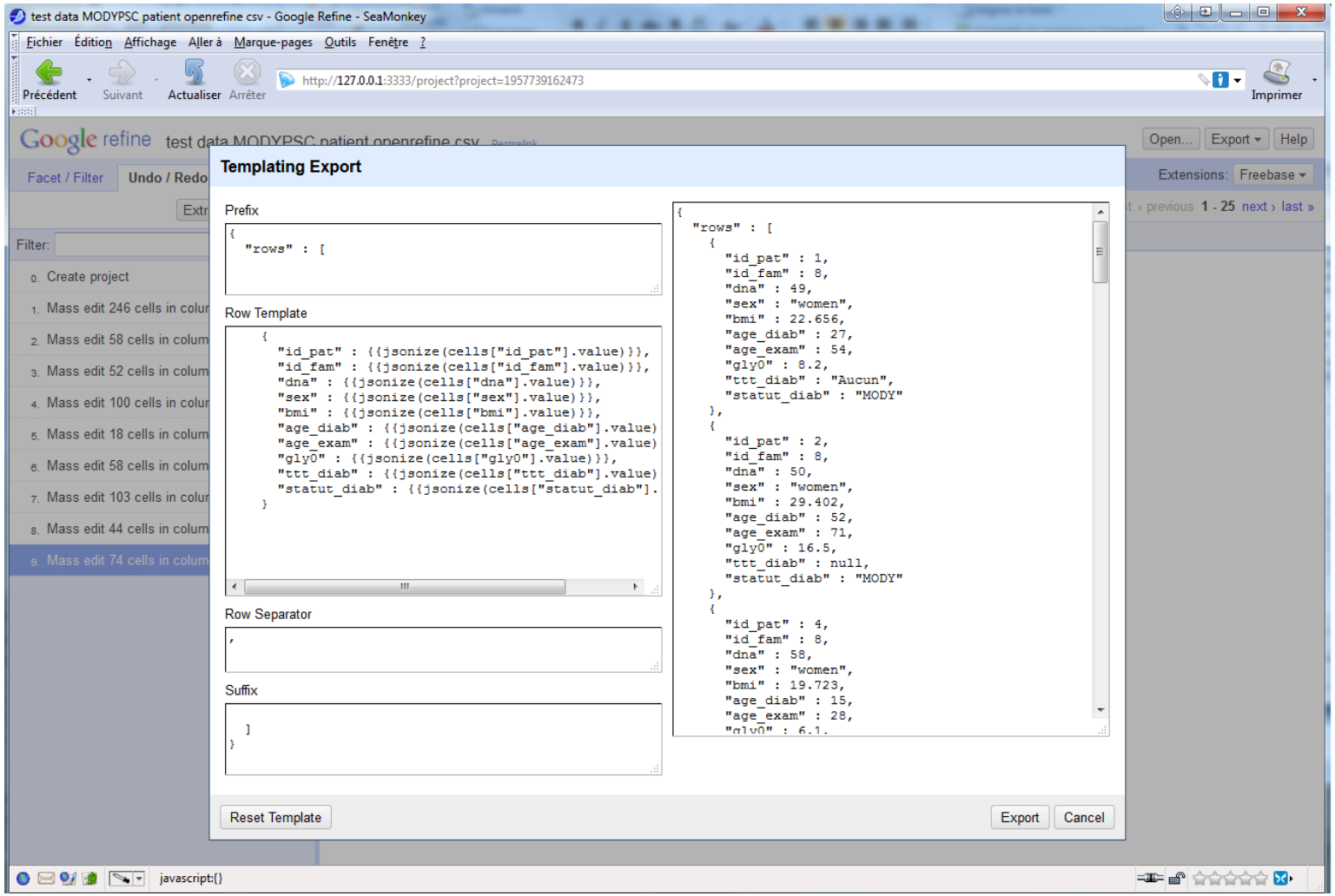


Exporter

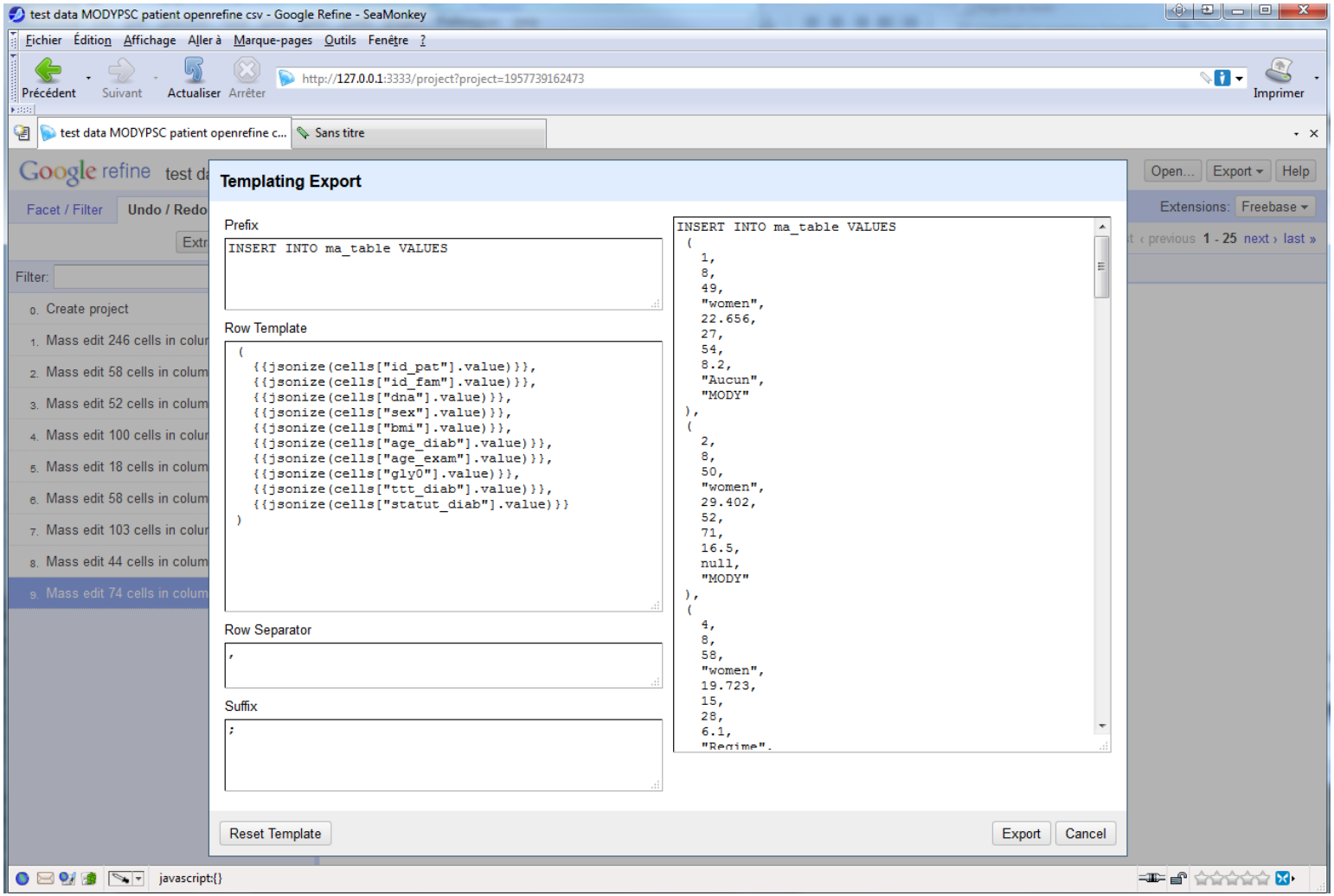


All		id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	
1.	1	8	49	women	22.656	27	54	8.2	Aucun	M
2.	2	8	50	women	29.402	52	71	16.5		M
4.	4	8	58	women	19.723	15	28	6.1	Regime	M
5.	5	8	59	women	21.99	33	51	7.4	Sulfamides	M
8.	8	8	1239	male	21.107	60	79	5.1	Sulfamides	M
9.	9	8	1248	male	23.457	38	44	5.5	Regime	M
11.	11	8	1270	women	20.761	13	26	7	Regime	M
13.	13	8	1535	women	20.83	28	55	8.4	Regime	M
14.	14	8	1911	women	22.835	28	28	8.2	Aucun	MODY
15.	15	8	1990	male	29.136	56	60	7.7	Bs	MODY
21.	21	28	115	women	14.605	13	13	6.8	Aucun	MODY/IG
25.	25	28	1147	male	22.275	33	37	6.8	Regime	MODY/IG
30.	30	28	2750	women	16.529		5	4.7	Aucun	NonDiab
31.	31	28	2787	male	16.627	5	5	6.77		MODY/IFG
35.	35	51	221	women	22.409	38	55	7.8	Aucun	MODY
36.	36	51	222	women	23.255	50	50	7.2	Regime	MODY/IFG
41.	41	51	229	women	19.628	38	68	8	Regime	MODY
47.	47	51	242	women	20.37	7	15	6.6	Aucun	MODY/IG
56.	56	51	961	women	16.276	4	4	6.1	Aucun	MODY/IFG
57.	57	51	965	male	26.573	30	46	6.4	Aucun	MODY/IFG
59.	59	51	972	women	13.646	8	8	6.3	Aucun	MODY/IFG
61.	61	51	990	women	0	35	41	6.5		MODY/IFG
63.	63	51	1000	women	0	14	14	6.7		MODY/IFG
72.	72	51	1168	women	0	33	35	6.3	Aucun	MODY/IG
82.	82	159	819	women	19.467	19	58	10.3	Bs	MODY

