

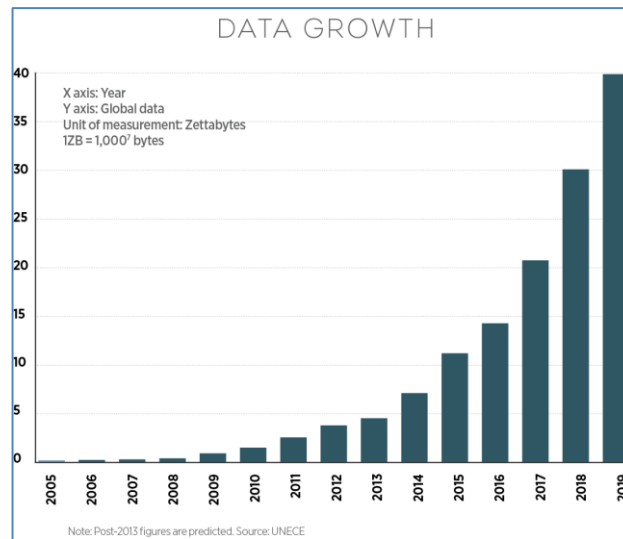
Data cleaning avec OpenRefine : nettoyer et enrichir ses données



Stefan Gaget, stefan.gaget@cnr.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

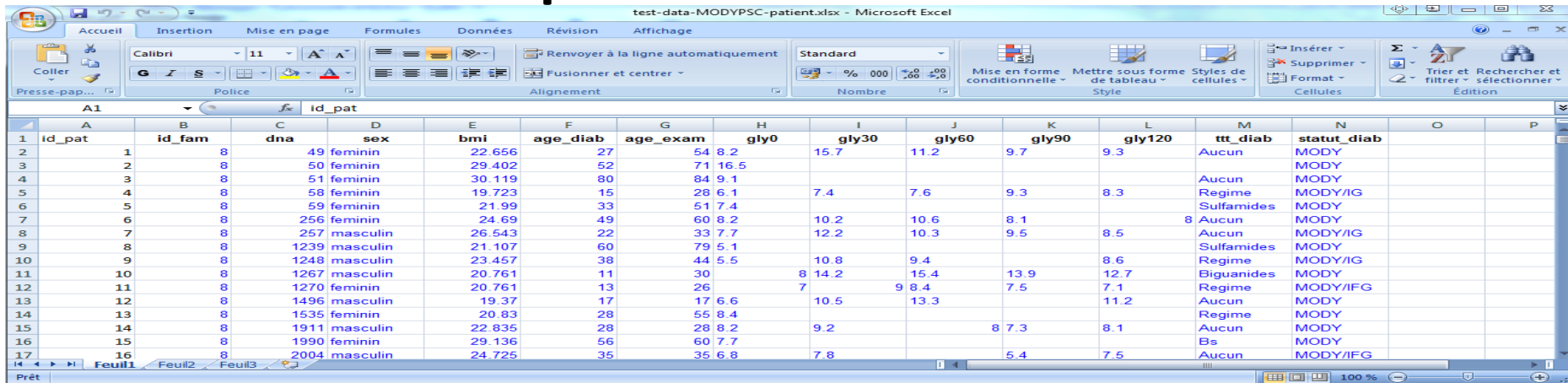
Les solutions

- « **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



The screenshot shows a Microsoft Excel spreadsheet with the following data:

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

The screenshot shows an Excel spreadsheet with the following data:

A	B	C	D	E	F	G	H	I	J	K	L	M	N		
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	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

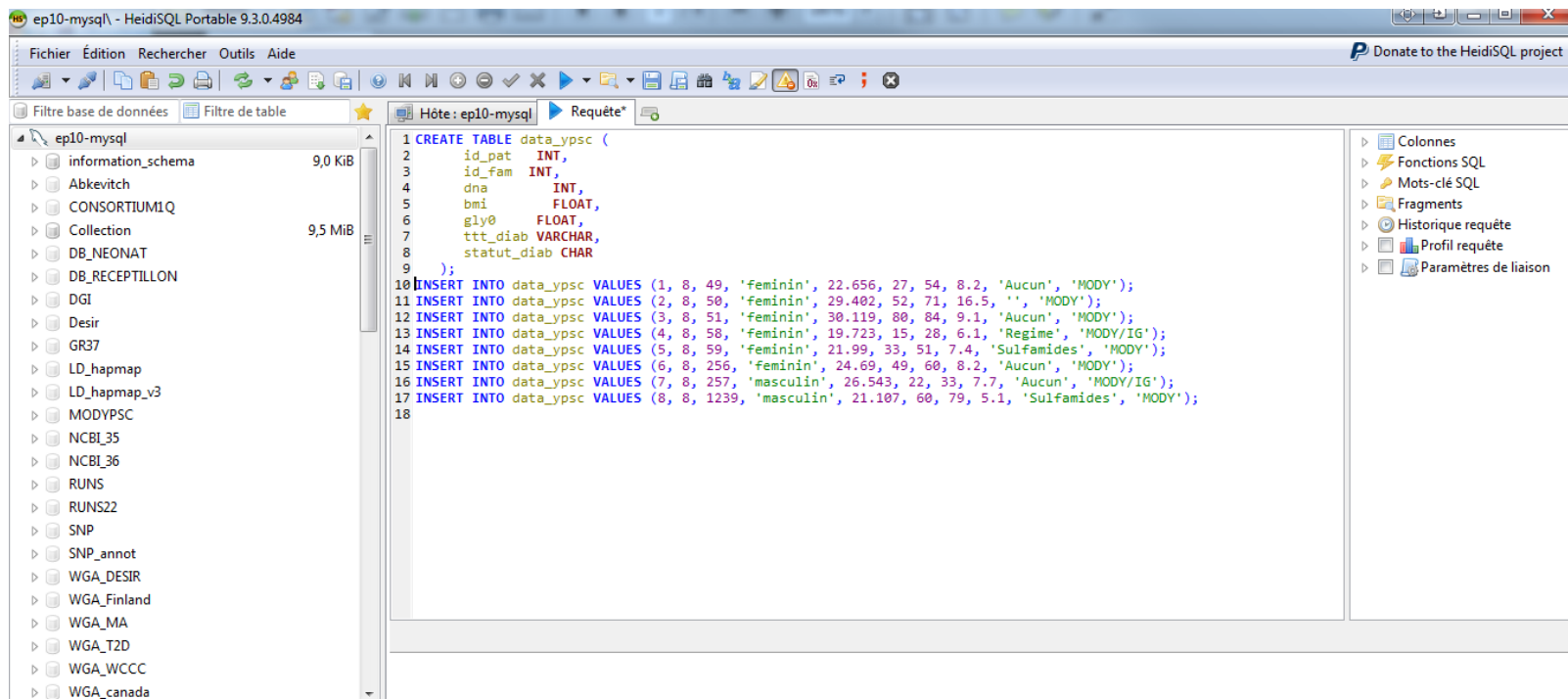
VALUES (";A9;" , ";B9;" , ";C9;" , "";D9;" , ";E9;" , ";F9;" , ";G9;" , ";H9;" , "";I9;" , "";J9;"");")

