Moteur de recherche sémantique au sein du dossier du patient informatisé : langage de requêtes spécifique

Romain Lelong[1] Tayeb Merabti[1] Julien Grosjean[1] Mher B. Joulakian[1] Nicolas Griffon[1,2] Badisse Dahamna[1] Marc Cuggia[3] Suzanne Pereira[4] Natalia Grabar[5] Frantz Thiessard[6] Philippe Massari[1] Stefan J. Darmoni[1,2]

[1]Service d'Informatique Biomédicale, CHU de ROUEN, Haute-Normandie & TIBS, LITIS EA 4108, France
[2]LIMICS, INSERM, U1142, Paris, France.
[3]Inserm U936 – Université de Rennes 1, France
[4]Société Vidal, 92 Issy les Moulineaux, France
[5]STL UMR8163 CNRS, Université Lille 1&3, France
[6]LESIM Université Bordeaux II (Victor Segalen), France

Résumé

La recherche d'information (RI) au sein du Dossier du Patient Informatisé (DPI) doit fournir aux professionnels de santé la bonne information à la bonne personne, au bon moment, et au bon endroit, et de réduire les tâches lourdes de recherche d'information manuelle dans des dossiers papiers voire informatiques. Dans ce contexte, l'objectif de ce travail a été de décrire les fonctionnalités d'un moteur de recherche sémantique au sein d'un DPI. Dans ce papier, nous décrivons un langage de requête orienté objet, conçu pour la consultation de données, flexible et évolutif permettant de prendre en charge n'importe quelle modèle de données. Ce moteur permet des requêtes sur des données structurées et non structurées, sur un patient unique à visée soin ou N patients à visée épidémiologique. Nous avons testé différents types de requêtes sur une base de test de 2 000 patients anonymisés, contenant environ 200 000 comptes rendus.

Abstract

Information Retrieval (IR) in the Electronic Health Record (EHR) should provide healthcare professionals with the right information to the right person at the right time and place and should reduce the hard tasks of manual information retrieval from papers or from computer. In this context, the objective of this study was to describe the features of a semantic search engine implemented in an EHR. In this paper, we describe a flexible and scalable object-oriented query language designed for retrieving and viewing data which support any data model. This search engine deals with structured and unstructured data, on a unique patient in the context of care, and on N patients in the context of epidemiology. In this study, we tested several types of queries on a test databases containing 2,000 anonymized patients and about 200,000 records.

Mots-clés : Dossier du patient informatisé ; recherche d'information ; indexation automatique

Keywords: *Electronic Health Record;information retrieval; automatic indexing*

1 Introduction

Le Dossier du Patient Informatisé (DPI) est une « version informatisée du *dossier du patient papier* » [1]. Hebda and Czar [2] décrivent le DPI comme une ressource d'informations informatisées utilisées en santé pour capturer des données du patient. L'International Organization for Standardization (ISO) a défini le DPI comme « un outil de dépôt d'informations de santé dans une forme informatisable, archivée, et transmissible à des utilisateurs authentifiés ». Son objectif principal est de garantie un soin de qualité, efficace et intégré ; le DPI contient des informations à la fois rétrospectives, actuelles et prospectives [3], utiles à tous les professionnels de santé, avec des prescriptions, de la planification et des évaluations [4]. Ces informations permettent par exemple l'aide à la décision et ou la création de cohortes.

Dans ce contexte, l'objectif de la recherche d'information (RI) au sein du DPI est de fournir aux professionnels de santé la bonne information à la bonne personne, au bon moment et au bon endroit [5]. Dans la pratique, utiliser un outil de recherche dans le DPI doit permettre de réduire les tâches lourdes de recherche d'information manuelle dans des dossiers papiers voire informatiques, et par ce biais, réutiliser ce temps de professionnel de santé pour améliorer la qualité des soins. Pour González-González et al. [6], les professionnels de santé ont besoin de différentes informations et de connaissances pour réaliser leurs tâches :

- Information sur le patient (information sur l'œil du patient s'il est diabétique) ;
- Connaissances (par exemple, sur les recommandations sur une pathologie donnée).

Terry et al. [7] ont décrit cinq options de recherche d'information dans le DPI :

1. Requêtes prédéfinies : l'utilisateur choisit une requête dans un menu ;
2. Requêtes simples personnalisables : l'utilisateur peut écrire une requête simple pour obtenir des résultats souvent hétérogènes, mais pas nécessairement tous pertinents ;
3. Requêtes avancées personnalisables : l'utilisateur peut saisir une grande variété d'informations dans sa requête, le plus souvent séparés par des opérateurs Booléens (ET, OU, SAUF) ;
4. Interface de langage de requêtes structuré : utilisant une interface spécifique pour saisir les requêtes ;
5. Analyse et extraction d'information avec des outils de bases de données : ces outils de bases de données fournissent le plus haut niveau de possibilité pour réaliser des requêtes complexes.

