

APTEEUS

Table des matières

Description de l'activité de l'entreprise	2
Etude générale :	3
Matériel utilisé.....	3
<i>Les valisettes:</i>	3
<i>Les plaques:</i>	3
<i>Les molécules</i>	6
<i>Les machines</i>	6
Processus d'activité de l'entreprise.....	7
Comparaison entre les deux processus d'analyse.....	9
Description de la structure de la base de données proposée :	14
Schéma général de base de données proposé.....	16
Eléments à prendre en compte.....	20
<i>Normalisations des fichiers</i>	20
<i>Evolutivité du programme</i>	20
Etude détaillée	22
Définition du screening :	22
Situation actuelle	23
Besoins	26
Propositions.....	26
Base de données proposée :	Erreur ! Signet non défini.
Structure de base de données proposée pour le screening	Erreur ! Signet non défini.
Mise en pratique	27
<i>Le processus de traitement</i> :.....	27
<i>Les outils:</i>	30

Description de l'activité de l'entreprise

APTEUS développe des traitements pour les patients atteints de maladies orphelines, elle effectue des tests sur des cellules de personnes malades avec des molécules de médicaments.

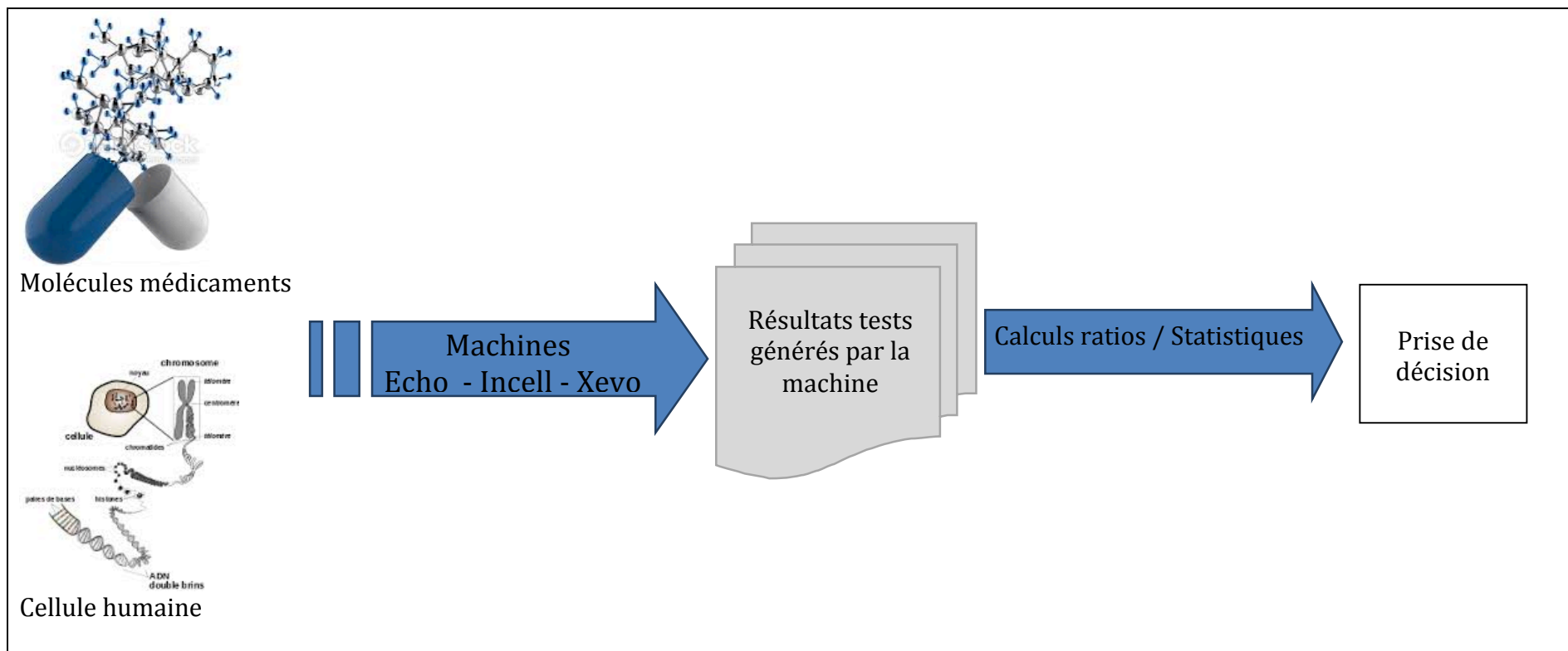


Fig. Description générale de l'activité

Les expériences réalisées par Apteeus se font en deux étapes principales :

- Le screening : tests sur des cellules avec toutes les molécules.
- DRC hit : Prendre les résultats positifs du screening et les tester à différentes concentrations de molécules.

Ces deux tests se font via trois machines : Echo puis Incell et Xevo.

Dans un premier temps Apteeus nous demande d'automatiser la partie screening, mais il faut également envisager l'intégration de la partie DRC dans notre modèle.

Etude générale :

Avant de décrire le processus d'activité de l'entreprise, Nous allons clarifier quelques termes techniques, notamment le matériel utilisé.

Matériel utilisé

Les valisettes:

Ce sont des boîtes contenant chacune des tubes de molécules en poudre

L'ensemble des valisettes constitue la chimiothèque (toutes les molécules à disposition utilisées dans le test du screening)

Une valisette est dite complète si elle contient 80 tubes (76 tubes molécules + 4 tubes vides), dans ce cas elle peut entrer dans le processus de traitement.

Les plaques:

Il y a 2 types de plaques : Plaques mères 96 pour stocker les solutions de molécules