CONCATENER																		
=CONCATENER("INSERT INTO sites VALUES (";A9;" , ";B9;" , ";C9;" , "";D9;" , ";E9;" , ";F9;" , ";G9;" , ";H9;" , "";I9;" , "";J9;"");")																		
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
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;" , "";I9;" , "";J9;"");")							
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...)



```
1 CREATE TABLE data_ypsc (
2   id_pat INT,
3   id_fam INT,
4   dna INT,
5   bmi FLOAT,
6   gly0 FLOAT,
7   ttt_diab VARCHAR,
8   statut_diab CHAR
9 );
10 INSERT INTO data_ypsc VALUES (1, 8, 49, 'feminin', 22.656, 27, 54, 8.2, 'Aucun', 'MODY');
11 INSERT INTO data_ypsc VALUES (2, 8, 50, 'feminin', 29.402, 52, 71, 16.5, '', 'MODY');
12 INSERT INTO data_ypsc VALUES (3, 8, 51, 'feminin', 30.119, 80, 84, 9.1, 'Aucun', 'MODY');
13 INSERT INTO data_ypsc VALUES (4, 8, 58, 'feminin', 19.723, 15, 28, 6.1, 'Regime', 'MODY/IG');
14 INSERT INTO data_ypsc VALUES (5, 8, 59, 'feminin', 21.99, 33, 51, 7.4, 'Sulfamides', 'MODY');
15 INSERT INTO data_ypsc VALUES (6, 8, 256, 'feminin', 24.69, 49, 60, 8.2, 'Aucun', 'MODY');
16 INSERT INTO data_ypsc VALUES (7, 8, 257, 'masculin', 26.543, 22, 33, 7.7, 'Aucun', 'MODY/IG');
17 INSERT INTO data_ypsc VALUES (8, 8, 1239, 'masculin', 21.107, 60, 79, 5.1, 'Sulfamides', 'MODY');
18
```

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

```
D:\Profils\gaget\Desktop\test-data-MODYPSC-patient.csv - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
index.html dbConfig.php test-data-MODYPSC-patient.csv x
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
3 2;8;50;feminin;29.402;52;71;16.5;;MODY
4 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
6 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
8 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
10 9;8;1248;masculin;23.457;38;44;5.5;Regime;MODY/IG
11 10;8;1267;masculin;20.761;11;30;8;Biguanides;MODY
12 11;8;1270;feminin;20.761;13;26;7;Regime;MODY/IFG
13 12;8;1496;masculin;19.37;17;17;6.6;Aucun;MODY
14 13;8;1535;feminin;20.83;28;55;8.4;Regime;MODY
15 14;8;1911;masculin;22.835;28;28;8.2;Aucun;MODY
16 15;8;1990;feminin;29.136;56;60;7.7;Bs;MODY
17 16;8;2004;masculin;24.725;35;35;6.8;Aucun;MODY/IFG
18 17;8;2006;masculin;17.313;8;8;6.6;Aucun;MODY/IFG
19 18;8;2085;feminin;14.863;5;6;6.8;Aucun;MODY/IFG
20 19;8;2101;masculin;14.863;5;6;6.9;Aucun;MODY/IFG
21 20;8;2978;feminin;0;33;33;6.7;;MODY/IG
22 21;28;115;feminin;14.605;13;13;6.8;Aucun;MODY/IG
23 22;28;117;masculin;17.746;7;15;6.5;Regime;MODY/IFG
24 23;28;119;feminin;18.105;25;25;7.1;Aucun;MODY/IFG
Line 1, Column 1 Tab Size: 4 Plain Text
```

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

```
1 id_pat;id_fam;dna;sex;bmi;age_diab;age_exam;gly0;ttdiab;statut_diab
2 1;8;49;feminin;22.656;27;54;8.2;Aucun;MODY
3 2;8;50;feminin;29.402;52;71;16.5;;MODY
4 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
6 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
8 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
10 9;8;1248;masculin;23.457;38;44;5.5;Regime;MODY/IG
11 10;8;1267;masculin;20.761;11;30;8;Biguanides;MODY
12 11;8;1270;feminin;20.761;13;26;7;Regime;MODY/IFG
13 12;8;1496;masculin;19.37;17;17;6.6;Aucun;MODY
14 13;8;1535;feminin;20.83;28;55;8.4;Regime;MODY
15 14;8;1911;masculin;22.835;28;28;8.2;Aucun;MODY
16 15;8;1990;feminin;29.136;56;60;7.7;Bs;MODY
17 16;8;2004;masculin;24.725;35;35;6.8;Aucun;MODY/IFG
18 17;8;2006;masculin;17.313;8;8;6.6;Aucun;MODY/IFG
19 18;8;2085;feminin;14.863;5;6;6.8;Aucun;MODY/IFG
20 19;8;2101;masculin;14.863;5;6;6.9;Aucun;MODY/IFG
21 20;8;2978;feminin;0;33;33;6.7;;MODY/IG
22 21;28;115;feminin;14.605;13;13;6.8;Aucun;MODY/IG
23 22;28;117;masculin;17.746;7;15;6.5;Regime;MODY/IFG
```

