Telecom Nancy 2A
Adrien Coulet, Mario Lezoche, Gabin Personeni

Module GMD - Projet :
"Réalisation d'un système d'intégration de données autour des maladies"

Contexte du projet et définitions
Il serait intéressant pour les acteurs du domaine biomédical de disposer d'un système qui regroupe des données diverses sur les maladies qui sont actuellement réparties dans plusieurs sources de données.

Dans le cadre de ce projet nous considérons trois types d'**entités** :
* des **maladies**,
* des **signes et symptômes** (qui peuvent être considérés comme des maladies) et
* des **médicaments.**

Définitions : Les **maladies** sont des manifestations d'un dysfonctionnement (d'origine psychologique, physique ou/et sociale) de l'état d'un individu. Ce dysfonctionnement a pour conséquence des modifications plus ponctuelles qui peuvent être observés de façon subjective ou objective par l'individu lui même ou par un clinicien. Ces observations sont les **signes et symptômes** d'une maladie. Un **médicament** est une substance (ou une composition de substances) qui est prescrite à un individu dans un but de soigner la maladie ou ses manifestations. On parle dans ce cas de l'**indication** du médicament. Cependant, un médicament peut avoir des **effets secondaires** indésirables qui sont eux même associés à un ensemble de signes et symptômes.

Nous nous intéressons aux types de relations suivantes entre ces entités :
** une maladie se **manifeste** sous la forme de signes et symptômes,
** une maladie peut être l'**indication** d'un médicament
** une maladie peut-être l'**effets secondaire** d'un médicament,

Pour plus de simplicité et comme indiqué plus haut, nous considérons que les signes et symptômes sont eux-mêmes des maladies.

Objectif du projet
Chaque *trinôme* doit développer un système d'intégration de données de type médiateur qui permet de retrouver :
* l'ensemble des données disponibles concernant une maladie (c'est à dire ses signes et symptômes, les médicaments qui la traitent, les médicaments qui la causent),
* la liste des maladies qui partagent une ou plusieurs propriétés, par exemple la liste des maladies qui ont pour symptôme la fièvre et qui sont traitées par la pénicilline.

Ce système doit permettre, à partir d'une requête unique, de considérer le contenu de quatre sources de données hétérogènes.
Une requête est composée d'un ou plusieurs noms de maladie, signes et symptôme ou médicament.

Exigences du projet

1. La langue du projet sera l'anglais (pour les commentaires du code, la documentation et les éventuelles interfaces).
2. Le système doit considérer simultanément les sources de données suivantes :

* **DrugBank**

format : XML
localisation : `http://www.drugbank.ca/system/downloads/current/drugbank.xml.zip`
contenu :

Cette source contient, entre autre, des données sur l'indication du médicament (attribut *Indication*), ses effets secondaires (attribut *Toxicity*). Attention, car ces données sont présentes sous la forme de phrases qui contiennent des noms de maladies. Pour simplifier, nous considérerons qu'un nom de maladie présent dans l'attribut *Indication* (respectivement *Toxicity*) est une indication (respectivement un effet secondaire).

* **Sider 2**

format : MySQL
localisation : *host* : `neptune.telecomnancy.univ-lorraine.fr`
database : `gmd`
login : `gmd-read` ; *pwd :* `esial`
contenu :

Cette source est composée de trois tables (adverse_effects_raw, indications_raw, label_mapping) qui contiennent les indications et les effets secondaires données par les notices d'utilisations de médicaments (*drug label* en anglais). Vous verrez que Sider propose des références aux concepts de l'UMLS appelés CUI. Cet UMLS propose, entre autre, des identifiants uniques appelées CUI pour de nombreux concepts biomédicaux.

* **OMIM**

format : text
localisation : `/home/depot/2A/gmd/projet_2014-15/omim/omim.txt`
et `/home/depot/2A/gmd/projet_2014-15/omim/omim_onto.csv`
contenu :

Cette source contient des données sur les maladies génétiques, notamment leur signes et symptômes, dans les sections marquées par la balise *FIELD* CS. Vous trouverez également le fichier omim_onto.csv qui permet d'associer des CUI à certains éléments d'OMIM.

* **OrphaData**

format : CouchDB
localisation : *host* : `couchdb.telecomnancy.univ-lorraine.fr`
database : `orphadatabase`
login : `votre_login` ; *pwd :* `CouchDB2A`
et `/home/depot/2A/gmd/projet_2014-15/omim/ ORPHA_Mappings.obo`
contenu :
Elle contient des données sur les maladies orphelines ou rares, notamment leur signes et symptômes. Elle contient également des références croisées avec les identifiants d'OMIM et de l'UMLS. Vous pourrez utiliser les vues `GetDiseaseClinicalSignsNoLang` pour obtenir les signes et symptômes d'une ou plusieurs maladies ; `GetDiseases` pour avoir des références croisées de OrphaData avec OMIM et UMLS.