Plaques 384 pour les tests machine

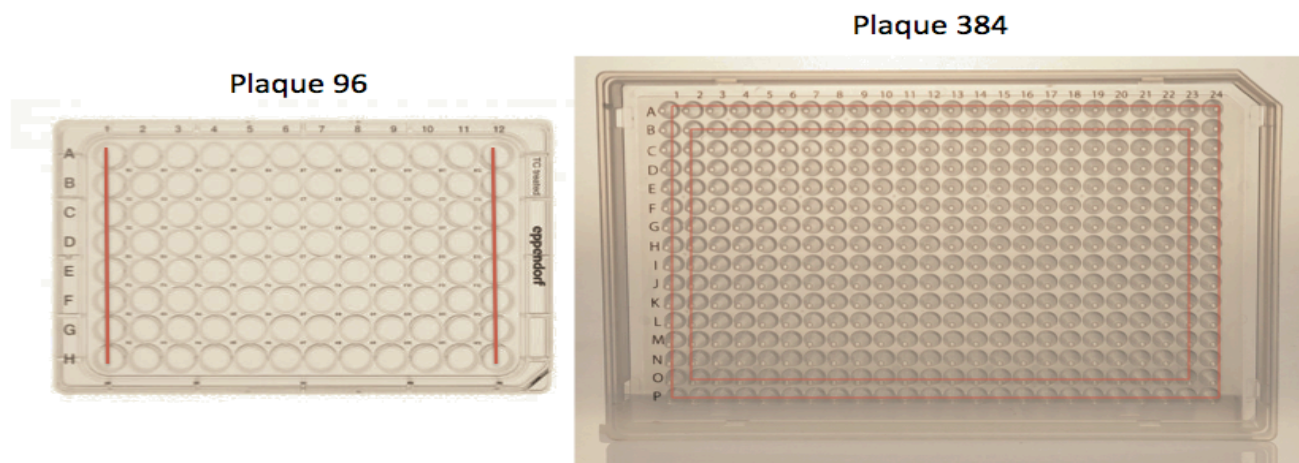
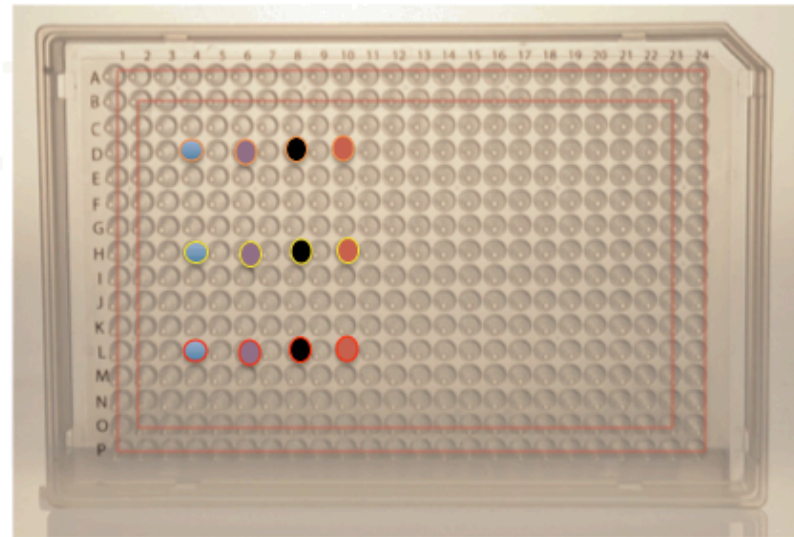
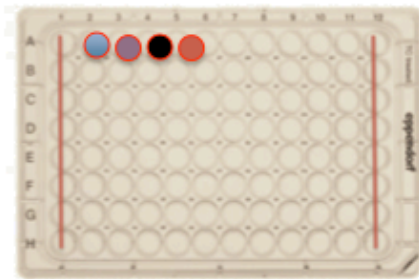
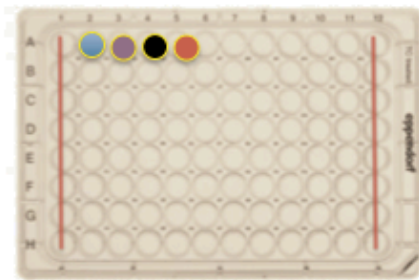


Fig. Les 2 types de plaques

- **Les plaques mères (96 puits) :**
Elles possèdent les mêmes positions que les valisettes, chaque puits contient une solution (poudre molécule + solvant)
Le DMSO : une solution (neutre) permet des tests de contrôle, ses quatre positions dans la plaque mère correspondent aux tubes vides dans les valisettes.
- **Les plaques machines (384 puits) :**
Les plaques 384 servent aux expérimentations dans les trois machines. Une plaque 384 est remplie par trois plaques 96.

Ces plaques sont en fait utilisées au format 80 pour la plaque 96 et au format 240 pour les plaques 384, ceci pour des raisons expérimentales, en effet lors d'une expérience l'analyse des éléments se trouvant au bord des plaques peut se montrer faussée par l'intervention d'éléments extérieurs comme la température.

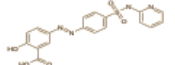


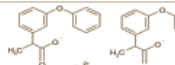
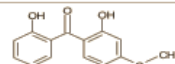
Traduction plaques 96 en plaque 384



Les molécules

L'entreprise possède un fichier recensant des molécules (PRCD), ce fichier nous donne des informations sur la structure, l'état, la disponibilité de la molécule mais aussi sa place et la quantité si celle-ci est disponible.

Dans ce fichier 3500 molécules sont recensées et plus de 1500 sont stockées physiquement dans des valisettes et 2500 sont désirées (les molécules désirées sont celles qui ont une place prédéfinie dans des valisettes).

ID	Structure	ID MotherPlate	PlatePosition	Desired	Mass mg
1		PR019	B02	Yes	1.02
2		PR001	B02	Yes	0.93
3		PR001	C02	Yes	1.05
4		PR001	D02	Yes	1.04
5		PR001	E02	Yes	1.06

Eléments importants du fichier PRCD

Les machines

Dans son processus de test l'entreprise utilise trois machines qui génèrent trois dossiers d'informations.

Echo : Au niveau de cette machine que commence l'association molécules/cellules dans des plaques Echo 384.

Incell : Un microscope qui donne des informations sur les cellules (taille, nombre...), retranscrit sous forme d'un fichier .txt ou .excel.

Xevo : Spectromètre de masse qui retourne les différentes valeurs des analytes observés dans le processus screening et DRC.