- **COPY data_ypsc (id_pat, id_fam, dna, bmi, gly0, ttdiab, 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

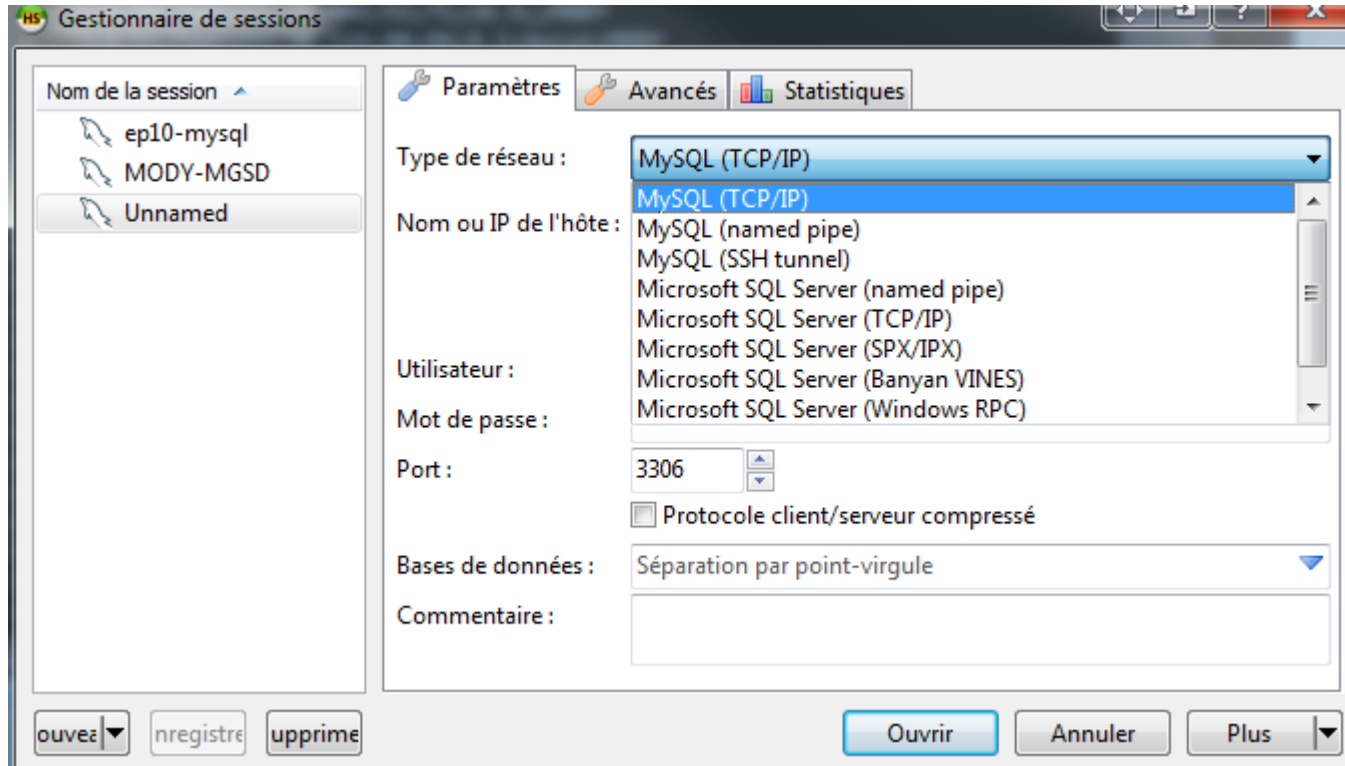
- **Création, éditions de bases, tables, ...**
- **Edition, exécution d'instructions SQL**
- **Constructeur de requête**
- **Importer/Exporter données et structures**

- **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)

The screenshot shows the HeidiSQL interface for a MySQL database named 'ep10-mysql\NCBI_35'. The main window displays a table with the following columns: Nom, Lignes, Taille, Créé(es), Mis à jour, Moteur, Commentaire, and Type. The table lists various tables within the NCBI_35 database, including puce_1M, snp, snp2gene, and snp_gene.

Nom	Lignes	Taille	Créé(es)	Mis à jour	Moteur	Commentaire	Type
puce_1M	1 145 510	57,1 MiB	2009-06-29 11:12:06	2014-10-02 01:27:29	MyISAM		Table
snp	10 752 546	664,5 MiB	2008-06-10 11:27:48	2014-10-02 01:28:30	MyISAM		Table
snp2gene	10 752 546	1,1 GiB	2007-05-02 18:09:05	2014-10-02 01:34:00	MyISAM		Table
snp2gene_affy...	115 276	10,9 MiB	2007-05-03 10:50:03	2014-10-02 01:34:02	MyISAM		Table
snp2gene_affy...	493 059	46,9 MiB	2007-05-03 10:50:05	2014-10-02 01:34:12	MyISAM		Table
snp2gene_illu...	408 777	39,0 MiB	2007-05-03 10:50:17	2014-10-02 01:34:20	MyISAM		Table
snp2gene_illu...	109 344	10,6 MiB	2007-05-03 10:49:54	2014-10-02 01:34:22	MyISAM		Table
snp2gene_illu...	317 501	30,3 MiB	2007-05-03 10:49:55	2014-10-02 01:34:28	MyISAM		Table
snp2gene_jcc...	10 752 546	664,5 MiB	2008-06-10 10:45:38	2014-10-02 01:35:28	MyISAM		Table
snp_gene	44 794 579	1,5 GiB	2008-06-12 15:10:13	2014-10-02 01:41:03	MyISAM		Table

