Module GMD : Projet 2019

"Réalisation d'un système d'intégration de données biomédicales"

Contexte du projet et définitions

Vous développerez un système qui permet à un utilisateur, de proposer un ou plusieurs symptômes pour savoir *1)* quelles maladies pourraient causer ce(s) symptôme(s) ou *2)* quels médicaments pourraient en être la cause. Pour cela, votre système devra permettre d'intégrer des données (réelles) diverses sur les symptômes, maladies et médicaments qui sont actuellement réparties dans plusieurs sources de données.

Pour le projet nous considérons ainsi trois types d'**entités** : * des **signes et symptômes**, * des **maladies**, * des **médicaments.**

Définitions **:** les **signes et symptômes** sont des observations de modifications relativement ponctuelles de l'état d'un individu. Ces observations peuvent être la conséquence de dysfonctionnements plus globales que sont les **maladies**. Les signes et symptômes peuvent être observés de façon subjective ou objective par l'individu lui-même ou par un clinicien. Il arrive également que la survenue de signes et symptômes soit causée par un **médicament**. On parle alors des **effets secondaires** du médicament. De façon générale un médicament est une substance (ou une composition de substances) qui est prescrite à un individu dans le but de soigner une maladie ou ses manifestations. On parle dans ce cas de l'**indication** du médicament. Il arrive que 2 médicaments soient prescrits aux patients, le premier pour traiter une maladie, et le second pour limiter un effet secondaire du premier.

Nous nous intéressons aux types de relations suivantes entre nos 3 entités : ** un signe ou symptôme peut-être soit la **manifestation** d'une maladie, ** un signe ou symptôme peut également être l'**effet secondaire** d'un médicament, ** une maladie ou un symptôme peut être l'**indication** d'un médicament.

Objectif du projet

Chaque *trinôme* développera un système d'intégration de données de type médiateur qui permet de retrouver :

- à partir d'un signe ou symptôme, l'ensemble des maladies qui pourraient le causer, mais également les médicaments qui pourraient, de façon indésirable, en être la cause.

- Si le signe ou symptôme peut-être causé par une maladie, la liste des médicaments qui pourraient la traiter ; et si le c'est un effet indésirable, la liste des médicaments qui pourraient traiter l'effet indésirable. Ce système doit permettre, à partir d'une requête unique, de considérer le contenu de 7 sources de données hétérogènes. Une requête devra pouvoir être composée de la conjonction de plusieurs noms de signes ou symptôme. Alors les résultats seront l'intersection des résultats associés à chaque signe et symptôme.

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 différentes sources de données suivantes :

Source 1 : 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).

Source 2 : OrphaData

format : base de données orientée documents, CouchDB localisation :

http://couchdb.telecomnancy.univ-lorraine.fr/orphadatabase/

contenu :

Elle contient des données sur les maladies rares, notamment leur signes et symptômes (appelés *clinical signs*). Elle contient également des références croisées avec les identifiants d'OMIM et de l'UMLS.

Vous pourrez importer les documents JSON avec le système de votre choix (CouchDB, MongoDB, etc.)

Quelques vues disponibles :

GetDiseaseByClinicalSign retourne les paires (label du signe clinique, données JSON sur la maladie associée) GetDiseaseClinicalSignsNoLang retourne les paires (id du signe clinique, données JSON sur la maladie associée)

Source 3 : OMIM

format : text localisation :

/home/depot/2A/gmd/projet_2017-18/omim/omim.txt

et

/home/depot/2A/gmd/projet_2017-18/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.

Source 4 : Sider 4.1

format : MySQL

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

contenu : Cette source est composée de quatre tables (meddra, meddra_all_indications, meddra_all_se*, meddra_freq) qui contiennent les indications et les effets secondaires données par les notices d'utilisations de médicaments. Vous verrez que Sider propose des identifiants appelés CUI que l'on retrouve dans de nombreuses bases de données. *se : side effect

Source 5 : HPO et HPO Annotations

format : OBO et SQLite localisation :

/home/depot/2A/gmd/projet_2017-18/hpo/hpo.obo

et

/home/depot/2A/gmd/projet_2017-18/hpo/hpo_annotations.sqlite

contenu : HPO (Human Phenotype Ontology) est un vocabulaire contrôlé de référence pour les

signes et symptômes. Il contient notamment une liste de synonymes pour les signes et symptômes. HPO Annotations contient des associations entre les identifiants des signes et symptômes de HPO et les maladies de OrphaData et OMIM.

Source 6 : STITCH et ATC

format : Texte localisation :

http://stitch.embl.de/download/chemical.sources.v5.0.tsv.gz

et

/home/depot/2A/gmd/projet_2017-18/atc/br08303.keg

contenu : Ces sources sur les médicaments vont vous permettre d'associer un label aux médicaments par le cheminement suivant : SIDER : stitch_compound_id -> STITCH:compound_id -> Code ATC -> Label.

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 sqlite en local.

- Il est recommandé de faire des indexes *full text* pour les fichiers textes.

A minima, l'interaction entre l'utilisateur et le système se fera via la console.

4. L'utilisateur doit pouvoir faire une requête par nom de signes et symptômes pour retrouver maladies et médicaments associés.

5. L'utilisateur doit pouvoir écrire une requête avec l'opérateur logique ET. Alors la liste de résultats correspond à l'intersection des éléments associés.

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

La soutenance

Durée : 20 minutes / trinômes 5 minutes de présentation des mappings (2 transparents au maximum) 15 minutes de démonstration du système + questions du jury

Dans un premier temps vous présenterez les mappings entre les différentes sources de données que vous aurez définies 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 10/20 au groupe. Cette première note prendra notamment en considération : la qualité de la présentation ; la cohérence des mappings ; la facilité et la rapidité d'utilisation du système.

Sans les fonctionnalités de base, 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 :

**un quantification de la qualité des mappings :* +2 Vous comptez et présentez le nombre de correspondances entre les sources. Par exemple si certains Stitch ID ne correspondent à aucun code ATC, nous perdrons certains résultats. Combien y a-t-il de Stitch ID au total, et combien correspondent à un Code ATC via votre mapping ? Et dans le sens inverse, combien de code ATC et combien correspondent à un Stitch ID ?

* *une interface graphique* : +1 Le système propose une interface graphique complète et intuitive.

* *requête avec des OU et des ET* : +1 Il est possible de faire une requête avec des signes et symptômes articulés avec des OU et des ET et que leurs priorités soient respectées.

* *utilisation des synonymes* :+1 L'utilisateur doit pouvoir faire une requête avec des synonymes de noms de signes et symptômes.

* *Utilisation de jokers dans la requête* : +2 L'utilisateur doit pouvoir écrire une requête partielle en utilisant des caractères joker. Par exemple « hypo* », pour obtenir les résultats associés avec tous les signes et symptômes dont le nom commence par « hypo », par exemple « hypotonia » et « hypoglycemia ». De la même façon «* fever» permet de trouver « episodic fever », et « bl*ing » permet de trouver « bleeding »

* *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.

* *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, dans certains cas, provenir de deux sources distinctes.

* *Visualisation des résultats :* +1 Un point pourra être accordé si le groupe propose une visualisation intéressante des résultats.

Dates importantes

Envoie à sabeur.aridhi@loria.fr les liens vers les slides et la code, 48h avant votre soutenance