Processus d'activité de l'entreprise

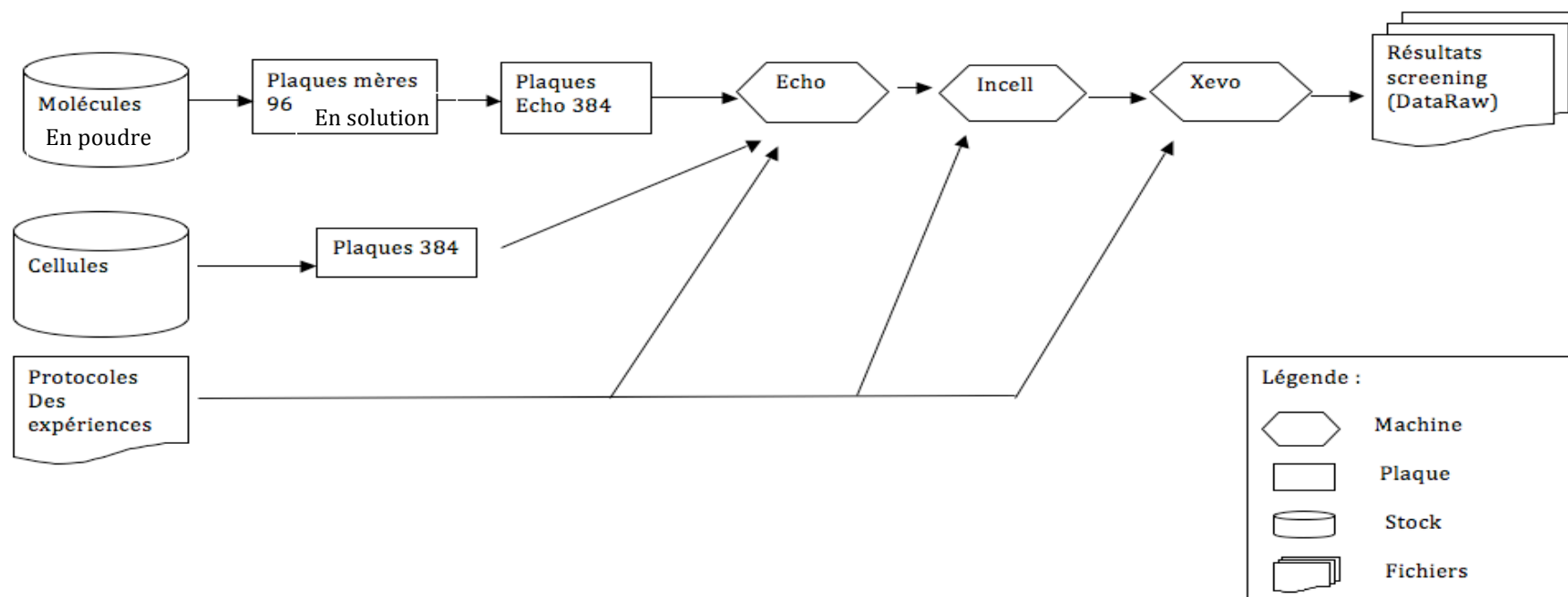


Fig1. Processus Screening

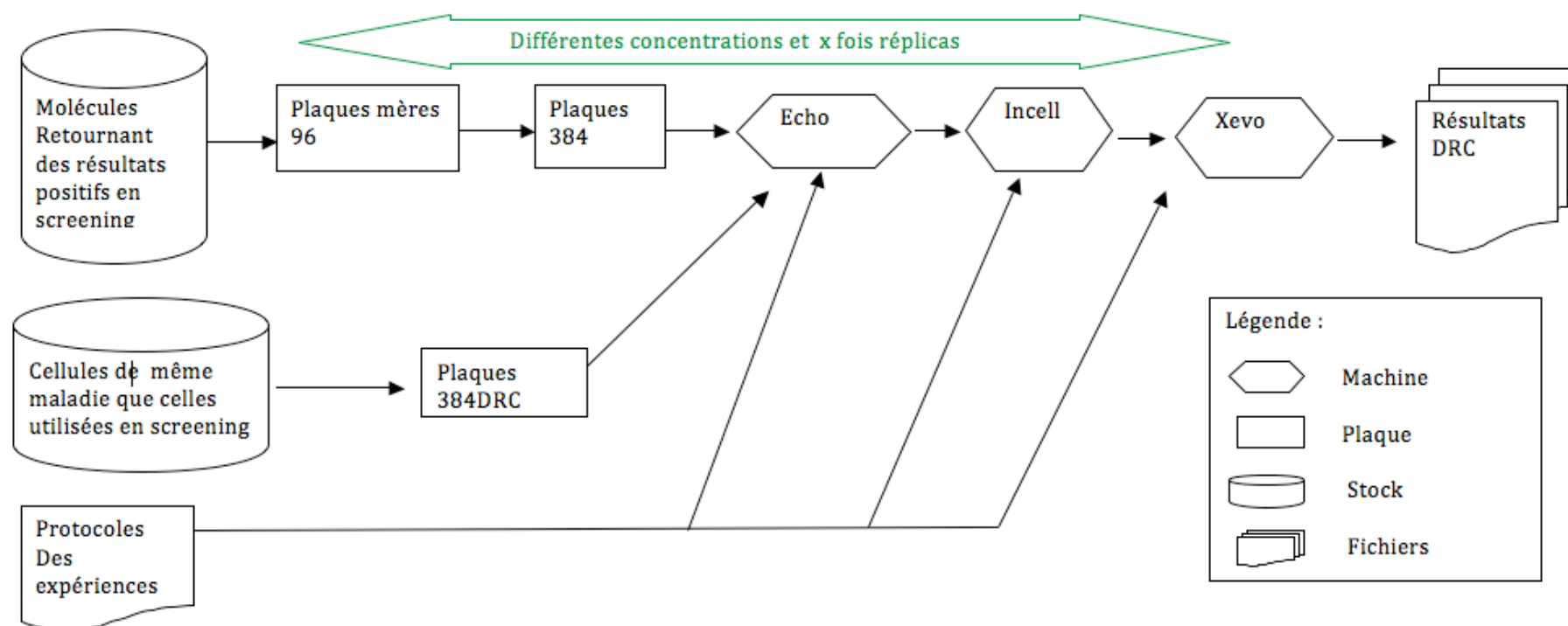


Fig2. Processus DRC Hit

Comparaison entre les deux processus d'analyse

Le but de cette comparaison est savoir si le parseur, qu'on créera pour la partie screening, pourra être adapté, plus tard, à la partie DRC puisqu'ils utilisent les mêmes machines et génèrent les mêmes types de fichiers.

Donc on commence par une description des deux processus puis une comparaison.

Le test screening se fait sur un type de cellule par toutes les molécules en disposition, le but est de déterminer quelles molécules réagissent le plus avec une certaine maladie ; ce sont les résultats dites positifs. A partir de ces résultats, on passe au test DRC-Hit ; il s'agit d'une deuxième étape où on teste des nouvelles cellules de même type (maladie) que les cellules utilisées en screening mais cette fois-ci on n'utilise que les molécules ayant donné des résultats positifs à différentes concentrations et ce test peut se répéter, de 1 à n fois, jusqu'à avoir des résultats suffisants pour les statistiques.

Points communs	Différences
<ul style="list-style-type: none">- Les 2 tests passent par les machines Echo puis Incell puis Xevo	<ul style="list-style-type: none">- Screening utilise toutes les molécules à disposition- DRC utilise un certain nombre de molécules à différentes concentrations
<ul style="list-style-type: none">- Les fichiers générés par les différentes machines ont la même structure	<ul style="list-style-type: none">- La différence est au niveau de l'interprétation des positions des plaques (puits) Screening : chaque puits représente une molécule différente DRC : chaque puits représente une molécule + concentration différentes

Les deux figures suivantes représentent deux captures des fichiers générés par la machine Xevo suite à test Screening et à un test DRC

Quantify Compound Summary Report

Printed Thu Nov 05 11:16:57 2015

Compound 1: C22estertriMeAmonEthan

	#	Name	RT	Area	Response	S/N	Vial	
1	1	10-12 MAP TEE2 E1			0.264549.567	4549.567	395.103	3:C,3
2	2	10-12 MAP TEE2 E2			0.26 5702.965		5702.965	510.454 3:C,4
3	3	10-12 MAP TEE2 E3			0.26 4549.119		4549.119	744.429 3:C,5
4	4	10-12 MAP TEE2 E4			0.26 6438.124		6438.124	331.129 3:C,6
5	5	10-12 MAP TEE2 E5			0.26 5530.448		5530.448	369.552 3:C,7
6	6	10-12 MAP TEE2 E6			0.26 6507.181		6507.181	305.736 3:C,8
7	7	10-12 MAP TEE2 E7			0.26 4743.260		4743.260	373.990 3:C,9
8	8	10-12 MAP TEE2 E8			0.26 5855.409		5855.409	489.165 3:C,10
9	9	10-12 MAP TEE2 E9			0.26 5272.920		5272.920	437.241 3:C,11
10	10	10-12 MAP TEE2 E10			0.26 5350.667		5350.667	231.048 3:C,12
11	11	10-12 MAP TEE2 E11			0.26 5035.833		5035.833	587.543 3:C,13
12	12	10-12 MAP TEE2 E12			0.26 3472.300		3472.300	215.623 3:C,14
13	13	10-12 MAP TEE2 E13			0.26 4106.547		4106.547	309.635 3:C,15
14	14	10-12 MAP TEE2 E14			0.26 5282.983		5282.983	357.451 3:C,16
15	15	10-12 MAP TEE2 E15			0.26 4899.554		4899.554	601.933 3:C,17

Fig. Capture d'une partie d'un fichier Screening généré par la machine Xevo

Les données concernant les molécules :

- Chaque fichier représente les résultats d'une plaque définie par son numéro.
Exemple : Ce fichier est le résultat de la plaque numéro 2, TEE'2'= plaque numéro2
- Chaque ligne d'un fichier représente un puits de la plaque 384 ; chaque puits est défini par sa position (la dernière colonne).
Exemple : C, 4= ligne C colonne 4
- ⇒ A partir des ces deux données on peut déterminer le nom de la molécule utilisé dans chaque puits.