The bottom panel shows the following SQL commands and their execution status:

```
51 SELECT `DEFAULT_COLLATION_NAME` FROM `information_schema`.`SCHEMATA` WHERE `SCHEMA_NAME`='NCBI_35';
52 SHOW TABLE STATUS FROM `NCBI_35`;
53 SHOW FUNCTION STATUS WHERE `Db`='NCBI_35';
54 /* Erreur SQL (1547) : Column count of mysql.proc is wrong. Expected 20, found 16. The table is probably corrupted */
55 SHOW PROCEDURE STATUS WHERE `Db`='NCBI_35';
56 /* Erreur SQL (1547) : Column count of mysql.proc is wrong. Expected 20, found 16. The table is probably corrupted */
57 SHOW TRIGGERS FROM `NCBI_35`;
58 SELECT *, EVENT_SCHEMA AS `Name` FROM information_schema.`EVENTS` WHERE `EVENT_SCHEMA`='NCBI_35';
59 /* Erreur SQL (1577) : Cannot proceed because system tables used by Event Scheduler were found damaged at server start */
60 SHOW /*!50002 GLOBAL */ STATUS LIKE 'Com\%';
```

The status bar at the bottom indicates: Connecté: 00:01 h, MySQL 5.5.38, Disponibilité: 6 jours, 21:23 h, En attente.

Vue générale (2)

The screenshot shows the HeidiSQL interface with the following components:

- Left Panel:** A tree view of the database structure. The 'NCBI_35' database is selected, showing a list of tables with their sizes. The 'snp2gene' table is highlighted, showing a size of 1,1 GiB.
- Top Panel:** The 'Table : snp2gene' tab is active, showing the table's structure. The 'snp_id' column is highlighted as the primary key.
- Table Structure:** A table with 8 columns: #, Nom, Type de données, Taille/Ensem..., Non signé, NULL autorisé, Compl..., Par défaut, and Commentaire. The columns are: 1 snp_id (VARCHAR, 14), 2 chrom (VARCHAR, 15), 3 start (INT, 11), 4 end (INT, 11), 5 gene (VARCHAR, 100), 6 gene_gauche (VARCHAR, 100), 7 gene_droit (VARCHAR, 100), and 8 additional (TINYINT, 4).
- Bottom Panel:** A SQL query window showing the following commands:

```
65 SHOW CREATE TABLE `NCBI_35`.`snp2gene`;  
66 SHOW TABLE STATUS LIKE `snp2gene`;  
67 SHOW CREATE TABLE `NCBI_35`.`snp`;  
68 SELECT * FROM `NCBI_35`.`snp` LIMIT 1000;  
69 SHOW CREATE TABLE `NCBI_35`.`snp`;  
70 SHOW TABLE STATUS LIKE `snp`;  
71 SHOW CREATE TABLE `NCBI_35`.`snp2gene_affy_500K`;  
72 SHOW CREATE TABLE `NCBI_35`.`snp_gene`;  
73 SHOW CREATE TABLE `NCBI_35`.`snp2gene_affy_100K`;  
74 SHOW CREATE TABLE `NCBI_35`.`snp2gene`;
```
- Status Bar:** Shows 'Connecté: 00:04 h', 'MySQL 5.5.38', 'Disponibilité: 6 jours, 21:26 h', and 'En attente.'

Vue générale (3)

The screenshot shows the HeidiSQL interface with the following components:

- Database Tree (Left):** Shows the hierarchy of databases and tables. The 'NCBI_35' database is selected, and the 'snp2gene' table is highlighted. Other tables include 'snp2gene_affy_100K', 'snp2gene_affy_500K', 'snp2gene_illumina', 'snp2gene_illumina_100K', 'snp2gene_illumina_317K', 'snp2gene_jcc_method', and 'snp_gene'.
- Query Editor (Top Center):** Contains the SQL query: `1 select * from snp2gene where chrom='chr22'`
- Results Panel (Middle):** Displays the query results in a table with 8 columns: `snp_id`, `chrom`, `start`, `end`, `gene`, `gene_gauche`, `gene_droit`, and `additional`. The table contains 154,936 rows of data.
- Command Window (Bottom):** Shows the execution of the query and its output, including the number of affected rows and execution time.

```
72 SHOW CREATE TABLE `NCBI_35`.`snp_gene`;
73 SHOW CREATE TABLE `NCBI_35`.`snp2gene_affy_100K`;
74 SHOW CREATE TABLE `NCBI_35`.`snp2gene`;
75 select * from snp2gene where chrom=22;
76 /* Affected rows: 0 Lignes trouvées: 0 Avertissements: 0 Durée pour 1 query: 5,476 sec. */
77 SELECT * FROM `NCBI_35`.`snp2gene` LIMIT 1000;
78 SHOW CREATE TABLE `NCBI_35`.`snp2gene`;
79 SHOW TABLE STATUS LIKE 'snp2gene';
80 select * from snp2gene where chrom='chr22';
81 /* Affected rows: 0 Lignes trouvées: 154 936 Avertissements: 0 Durée pour 1 query: 0,499 sec. (+ 1,887 sec. network) */
```