3. Le système d'intégration doit suivre l'architecture de type **médiateur**. C'est à dire que les données doivent rester dans leurs sources d'origine. Lorsque l'utilisateur pose une requête, celle-ci est traduite pour être posée aux différentes sources de données, les résultats de chaque source de données sont ensuite regroupés de façon cohérente avant d'être présentés à l'utilisateur
Cependant, nous recommandons quelques entorses à ce principe pour améliorer les performances de votre système :
- Vous pouvez copier les fichiers texte et XML en local.
- Il est recommandé de faire des indexes *full text* pour ces deux fichiers. Il faut dans ce cas vérifier au lancement du système que les données d'origine sont inchangées et si non, mettre l'index à jour.

4. L'utilisateur doit pouvoir faire une requête par nom de maladie pour retrouver ses signes et symptômes, et les médicaments associés à cette maladie.
L'utilisateur doit pouvoir écrire une requête avec les opérateurs logiques ET et OU. Alors les listes de résultats seront soit l'intersection soit l'union de ceux associés aux maladies d'une requête.

6. La présentation des résultats à l'utilisateur doit lui permettre de distinguer clairement à quoi correspondent les résultats.

7. Dans le fonctionnement par défaut, la chaîne de caractère de la requête doit correspondre exactement au nom du médicament.
Exemple : si l'utilisateur fait une requête avec le nom de maladie *cancer*, uniquement les données qui sont associés avec *cancer* doivent être retournées et pas associées à *breast cancer*.

La soutenance
Durée : 20 minutes / trinômes
5 minutes de présentation des mappings (2 transparents au maximum)
10 minutes de démonstration du système

5 minutes de questions en français.

Dans un premier temps vous présenterez clairement les mappings entre les différentes sources de données que vous aurez définis pour assurer la cohérence (et la complétude) des résultats renvoyés par votre système. Vous ferez ensuite une démonstration de votre système d'intégration en montrant comment il répond aux exigences du projet.
Le jury vous interrogera sur vos choix techniques et vous demandera de les motiver.

Barème

1)La base

La création d'un système qui répond aux exigences précédentes assurera une note de 09/20 au groupe.
Cette première note prendra notamment en considération :
- la qualité de la présentation lors de la soutenance,
- la cohérence des mappings présentés lors de la soutenance,
- la facilité et la rapidité d'utilisation du système.

Si aucune des fonctionnalités de base n'est développée le jour de la soutenance, la note attribuée au groupe sera 0/20.

2)Les fonctionnalités supplémentaires

Pour gagner plus de points, il vous est proposé de développer les fonctionnalités supplémentaires suivantes :

* *une interface graphique* : +1

Le système propose une interface graphique complète et intuitive.

* *requête par nom de médicament* : +1

L'utilisateur peut interroger le système avec un nom de médicament pour retrouver l'ensemble des données disponibles concernant un médicament (c'est à dire les maladies qu'il traite et les maladies qu'il cause).

* *utilisation des synonymes* : +1

L'utilisateur doit pouvoir faire une requête avec des synonymes de noms de maladie (*cancer* et *neoplasm*). A vous de trouver une source de synonymes fiable pour cette fonctionnalité.

* *tri des résultats* : +1

Les résultats sont triés suivant un score. Le score est plus grand notamment si plusieurs sources proposent le même résultat à une requête.

* *Utilisation de jokers dans la requête* : +1

L'utilisateur doit pouvoir écrire une requête partielle en utilisant des caractères joker.
Par exemple « tamoxifen* », pour obtenir les listes de gènes associés avec tous les médicaments dont le nom est ou commence par tamoxifen (par exemple tamoxifen et tamoxifen N-oxide). De la même façon « ep*fever » permet de trouver « episodic fever »

* *Fournir la provenance des données :* +1

L'utilisateur doit pouvoir savoir de quelle source proviennent les résultats d'une requête. Attention car un même résultat peut provenir de deux sources distinctes.

* *Interrogation par famille de maladies :* +2

Le système permet de faire une requête par famille de maladies pour trouver alors les données communes à tous les membres d'une famille de maladie. A vous de trouver une source de familles de maladie fiable pour cette fonctionnalité.

* *Visualisation des résultats :* +1

Un point pourra être accordé si le groupe propose une visualisation intéressante des résultats.

« Chi va piano va sano e va lontano »

2 points seront accordés aux étudiants qui progressent régulièrement. Pour cela il faut valider une étape (*milestone* en anglais) en présentant les fonctionnalités suivantes à la date suivante à votre enseignant de TP. Ce sera à vous de demander à l'enseignant de venir valider vos fonctionnalités et pas l'inverse.

- Milestone 1 : avant le 10/3 (pour les IAMD) une seule requête permet de récupérer des résultats des quatres sources
- Milestone 2 : avant le 3/4 (pour les IAMD) les résultats sont cohérents (il n'y a pas de doublons)

Dates importantes

**Envoie du sujet : semaine du 9 février 2015

**Envoie à adrien.coulet@loria.fr et mario.lezoche@univ-lorraine.fr un lien vers la démo ou son code : 48h avant votre soutenance

**Soutenance du projet semaine du 27 avril 2015