Les données concernant les cellules :

Dans la base des données actuelle :

- Une cellule possède un identifiant composé de 3 lettres et 3 chiffres (exemple MAP003).

Exemple : Ici, on voit les 3 lettres MAP mais pas les 3 chiffres car tout est entré à la main au niveau de la machine Xevo ; le nom de la cellule, le nom des métabolites, le numéro de l'expérience...

De plus, la table des cellules est encore primaire, pour le moment, pour les cellules MAP il n'y a que MAP007, il n'y en pas d'autres qui commencent par MAP comme MAP001 ou MAP002 c'est pour ça qu'on a tapé MAP tout court.

Il y a un certaines cellules qui commencent par GM puis une suite de chiffres (GM14476, GM07469, GM04488A...) dans ce cas il faut écrire tous l'identifiant.

Généralement, le numéro de l'expérience indique le numéro de la maladie

Exemple : 10-12 ; ici on peut savoir qu'il s'agit de la maladie numéro 10= « G-CPT2 » mais ce n'est pas toujours vrai car il y a des expériences, comme l'expérience numéro 5 faites sur des cellules saines donc le numéro 5 ne signifie pas la maladie numéro 5 « Lung Carcinoma ».

Les données concernant les analytes :

Ce fichier contient aussi les résultats du test ; un certain nombre d'analytes avec des valeurs associées pour chaque activité (Area, RT, S/N...).

Pour chaque analyte (compound) il y a 240 lignes au nombre des puits (dans cette capture on voit que 15 lignes du compound numéro1)

Le nombre et le type des analytes varient selon la maladie (à paramétrer dans la machine avant de démarrer le test).

A chaque expérience, on peut trouver des analytes déjà observées, comme on peut tirer de nouveaux analytes.

⇒ A partir de ces résultats on tire les données qui vont servir à l'étape suivante ; Le DRC-Hit.

Quantify Compound Summary Report

Printed Wed Mar 16 09:18:37 2016

Compound 1: carnitine

	#	Name	RT	Area	Response	S/N	Vial	
1	1	11-08-GAU1-DRC-1			0.19	8813.329	8813.329	1723.417
2	2	11-08-GAU1-DRC-2			0.19	1973.505	1973.505	954.150 4:C,4
3	3	11-08-GAU1-DRC-3			0.19	6363.998	6363.998	11519.537
4	4	11-08-GAU1-DRC-4			0.19	5799.194	5799.194	784.784 4:C,6
5	5	11-08-GAU1-DRC-5			0.19	11799.691	11799.691	514.134 4:C,7
6	6	11-08-GAU1-DRC-6			0.19	644.477 644.477	58.073	4:C,8
7	7	11-08-GAU1-DRC-7			0.19	4327.037	4327.037	7098.559
8	8	11-08-GAU1-DRC-8			0.19	9903.825	9903.825	995.115 4:C,10
9	9	11-08-GAU1-DRC-9			0.19	11392.818	11392.818	2521.742
10	10	11-08-GAU1-DRC-10			0.19	10886.478	10886.478	2555.549
11	11	11-08-GAU1-DRC-11			0.19	1618.677	1618.677	344.356 4:C,13
12	12	11-08-GAU1-DRC-12			0.19	280.988 280.988	441.029	4:C,14
13	13	11-08-GAU1-DRC-13			0.19	8518.062	8518.062	11820.935
14	14	11-08-GAU1-DRC-14			0.19	5465.101	5465.101	775.252 4:C,16
15	15	11-08-GAU1-DRC-15			0.19	10788.033	10788.033	1730.276

Fig. Capture d'une partie d'un fichier DRC généré par la machine Xevo

Les données concernant les molécules et les concentrations :

- Chaque position (la dernière colonne) correspond à une molécule à un certain niveau de concentration.
- On trouve les informations nécessaires dans un fichier Excel appelé « Template DRC » ; Ce dernier est rempli manuellement suite aux calculs des ratios du premier test (les résultats positifs) sous forme de tableau de correspondance entre position plaque et molécule/concentration.

Exemple : Dans le fichier « TemplateDRC », on aura un index incrémental comme la colonne tout à gauche, une deuxième colonne contenant le nom de la molécule et une troisième colonne contenant le niveau de la concentration.

Les données concernant les cellules :

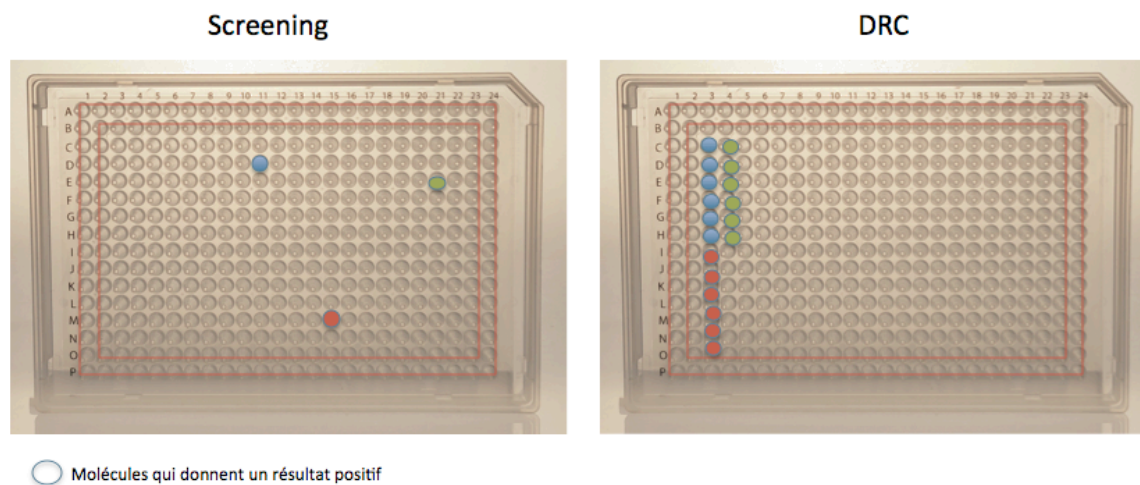
- Comme dans le fichier précédent.

Les données concernant le numéro réplicas :

- C'est le chiffre qui suit le nom de la cellule.
Exemple : GAU1; GAU est l'identifiant de la cellule et 1 est le numéro du réplica

Les données concernant les analytes :

Tout comme le fichier Xevo du Screening, ce fichier contient aussi les résultats du test ; analytes et valeurs associées (Area, RT, S/N...). Les analytes observés suite au test DRC peuvent être différentes des analytes observés au niveau du test screening.



La DRC nous pose un problème pour retrouver les molécules testées, en effet pour le screening on peut retrouver facilement la molécule qui a réagit positivement grâce à sa position, mais dans le cas de la DRC les positions ne sont plus liées directement à une molécule dans la valisette de départ mais à une concentration et une molécule sauvegardées dans le fichier « TemplateDRC ».