snp_id	chrom	start	end	gene	gene_gauche	gene_droit	additional
rs12160885	chr22	49 479 289	49 479 290		ACR/5419	LOC150417/3726	0
rs9680581	chr22	39 466 015	39 466 016		MCHR1/62698	SLC25A17/24119	0
rs12158589	chr22	20 870 069	20 870 070	IGL@	BMP6P1/4010	IGLV6-57/4864	0
rs9618163	chr22	15 378 007	15 378 008		LOC644773/34248	LOC644784/49500	0
rs9618165	chr22	15 378 248	15 378 249		LOC644773/34489	LOC644784/49259	0
rs3888438	chr22	15 729 465	15 729 466		ZNF402P/13291	LOC644845/22121	0
rs2885426	chr22	15 735 233	15 735 234		ZNF402P/19059	LOC644845/16353	0
rs4006206	chr22	15 752 127	15 752 128	LOC644845	ZNF402P/35953	LOC644851/7509	0
rs4023016	chr22	15 761 668	15 761 669	LOC644851	LOC644845/5792	LOC440786/7452	0
rs2885426	chr22	15 735 233	15 735 234		ZNF402P/19059	LOC644845/16353	0
rs4239862	chr22	15 729 525	15 729 526		ZNF402P/13351	LOC644845/22061	0
rs3888438	chr22	15 729 465	15 729 466		ZNF402P/13291	LOC644845/22121	0
rs3865632	chr22	15 379 407	15 379 409		LOC644773/35648	LOC644784/48099	0
rs4010369	chr22	15 313 031	15 313 031	LOC644768	LOC644758/31899	LOC644773/29403	0
rs4010475	chr22	15 280 494	15 280 501		ABCD1P4/34300	LOC644758/401	0
rs4010493	chr22	15 279 936	15 279 936		ABCD1P4/33742	LOC644758/966	0
rs11551383	chr22	36 598 352	36 598 353	EIF3S61P	ANKRD54/33550	MICAL-L1/28492	0
rs11554923	chr22	37 443 499	37 443 500	GTPBP1	JOSD1/22033	ENST00000216044/5134	0
rs11554915	chr22	37 443 735	37 443 736	GTPBP1	JOSD1/22269	ENST00000216044/4898	0
rs12172224	chr22	33 400 365	33 400 366		LOC646680/16795	RAXLX/386317	0

L'outil Importer dans HeidiSQL

ep10-mysql - HeidiSQL Portable 9.3.0.4984

Fichier Édition Rechercher Outils Aide

ep10-mysql

information_schema 9,0 KiB

Abkevitch

CONSORTIUM1Q

Collection 9,5 MiB

DB_NEONAT

DB_RECEPTILLON

DGI

Desir

GR37

LD_hapmap

LD_hapmap_v3

MODYPSC

NCBI_35

NCBI_36

RUNS

RUNS22

SNP

SNP_annot

WGA_DESIR

WGA_Finland

WGA_MA

WGA_T2D

WGA_WCCC

WGA_canada

```
1 CREATE TABLE data_ypsc (  
2 id  
3 id  
4 dna  
5 bmi  
6 gly  
7 ttt  
8 sta  
9 );  
10 INSERT IN  
11 INSERT IN  
12 INSERT IN  
13 INSERT IN  
14 INSERT IN  
15 INSERT IN  
16 INSERT IN  
17 INSERT IN  
18
```

Importer fichier texte

Fichiers en entrée

Nom fichier : D:\Profils\gaget\Desktop\IndexHaloplex.csv

encodage : Laisser serveur/base de données décider (utf8)

Options

Ignorer 1ère 0 lignes

Priorité passe, pour éviter une erreur de la langue, par ex. 1234.56 en

nombres formatés en fonction de la langue, par ex. 1234.56 en

Tronquer la table de destination à

Gestion des lignes dupliquées

INSERT (peut provoquer des erreurs)

INSERT IGNORE (lignes dupliquées)

REPLACE (lignes dupliquées)

Méthode

Interprétation du contenu par le serveur (LOAD DATA)

Interprétation du contenu par le client

Caractères de contrôle

Champs terminés par ;

Champs délimités par " facultatif

Champs échappés par "

Lignes terminées par \r\n

Destination

Base de données : information_schema

Table : CHARACTER_SETS

Colonnes :

CHARACTER_SET_NAME

DEFAULT_COLLATE_NAME

DESCRIPTION

MAXLEN

Importer Annuler

```
25 SHOW TRIGGERS FROM `information_schema`;  
26 SHOW EVENTS FROM `information_schema`;  
27 /* Erreur SQL (1577) : Cannot proceed because system tables used by Event Scheduler were found damaged at server start */  
28 SELECT *, EVENT_SCHEMA AS `Db`, EVENT_NAME AS `Name` FROM information_schema.EVENTS` WHERE EVENT_SCHEMA='Collection';  
29 /* Erreur SQL (1577) : Cannot proceed because system tables used by Event Scheduler were found damaged at server start */  
30 SHOW CREATE DATABASE `information_schema`;  
31 SHOW CHARSET;  
32 SHOW CREATE TABLE `information_schema`.`CHARACTER_SETS`;  
33 SHOW COLLATION;  
34 SHOW CREATE DATABASE `information_schema`;
```

L'outil Importer dans phpMyAdmin

phpMyAdmin

Seigneur: production

Bases de données SQL État Exporter Importer Paramètres Variables Jeux de caractères Moteurs

Importation dans le serveur actuel

Fichier à importer :

Le fichier peut être comprimé (gzip, bzip2, zip) ou non.
Le nom du fichier comprimé doit se terminer par `.[format].[compression]`. Exemple: `.sql.zip`

Parcourir : Aucun fichier sélectionné. (Taille maximum: 100Mio)

Choisissez depuis le répertoire de téléchargement du serveur web `/tmp/` :

Jeu de caractères du fichier :

Importation partielle :

Permettre l'interruption de l'importation si la limite de temps configurée dans PHP est sur le point d'être atteinte. *(Ceci pourrait aider à importer des fichiers volumineux, au détriment du respect des transactions.)*

Ignorer ce nombre de requêtes (pour SQL) ou de lignes (autres formats), à partir du début :

Format :