Le template


test data MODYPSC patient openrefine csv - Google Refine - SeaMonkey
Fichier Édition Affichage Aller à Marque-pages Outils Fenêtre ?
Précédent Suivant Actualiser Arrêter
http://127.0.0.1:3333/project?project=1957739162473
Imprimer
Google refine test data MODYPSC patient openrefine csv
Open... Export Help
Facet / Filter Undo / Redo
Extensions: Freebase
Filter:
0. Create project
1. Mass edit 246 cells in colur
2. Mass edit 58 cells in colum
3. Mass edit 52 cells in colum
4. Mass edit 100 cells in colur
5. Mass edit 18 cells in colum
6. Mass edit 58 cells in colum
7. Mass edit 103 cells in colur
8. Mass edit 44 cells in colum
9. Mass edit 74 cells in colum
Templating Export
Prefix
{
"rows" : [
Row Template
{
"id_pat" : {{jsonize(cells["id_pat"].value)}},
"id_fam" : {{jsonize(cells["id_fam"].value)}},
"dna" : {{jsonize(cells["dna"].value)}},
"sex" : {{jsonize(cells["sex"].value)}},
"bmi" : {{jsonize(cells["bmi"].value)}},
"age_diab" : {{jsonize(cells["age_diab"].value)
"age_exam" : {{jsonize(cells["age_exam"].value)
"gly0" : {{jsonize(cells["gly0"].value)}},
"ttt_diab" : {{jsonize(cells["ttt_diab"].value)
"statut_diab" : {{jsonize(cells["statut_diab"].
}
Row Separator
,
Suffix
]
}
{
"rows" : [
{
"id_pat" : 1,
"id_fam" : 8,
"dna" : 49,
"sex" : "women",
"bmi" : 22.656,
"age_diab" : 27,
"age_exam" : 54,
"gly0" : 8.2,
"ttt_diab" : "Aucun",
"statut_diab" : "MODY"
},
{
"id_pat" : 2,
"id_fam" : 8,
"dna" : 50,
"sex" : "women",
"bmi" : 29.402,
"age_diab" : 52,
"age_exam" : 71,
"gly0" : 16.5,
"ttt_diab" : null,
"statut_diab" : "MODY"
},
{
"id_pat" : 4,
"id_fam" : 8,
"dna" : 58,
"sex" : "women",
"bmi" : 19.723,
"age_diab" : 15,
"age_exam" : 28,
Reset Template
Export
Cancel
javascript:{}

Le code SQL


test data MODYPSC patient openrefine csv - Google Refine - SeaMonkey
Templating Export
Prefix
INSERT INTO ma_table VALUES
Row Template
(
{{jsonize(cells["id_pat"].value)}},
{{jsonize(cells["id_fam"].value)}},
{{jsonize(cells["dna"].value)}},
{{jsonize(cells["sex"].value)}},
{{jsonize(cells["bmi"].value)}},
{{jsonize(cells["age_diab"].value)}},
{{jsonize(cells["age_exam"].value)}},
{{jsonize(cells["gly0"].value)}},
{{jsonize(cells["ttt_diab"].value)}},
{{jsonize(cells["statut_diab"].value)}}
)
Row Separator
,
Suffix
;
INSERT INTO ma_table VALUES
(
1,
8,
49,
"women",
22.656,
27,
54,
8.2,
"Aucun",
"MODY"
),
(
2,
8,
50,
"women",
29.402,
52,
71,
16.5,
null,
"MODY"
),
(
4,
8,
58,
"women",
19.723,
15,
28,
6.1,
Reset Template
Export
Cancel

Quelques remarques

- Solutions rapidement réalisables
- Des solutions « one shoot »
- Utilisables en routine pour certaines
- Attention à la qualité des données
- Et les ETL dans tout cela (Extraction – Transformation – Loading) ?
  - Talend Open studio : Data Integration

Programme :

- Les données :
  - codification
  - contrôle de qualité
  - erreurs courantes dans la manipulation
- Alimenter une base de données :
  - vue d'ensemble
  - préparation des données
- Formater des données avec OpenRefine :
  - introduction
  - installation
  - processus général par l'étude d'un cas pratique
  - fonctionnalités
- Un projet de A à Z par la pratique