La question qui se pose : Si les deux processus passent par les mêmes machines et génèrent des fichiers de même structure, Peut-on stocker les résultats dans une même table dans la base de données ?

Description de la structure de la base de données proposée :

- Echo, Incell et Xevo donnent des résultats de structures similaires pour le screening et le DRC sauf que les positions des puits ont des significations différentes ; Il ne faut pas confondre position plaque screening et position plaque DRC.
- En effet, il n'y a pas un lien direct entre les deux tables Screening et DRC dans notre base de données car nous n'avons pas d'informations sur les résultats positifs ; ils ne peuvent être déterminés que suite au calcul des ratios et des statistiques.
- Une fois les résultats positifs obtenus, il serait possible de créer la table Template DRC qui servira à déterminer les positions de molécules ainsi que leurs concentrations dans les plaques DRC (les informations de cette table ne peuvent pas être générées automatiquement)
- Pour pouvoir déterminer le nom de la molécule correspondant à chaque résultat (à chaque puits plaque/ligne fichier), Nous aurons besoin d'une table molécule qui sera liée à la table Screening directement et à la table DRC en passant par la table Template DRC.
- Le DRC possède deux propriétés différentes du screening : La concentration et le réplica.
- Pour garder des informations sur la date et les conditions de l'expérience nous avons décidé de créer une table Expérience qui sera liée aux deux tables Screening et DRC par le numéro de l'expérience et permettra plus tard des recherches par type d'expérience.
- Le calcul des ratios nécessite des informations sur les cellules qu'on peut tirer à partir des fichiers Incell, donc nous avons intérêt à créer une table Incell liée aux deux tables principales (DRC et Screening)
- L'entreprise possède déjà une base de données contenant :
 - Des informations concernant les maladies : Nous nous intéressons qu'au numéro de la maladie
 - Les protocoles des expériences : Table analyte contenant des informations permettant de paramétrer la machine Xevo selon le type de la maladie ; Nous l'avons pas insérer dans notre modèle car cette étape se fait avant même de commencer le test .
 - L'état du stock des cellules et les fichiers Echo : ils serviront plus tard pour automatiser un processus indépendant du screening permettant le contrôle des valisettes (volume, copie...)

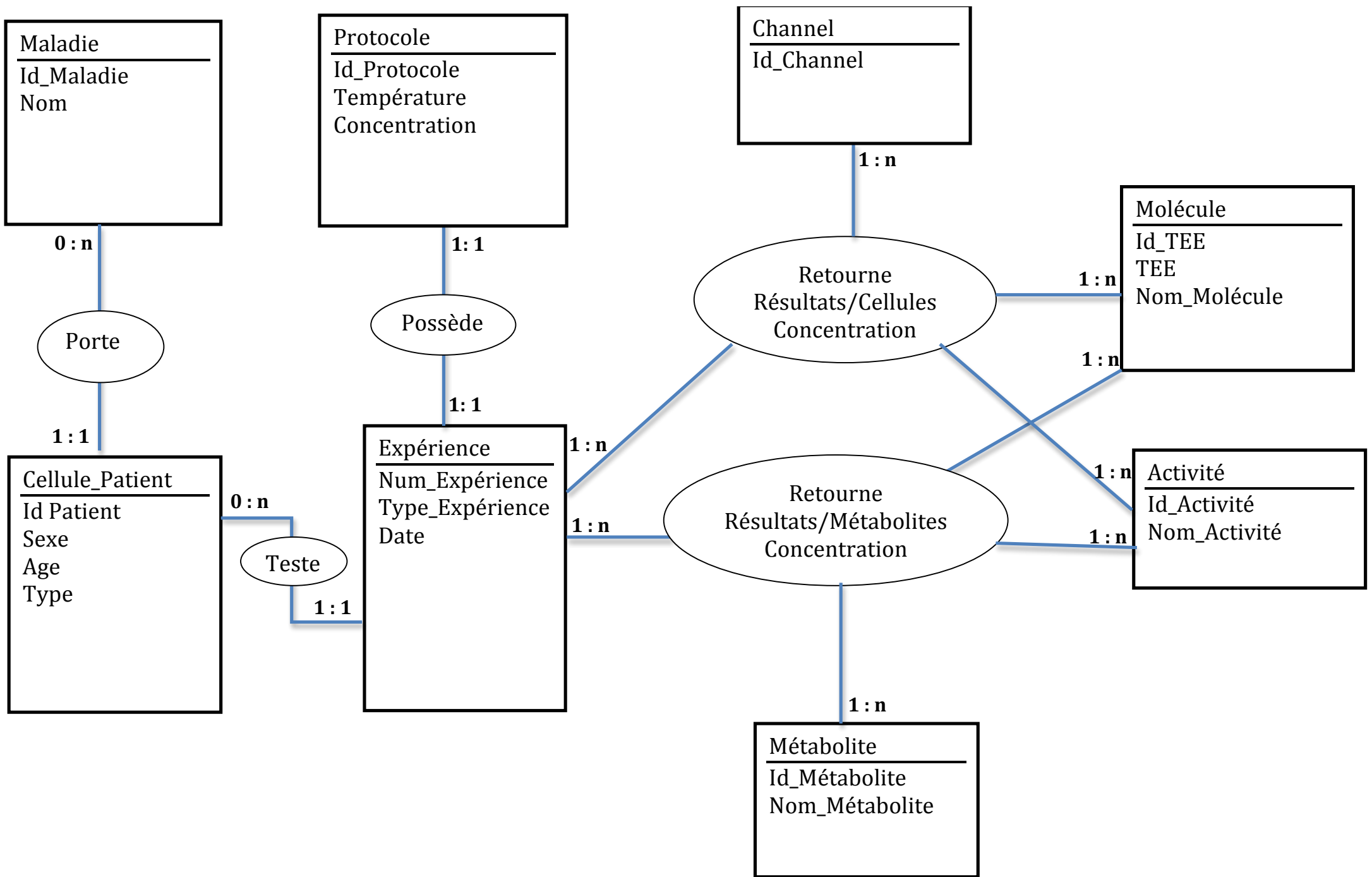
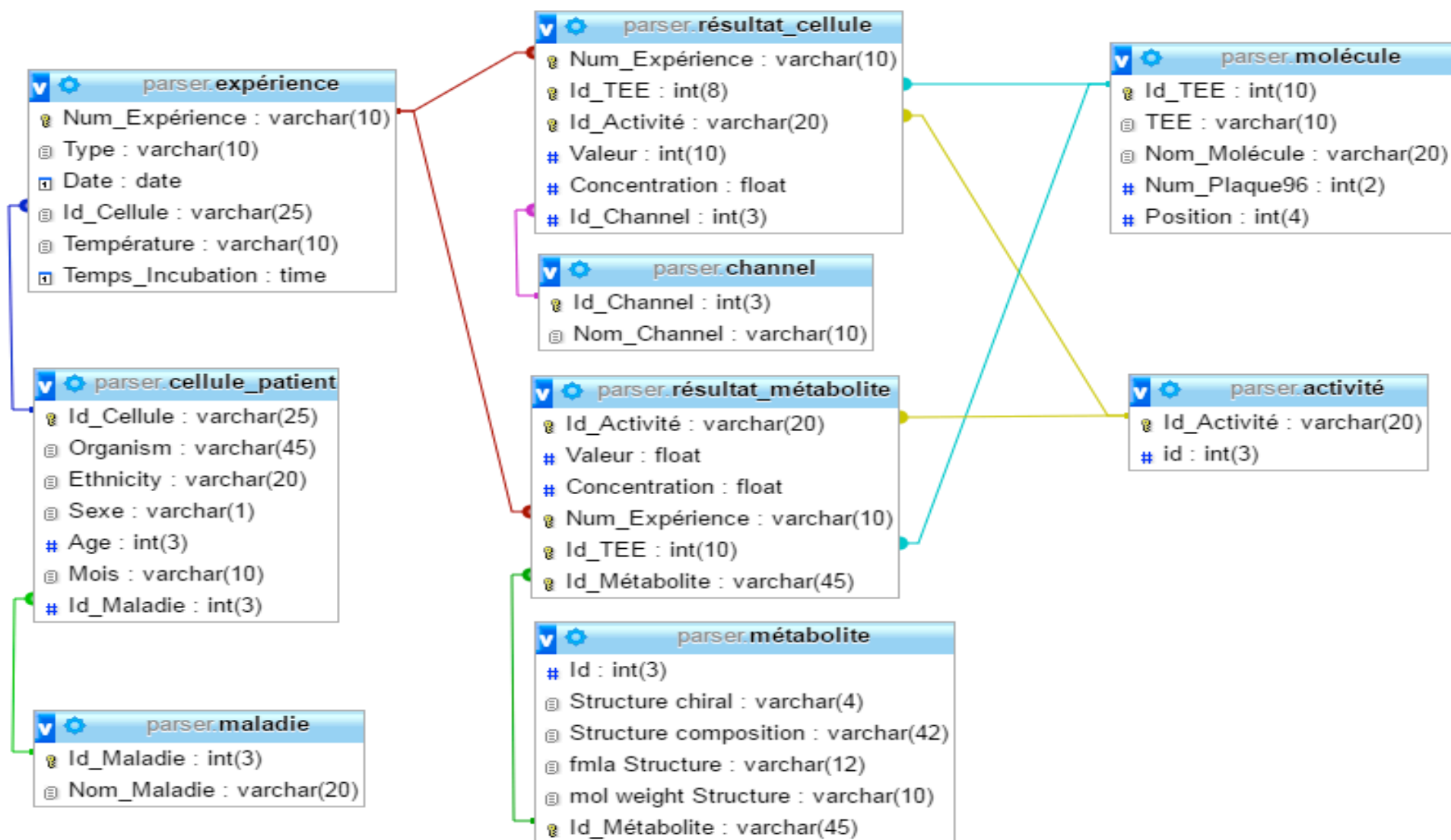