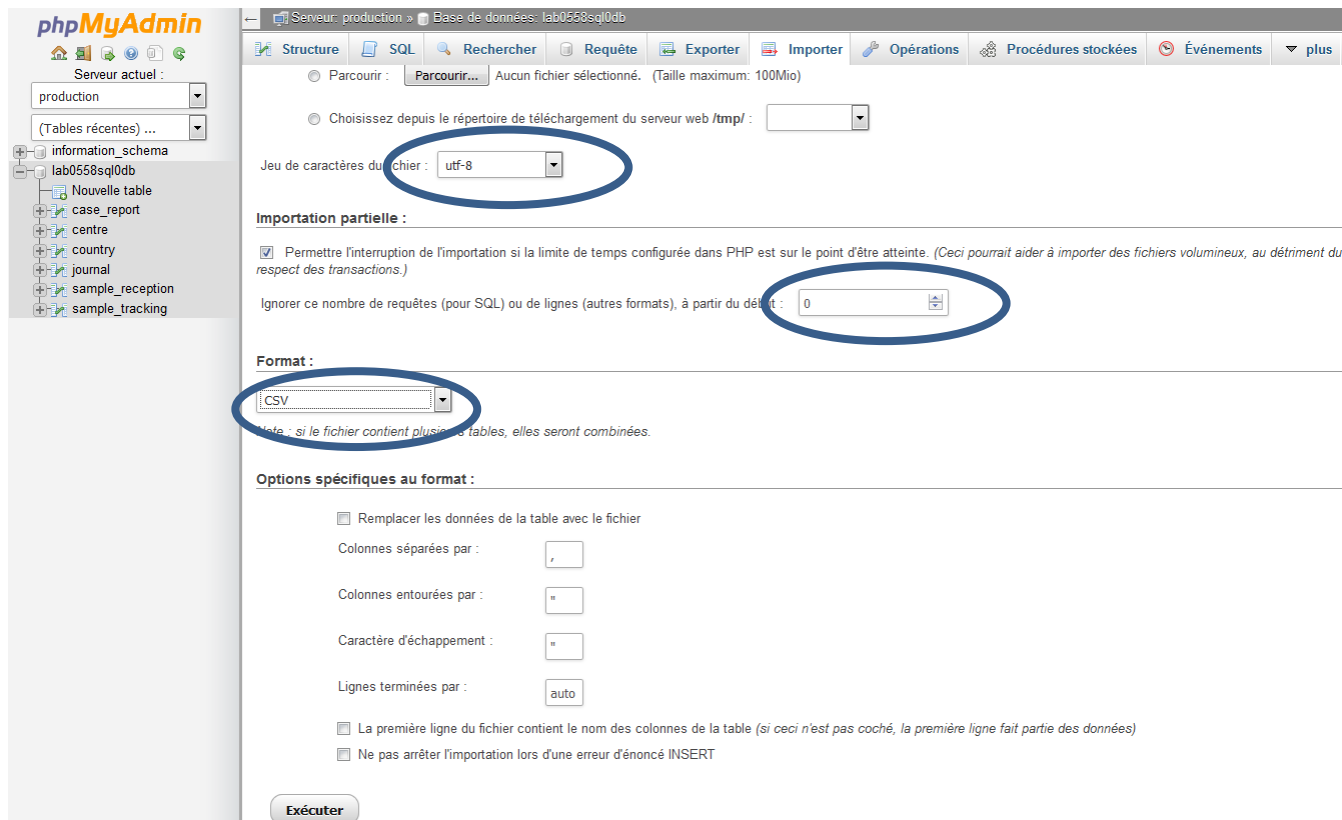
Options spécifiques au format :

Mode de compatibilité SQL :

Ne pas utiliser `AUTO_INCREMENT` pour la valeur zéro

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

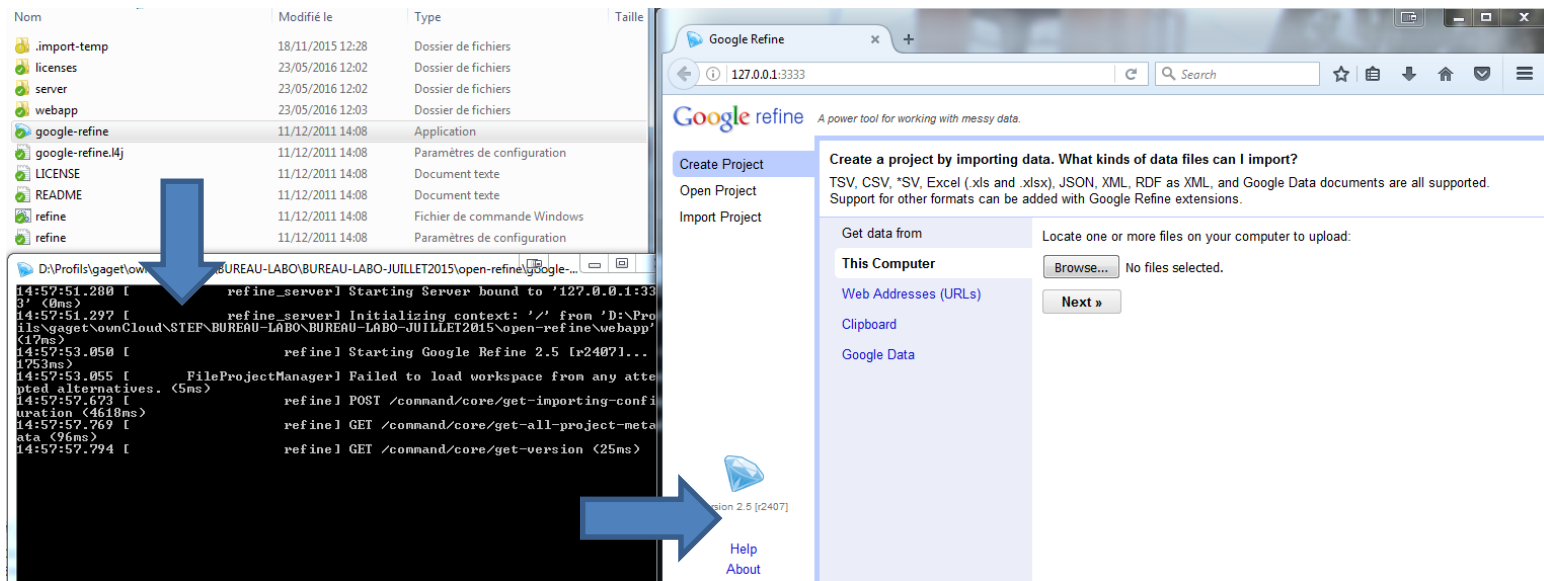
Weborama *OPEN Refine*

- <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



The image shows two screenshots illustrating the installation process. On the left, a Windows File Explorer window displays the contents of the OpenRefine installation folder, including files like .import-temp, licenses, server, webapp, google-refine, google-refine.l4j, LICENSE, README, refine, and refine. A blue arrow points from the 'google-refine' folder to a terminal window below it. The terminal shows the command 'refine_server.bat' being executed, with output indicating the server is starting on '127.0.0.1:3333'. A second blue arrow points from the terminal to a web browser window on the right. The browser shows the OpenRefine web interface at '127.0.0.1:3333', displaying the 'Create a project by importing data' page with options for 'This Computer', 'Web Addresses (URLs)', 'Clipboard', and 'Google Data'.

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	6.7		MODY/IG
22	21	28	115	women	14.605		13	6.8	Aucun	MODY/IG
23	22	28	117	Woman	17.746		7	6.5	Regime	MODY/IFG
24	23	28	118	Man	19.195		25	7.1	Regime	MODY/IFG
25	24	28	419	F	26.73		57	6.4	Bs	MODY
26	25	28	1147	Male	22.275		33	6.8	Regime	MODY/IG
27	26	28	1304	Men	20.196		33	7.4	Aucun	MODY/IFG
28	27	28	1324	men	22.34		34	6.1	Aucun	MODY/IFG
29	28	28	1917	female	21.5		32	7.2	Aucun	MODY/IFG
30	29	28	2062	Man	22.863		40	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, ...