Fig. Modèle Entité/Association

Schéma général de base de données proposé



Dictionnaire des données:

Expérience

Code	Type	Sémantique	Valeurs	Remarque
Num_Expérience	Numérique	Numéro de l'expérience	Ex: 10-12	Clé primaire
Type	Chaine de caractères	Type de l'expérience	DRC ou Screening	
Date	Date		JJ/MM/AAAA	
Id_Cellule	Numérique	Identifiant Cellule		Clé étrangère
Id_Protocole	Numérique	Identifiant Protocole		Clé étrangère

Resultat_Cellule

Code	Type	Sémantique	Valeurs	Remarque
Num_Expérience	Numérique	Numéro de l'expérience	Ex: 10-12	Clé primaire
Id_TEE	Numérique	Identifiant du TEE	Incremental	Clé primaire
Id_Activité	Numérique	Identifiant Activité étudiée		Clé primaire
Valeur	Numérique	Valeur de l'activité étudiée		Form_Factor : 1 rond 0 pas rond Cv hoechst_mean: faible: bien, élevé: toxicité du produit
Concentration	Numérique	Concentration molécule dans la solution	6 valeurs (30, 10, 3, 1,0.3, 0.1)	
Replica	Numérique		0,1,2,3..	Dans le cas du screening on met 0. 1,2,3... dans le cas DRC

Résultat Métabolite

Code	Type	Sémantique	Valeurs	Remarque
Num_Expérience	Numérique	Numéro de l'expérience	Ex: 10-12	Clé primaire
Id_TEE	Numérique	Identifiant du TEE	Incremental	Clé primaire
Id_Métabolite	Chaine de caractère	Nom de la métabolite	C16,C22...	Clé primaire
Id_Activité	Numérique	Identifiant Activité étudiée		Clé primaire
Valeur	Numérique	Valeur de l'activité étudiée		
Concentration	Numérique	Concentration molécule dans la solution	6 valeurs (30, 10, 3, 1,0.3, 0.1)	

Molécule

Code	Type	Sémantique	Valeurs	Remarque
Id_TEE	Numérique	Identifiant du TEE	Incremental	
TEE	Chaine de caractères	Identifiant standard de molécule	TEE+4 chiffres	
Nom_Molécule	Chaine de caractères	Nom de la molécule		
Num_Plaque96	Numérique	Numéro de la plaque 96		
Position	Chaine de caractères	Position dans la plaque 96		

Cellule Patient

Code	Type	Sémantique	Valeurs	Remarque
Id_Patient	Numérique	Identifiant du Patient	MAP...	
Sexe	Chaine de caractères	Sexe du Patient	0 ou 1	
Age	Numérique	Age du Patient		
Type	Chaine de caractères	Type de Cellule		

Id_Maladie	Numérique	Identifiant de la maladie		Clé étrangère
------------	-----------	---------------------------	--	---------------

Maladie

Code	Type	Sémantique	Valeurs	Remarque
Id_Maladie	Numérique	Identifiant de la maladie		Clé Primaire
Nom_Maladie	Chaine de caractères	Nom de la maladie		

Protocole

Code	Type	Sémantique	Valeurs	Remarque
Id_Protocole	Numérique	Identifiant du protocole		Clé primaire
Température	Numérique	Température en degrés	X°C	
Concentration	Numérique	Concentration molécule dans la solution	6 valeurs (30, 10, 3, 1,0.3, 0.1)	

Métabolite

Code	Type	Sémantique	Valeurs	Remarque
Id_Métabolite	Chaine de caractères	Identifiant du métabolite	C22,C24...	Clé primaire
Nom	Chaine de caractères	Nom du métabolite		

Activité

Id_Activité	Numérique	Identifiant Activité étudiée		Clé primaire
-------------	-----------	------------------------------	--	--------------

Activité	Chaine de caractère	Activité étudiée	Area_Mean, Count, Form_Factor, Cv_Hochest_Mean, RT, S/N, Area...	Aire calculé des noyaux des cellules : Area_Mean Nombre des cellules : Count Forme du noyau : Form_Factor Colorant des noyaux coefficient de variation de l'intensité des noyaux : Cv_Hochest_Mean
----------	---------------------	------------------	------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Éléments à prendre en compte

Normalisations des fichiers

Dans le fichier Analyte on dispose de la liste des composants et leurs propriétés, mais les noms ne sont pas les mêmes que ceux recensés dans le fichier du screening,

Exemple : C22estertriMeAmonEthan dans le screening.

C22 dimethylaminoethanol ester ou C22 Behenoylcarnitine butylé semblent correspondre dans le fichier Analyte.

Evolutivité du programme

En se développant, l'entreprise est amenée à tester de plus en plus de molécules, ce nombre augmente de 5 à 10 par mois. L'insertion des nouvelles molécules dans la base de données se fait tous les 6 mois, à peu près, car une valisette n'entre en processus que lorsqu'elle est complète (contient 80 molécules).