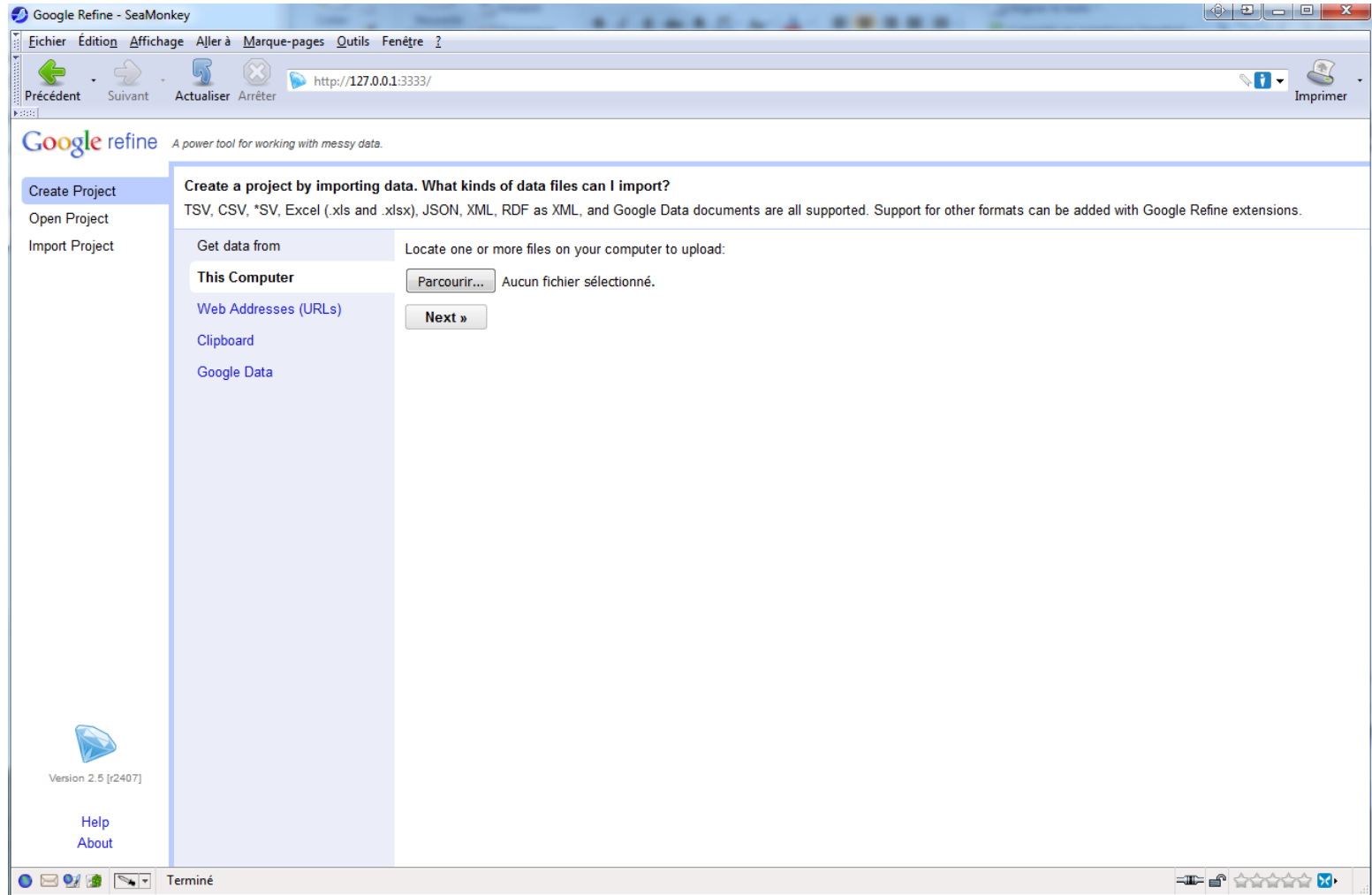
- Solution TGV (Très Geek et Vieux !) :

```
/ [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

The screenshot shows the Google Refine web interface. At the top, there's a navigation bar with 'Fichier', 'Édition', 'Affichage', 'Aller à', 'Marque-pages', 'Outils', and 'Fenêtre ?'. Below that, a browser address bar shows 'http://127.0.0.1:3333/'. The main content area features the 'Google refine' logo and the tagline 'A power tool for working with messy data.' On the left, there are buttons for 'Create Project', 'Open Project', and 'Import Project'. The central part of the interface displays a data table with columns: id_pat, id_fam, dna, sex, bmi, age_diab, age_exam, gly0, tt_t_diab, and statut_diab. The table contains 17 rows of data. Below the table, there are sections for 'Parse data as' (with options like CSV/TSV, JSON, etc.), 'Character encoding', and 'Configure Parsing Options' (including checkboxes for 'Ignore first', 'Parse next', 'Discard initial', 'Load at most', 'Parse cell text into numbers, dates, ...', 'Quotation marks are used to enclose cells containing column separators', 'Store blank rows', 'Store blank cells as nulls', and 'Store file source in each row'). A 'Create Project' button is visible on the right side of the interface.

	id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	tt_t_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

Apparence

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 Permalink Open... Export Help

Facet / Filter Undo / Redo

381 rows Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 25 next > last »

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	Aucun	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	Aucun	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

The screenshot displays the OpenRefine web interface. The browser address bar shows the URL `http://127.0.0.1:3333/project?project=1957739162473`. The page title is "test data MODYPSC patient openrefine csv". The interface includes a navigation bar with "Précédent", "Suivant", "Actualiser", and "Arrêter" buttons. A sidebar on the left contains a "Facet / Filter" section with a "Using facets and filters" tip. The main area shows a table with 381 rows. A context menu is open over the 'sex' column, listing actions like "Facet", "Text filter", "Edit cells", "Edit column", "Transpose", "Sort...", "View", and "Reconcile".

	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 »

The screenshot shows the OpenRefine web interface. The browser address bar displays `http://127.0.0.1:3333/project?project=1957739162473`. The main title is "test data MODYPSC patient openrefine csv". The interface includes a menu bar (Fichier, Édition, Affichage, Aller à, Marque-pages, Outils, Fenêtre ?), navigation buttons (Précédent, Suivant, Actualiser, Arrêter), and a toolbar (Imprimer, Open..., Export, Help). The "Facet / Filter" panel on the left shows a facet for "sex" with 14 choices, sorted by name count. The main table displays 381 rows of data with columns: id_pat, id_fam, dna, sex, bmi, age_diab, age_exam, gly0, ttt_diab, and statut_diab. The table is sorted by id_pat in ascending order.

	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)

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 Permalink Open... Export Help

Facet / Filter Undo

Refresh

sex

14 choices Sort by: name

f 31
F 27
female 22
Female 23
M 20
m 24
Male 30
Man 52
Men 29
men 19
w 18
Woman 35
Women 25
women 26

Facet by choice counts

Cluster & Edit column "sex"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision Keying Function fingerprint 5 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	45	<ul style="list-style-type: none">Female (23 rows)female (22 rows)	<input type="checkbox"/>	Female
2	58	<ul style="list-style-type: none">f (31 rows)F (27 rows)	<input type="checkbox"/>	f
2	51	<ul style="list-style-type: none">women (26 rows)Women (25 rows)	<input type="checkbox"/>	women
2	48	<ul style="list-style-type: none">Men (29 rows)men (19 rows)	<input type="checkbox"/>	Men
2	44	<ul style="list-style-type: none">m (24 rows)M (20 rows)	<input type="checkbox"/>	m

Rows in Cluster

Average Length of Choices

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Terminé

Exporter

The screenshot shows the OpenRefine web interface. The browser address bar displays `http://127.0.0.1:3333/project?project=1957739162473`. The main content area shows a table with 143 matching rows. The table columns are: `id_pat`, `id_fam`, `dna`, `sex`, `bmi`, `age_diab`, `age_exam`, `gly0`, and `ttt_diab`. The table data is as follows:

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

An 'Export project' menu is open, showing options: Tab-separated value, Comma-separated value, HTML table, Excel, ODF spreadsheet, Triple loader, MQLWrite, Custom tabular exporter..., and Templating... The 'Export' button in the top right is highlighted.

Le template

The screenshot shows the Google Refine web interface with a 'Templating Export' dialog box open. The dialog box contains the following fields and content:

- 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"].value)}}
}
```
- Row Separator:** ,
- Suffix:**] }

The preview window on the right shows the resulting JSON output:

```
{
  "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,
      "gly0" : 6.1
    }
  ]
}
```

Buttons at the bottom of the dialog include 'Reset Template', 'Export', and 'Cancel'.

Le code SQL

The screenshot shows the Google Refine interface with the 'Templating Export' dialog box open. The dialog box is titled 'Templating Export' and contains the following fields and content:

- Prefix:** A text box containing the SQL statement: `INSERT INTO ma_table VALUES`
- Row Template:** A text box containing a JSON template for each row: `{ {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:** A text box containing a comma: `,`
- Suffix:** A text box containing a semicolon: `;`
- Reset Template:** A button at the bottom left of the dialog.
- Export/Cancel:** Buttons at the bottom right of the dialog.

The right side of the dialog box shows the generated SQL code, which is a series of `INSERT INTO ma_table VALUES` statements for each row of the data. The data is as follows:

id_pat	id_fam	dna	sex	bmi	age_diab	age_exam	gly0	ttt_diab	statut_diab
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	Regime	

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