Le Xevo (la machine qui réalise le screening) est quant à elle capable d'accueillir 20 plaques contenant 240 molécules. Soit environ 4700 molécules (il y a 4 tubes vides par valisette)

Notre parseur doit donc être capable de traduire entre 1 et 20 fichiers tests pour les rentrer dans une base de données et sortir un fichier Excel contenant un nombre non déterminé à l'avance de feuilles.

Le nombre d'analytes n'est pas fixe, on doit donc être capable de rajouter des colonnes dans le fichier Excel, et de faire le formatage conditionnel de ces colonnes.

Au cas ou une molécule viendrait à changer de position dans une valisette on peut imaginer une interface derrière l'interface principale qui permettrait de réaliser ce changement.

Un dictionnaire de noms pour les analytes est également nécessaire pour recenser les différents synonymes (ex : C22estertriMeAmonEthan, C22)

Dans la deuxième partie de ce document, nous nous intéresserons à l'étude détaillée du processus Screening que nous voulons automatiser.

Volumétrie

Une autre question qui se pose est de savoir l'espace que prendra notre base de données, nos préoccupations se portent particulièrement sur les tables de résultats qui sont celles qui évolueront le plus au cours du temps.

Tables à évolution lente :

Actuellement :

Molécule – 1500 lignes
Métabolite – 100 lignes
Expérience – 100 lignes
Cellule_Patient – 50 lignes
Protocole – 50 lignes
Activité – 15 lignes
Channel – 10 lignes
Maladie – 10 lignes

Tables à évolution rapide :

Actuellement :

Resultat_Cellule – 100 expériences x 240 puits x 4 activités x 2 channels x 7 plaques = 1 300 000 lignes
Resultat_Métabolite – 100 expériences x 240 puits x 4 activités x 4 métabolites x 7 plaques = 2 700 000 lignes

Le nombre d'expérience par an est d'environ 50, le nombre de puits n'évolue pas et sera donc toujours de 240, le nombre d'activités peut évoluer très légèrement et passer à 5 ou 6 pour chacune des tables de résultats, le nombre de channels et de métabolites dépend de l'expérience, sa valeur ne devrait que très peu évoluer, le nombre de plaques peut quant à lui augmenter si Apteeus obtient de plus en plus de molécules.

Ajouts par an (actuellement) :

Resultat_Cellule – 50 expériences x 240 puits x 4 activités x 2 channels x 7 plaques = 700 000 lignes

Resultat_Métabolite – 50 expériences x 240 puits x 4 activités x 4 métabolites x 7 plaques = 1 300 000 lignes

Ajouts par an (prévisionnel haut) :

Resultat_Cellule – 75 expériences x 240 puits x 6 activités x 2 channels x 13 plaques = 2 800 000 lignes

Resultat_Métabolite – 75 expériences x 240 puits x 6 activités x 4 métabolites x 13 plaques = 5 600 000 lignes

Dans 10 ans nous aurions donc une base de données composées pour ces deux tables d'environ :

Resultat_Cellule - 15 000 000 lignes

Resultat_Métabolite – 40 000 000 lignes

Etude détaillée

Définition du screening :

Cette étape consiste à soumettre des milliers de molécules à une batterie de tests systématiques, afin d'étudier leurs propriétés chimiques et pharmacologiques, elle permet de repérer celles qui pourraient avoir un intérêt thérapeutique sur des cellules malades. Toutes les molécules subissent les tests initiaux, pour effectuer un premier tri par type de maladie.

Dans un premier temps, il faut utiliser un robot pour déplacer les solutions des plaques mères aux plaques machines.

Ensuite, ces molécules doivent être mélangées avec des cellules dans des plaques 384 adaptées aux machines de tests. Cette opération se fait au niveau de la machine Echo qui retourne par la suite les informations nécessaires sur le succès ou l'échec de cette étape comme le volume déplacé...???

Puis, ces plaques sont transmises dans une deuxième machine Incell qui retournera des informations sur l'analyse microscopique des cellules comme le nombre, l'aire moyenne des cellules...

La troisième et la dernière machine à utiliser pour le screening est le Xevo ; il retourne des analytes (C1, C5, C18, C22...) et les activités associées à chacune (Area, RT, S/N...).

Les trois machines génèrent chacune un certain nombre de fichiers correspondant au nombre de plaques utilisés (actuellement 7 plaques).

C'est à partir de ces fichiers que l'entreprise peut effectuer des calculs nécessaires à la prise de décisions ; ces fichiers seront nos sources de données pour remplir les tables de la base.

⇒ A chaque expérience :

- ~240 lignes (positions)*7 fichiers (plaques) * nombre de composants dans la table Xevo.
- ~240 lignes (positions)*7 fichiers (plaques) dans la table Incell.

⇒ La table conversion a un nombre de lignes fixe : ~1500 lignes

- L'évolution de cette table n'est pas liée aux expériences mais aux molécules possédées par Apteeus.

Situation actuelle

Les fichiers initiaux ne permettent pas une bonne lisibilité ni une structure adaptable aux calculs des ratios.

Pour analyser les résultats des tests les employés d'Apteeus ont besoin de retranscrire manuellement les fichiers texte dans un fichier Excel (comme décrit dans le schéma suivant).

Les résultats de ces tests de screening sont stockés dans des dossiers et non pas dans une base de données.

Réalisation du fichier screening

XEVO

C22				
	NAME	RT	AREA	POSITION
C24				
	NAME	RT	AREA	POSITION
C26				
	NAME	RT	AREA	POSITION



Fichier Final

		C22		C24		C26			
NAME	POSITION	RT	AREA	RT	AREA	RT	AREA	COUNT	RATIOS

INCELL

POSITION	MEAN AREA	COUNT	MEAN FORM FACTOR	MEAN CV HOECHST



Besoins

- Obtenir automatiquement le fichier final au format Excel tout en rajoutant des informations complémentaires sur les molécules.
- Garder un historique des tests.

Propositions

- Réalisation d'un parseur qui, à partir des fichiers textes, insère les données dans une base de données.
- Création d'une table de correspondance plaque 384/96 pour le rajout des informations complémentaires sur les molécules (position initiale, nom de la molécule).
- Ecriture d'un programme permettant d'extraire les données nécessaires pour les exporter dans un fichier Excel avec des mises en forme conditionnelles.
- Permettre à l'utilisateur de choisir des options de traitement (calculs, graphiques...).
- Permettre à l'utilisateur de faire des recherches par critères dans la base de données (num_exp, date...)

Le processus de traitement :

- Sélection des fichiers
- Zone de recherche par critère
- Zone d'affichage avant export
- Export Excel

Chercher les fichiers dans le répertoire

ou

Requêtes SQL

Requêtes SQL+
PHPExcel

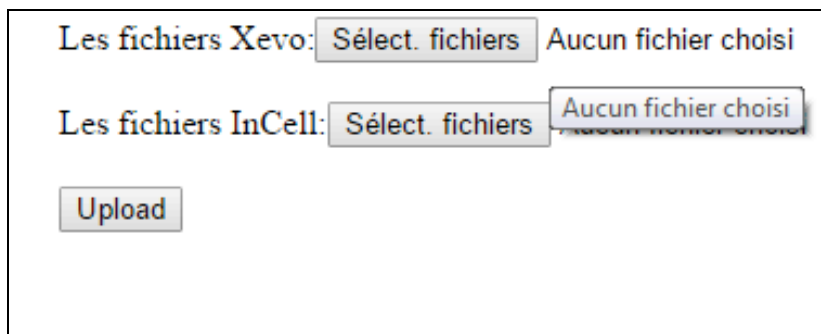
Fichiers de formatage conditionnel

Insertion dans les
fichiers .xml du fichier .xlsx

27

Le parseur :

1- Une Interface avec 2 boutons pour uploader les fichiers provenant de la machine Xevo et ceux de la machine InCell



The screenshot shows a web interface with two sections for file selection. The first section is labeled 'Les fichiers Xevo:' and contains a button 'Sélect. fichiers' and a status indicator 'Aucun fichier choisi'. The second section is labeled 'Les fichiers InCell:' and also contains a button 'Sélect. fichiers' and a status indicator 'Aucun fichier choisi'. Below these two sections is a single 'Upload' button.

-Ces fichiers seront déplacés dans deux sous répertoires sur le serveur.

-Prévoir aussi des champs de saisie permettant de taper des informations sur l'expérience tel que la température ou le type et permettant aussi des insertions dans la base de données

2- Parser les fichiers Xevo à l'aide des expressions régulières, pour optimiser le programme je crée certaines fonctions ;

- Une fonction qui stocke les noms des fichiers dans une liste et retourne le nombre des fichiers à lire

- Une fonction qui initialise l'Id_TEE de chaque fichier

- Une fonction qui lit 1 fichier ; extraire les variables pour les insérer dans la base des données.

- Créer la boucle à l'aide des fonctions précédentes pour lire tous les fichiers

- Une fois insérés, supprimer les fichiers du serveur ; Pour avoir le sous répertoire vide et ne contenant que les fichiers à insérer.

3- Parser les fichiers InCell (en cours).

Processus d'extraction

1- Dans la première étape nous allons *mettre en page le document Excel* grâce à des requêtes SQL sur la base de données, en effet grâce au :

- Nombre d'activités et de métabolites pour une expérience nous allons connaître le nombre de colonnes à créer.
- Numéro de l'expérience nous allons compléter une colonne avec le nom de l'expérience.

Puis nous allons mettre automatiquement à la main les éléments invariants de notre fichier Excel comme les positions et les Id.
Puis Extraction en fichier Excel de ce document qui servira de modèle.

Sample Number	Sample Name	Row	Col	C22 RT	C22 Area	C23 RT	C23 Area	C24 RT	C24 Area	C25 RT	C25 Area
1	10-12	C	3								
2	10-12	C	4								
3	10-12	C	5								
4	10-12	C	6								
5	10-12	C	7								
6	10-12	C	8								
7	10-12	C	9								
8	10-12	C	10								
9	10-12	C	11								
10	10-12	C	12								
11	10-12	C	13								

2- Dans une deuxième étape nous allons remplir un nouveau document Excel grâce à des requêtes SQL et grâce aux informations obtenues de l'étape précédente.

- Nombre d' Id_TEE pour une expérience pour savoir le nombre de feuilles à créer.
- Remplir les colonnes grâce aux informations sur les activités et les métabolites obtenues précédemment.

Puis on extrait de nouveau en format Excel.

Sample Number	Sample Name	Row	Col	C22 RT	C22 Area	C23 RT	C23 Area	C24 RT	C24 Area	C25 RT	C25 Area
1	10-12	C	3	300	4549,567	9900	1642,457	19500	1436,557	29100	313346,88
2	10-12	C	4	340	5702,965	9940	2331,269	19540	1839,196	29140	361487,22
3	10-12	C	5	380	4549,119	9980	1644,682	19580	1792,476	29180	321048,81
4	10-12	C	6	420	6438,124	10020	2309,041	19620	2312,422	29220	434663,09
5	10-12	C	7	460	5530,448	10060	1884,14	19660	2177,981	29260	400426,56
6	10-12	C	8	500	6507,181	10100	2363,643	19700	2404,739	29300	436540,66
7	10-12	C	9	540	4743,26	10140	1560,398	19740	1807,561	29340	331941,03
8	10-12	C	10	580	5855,409	10180	2207,062	19780	2453,446	29380	408212,59

3- Maintenant nous allons extraire notre fichier Excel comme nous le ferions pour une archive .zip pour aller chercher les fichiers .xml.

4- Dans les fichiers .xml de notre fichier Excel on va incorporer le formatage conditionnel.

5- On met nos nouveaux fichiers .xml dans le dossier, on compresse le tout, on renomme le fichier en .xlsx

Les outils:

Tâche	Description	PHP	Macros Excel	Python	Pipeline Pilot
Chargement des fichiers	Drag and drop Sélection du dossier	+	+	+	+
Parseur	Lecture des fichiers Expressions régulières	+	+ -	+	?
Insertion base des données	Requêtes SQL, Evolution de la BD	+	-	+	+
Calcul ratios	Automatique ou Sélectionné	+	+	+	+
Export Excel / Mise en forme conditionnelle	Echelle des couleurs Excel	+-	+	+	?
Interface graphique	Upload, export, zone de recherche, visualisation pré-export	+	-	+ ?	+
Recherche par critères	Recherche par num_exp, par date...	+	-	+ ?	+

PHP

- Upload (Drag and drop pour envoyer tous les fichiers d'un coup)
- Chercher les textes dans le répertoire (fenêtre de dialogue pour choisir le dossier qui contient tous les fichiers de l'expérience)
- Insertion de chaque fichier dans la BDD
- Extraction de la BDD (avec éventuellement une table temporaire)
- Export Excel

Avantages :

Facile à utiliser, interface minimaliste, langage maîtrisé, BDD facile à consulter.

Respecte le fait de créer une base de données qui mène à l'élaboration du fichier Excel.

Inconvénients :

Figé dans le temps dans le cas où l'utilisateur ne maîtrise pas php.

Outils envisagés:

Dropzone pour l'upload.

Phpmyadmin pour stocker la BDD.

PHPExcel pour l'export.

Python

- Upload (Drag and drop pour envoyer tous les fichiers d'un coup)
- Chercher les textes dans le répertoire
- Insertion de chaque fichier dans la BDD (avec éventuellement une table temporaire pour les nouveaux documents)
- Extraction de la BDD
- Export Excel (avec mise en forme conditionnelle)

Avantages :

Outils pour l'export Excel plus puissant que sur php, possibilité d'utiliser les macros.

On peut paramétrer des calculs.

Inconvénient :

Interface Python à lancer, moins bien maîtrisé.

Outils envisagés:

Serveur web python pour l'upload.

Mysql/sqlite pour stocker la BDD.

Xlsxwriter pour l'export.

Pipeline Pilot

- Upload
- Insertion de chaque fichier dans la BDD
- Export Excel (avec mise en forme conditionnelle ?)

Avantage :

Peut être mis à jour par les utilisateurs confirmés, interface web possible.

Inconvénients :

Outil non maîtrisé, doutes sur les possibilités d'Excel, très peu d'information en ligne à part la documentation de l'éditeur.

Visual Basic for Application (Macros)

- Chercher les textes dans le répertoire
- Export Excel (avec mise en forme conditionnelle)

Avantage :

Fait pour Excel.

Inconvénients :

Pas de création de BDD, donc pas d'historisation.

Pas d'options.