Ces cinq niveaux sont différents en termes d'usage et de complexité. Les trois premiers niveaux apparaissent comme des solutions assez pauvres pour l'utilisation de recherche d'information dans le DPI, car les questions de recherche des professionnels de santé sont de plus en plus complexes.

Dans ce contexte, l'objectif de ce travail a été de décrire les fonctionnalités d'un moteur de recherche sémantique au sein d'un DPI. Le langage de requêtes est plus complexe que le mode d'interrogation classique de recherche d'information documentaire, du fait de l'existence de plusieurs niveaux hiérarchiques (patient, établissement, séjour), puis le niveau plus classique (actes, procédures, codage PSMI, examens biologiques, métadonnées d'un compte-rendu). Le modèle de données utilisé a permis de définir un langage proche de la représentation médicale de la prise en charge d'un patient. Toutes les informations contenues dans le DPI peuvent ainsi pouvoir être affichées à ces différents niveaux. Ce moteur de recherche est en cours de développement dans le cadre du projet RAVEL [8], financé par le programme TecSan de l'Agence Nationale de la Recherche (ANR).

2 État de l'art

Plusieurs outils et plateformes pour la recherche dans le DPI ont été proposés. Nous intégrons ici les systèmes orientés population fondés sur un entrepôt de données, et les systèmes de recherche d'information dans le dossier (mono)patient. Ces outils sont souvent différents suivant le type de données à rechercher : structurées ou non structurées [9].
Dans les systèmes de recherche d'information (SRI) dans le dossier (mono)patient, plusieurs outils ont été décrits dans la littérature : CISearch [10] est l'outil développé et implémenté au sein du DPI de l'hôpital universitaire de Columbia, aux Etats-Unis. Il permet à l'utilisateur d'effectuer des recherches dans l'ensemble des notes en texte libre (comptes-rendus de radiologie, d'anatomo-pathologie, résumés de sortie, notes de soins…) du dossier médical qu'il consulte. Il exploite quelques fonctionnalités de Lucene. MIRS (Medical Information Retrieval System) [11] est également fondé sur Lucene. Citons également le projet LERUDI [12] (dont l'équipe Rouennaise était partenaire sur les terminologies de santé) qui avait comme objectif la RI au sein du DMP (Dossier Médical Partagé) dans un contexte d'urgence, avec une approche sémantique et la création d'une ontologie de domaine.

Dans les entrepôts de données permettant la recherche d'information (multi)patients, I2B2 (Informatics for Integrating Biology and the Bedside) [13] est une plateforme open source développée aux Etats-Unis et dédiée à la recherche translationnelle. I2B2 est implémenté dans plusieurs pays, dans environ 70 CHU, et peut déjà être considéré comme un standard de facto. I2B2 stocke les données dans un entrepôt de données centré sur l'exploitation de données structurées. Un des composants les plus visibles d'I2B2 est un outil open source de sélection des patients appelé « i2b2 workbench », qui est un outil modulaire, facile à utiliser, et qui permet l'interrogation et la visualisation graphique des données cliniques [14]. La plateforme utilise des données biologiques et d'autres données génomiques (surcouche Transmart).
La fonctionnalité de recherche d'information est aussi proposée par le système de requêtes d'OpenEHR [15], qui repose sur un langage dédié, dépendant de sa structure orientée "archétypes". Spécialement conçu pour interroger ce type de modèle sur le Dossier Patient Informatisé, l'AQL (Archetype Query Language)[1] se veut un langage sémantique et indépendant du système. Stanford Translational Research Integrated Database Environment (STRIDE) propose un outil de requêtes nommé « Anonymous Patient Cohort Tool » dédié à la création de cohortes de patients [16]. Le moteur de recherche EMERSE [17] permet de rechercher des termes en plein texte avec des options avancées, adaptées au DPI (exemple : recherches avec troncatures, synonymie, etc.). XOntoRank [18] est un moteur de recherche permettant de faire une RI sémantique dans des documents médicaux structurés conformes au standard HL7 CDA. Ces documents ont la particularité de contenir à la fois des données codées, structurées et des données textuelles. L'outil utilisé comprend deux phases :

- Une indexation sémantique à l'aide de la nomenclature SNOMED
- Une phase de requête dans laquelle les concepts SNOMED des termes extraits de la requête utilisateur sont mappés avec la base de documents XML indexés.

Contrairement aux premiers outils de recherches appliquées sur des données structurées, ces types d'outils exploitent le contenu textuel des DPI. La recherche en utilisant ces outils peut se faire de différentes manières: une recherche plein texte [19] ou une recherche fondée sur les métadonnées décrivant la sémantique du contenu textuel [20], après utilisation d'outils de traitement automatique de langues (TAL).

[1] http://www.openehr.org/wiki/display/spec/Archetype+Query+Language+Description

Roogle [21] est une plateforme française du CHU de Rennes dédiée à la recherche d'information au sein de DPI. Cet outil a été à l'origine du projet RAVEL [8]. Cette plateforme est constituée d'un entrepôt de données stockant les données du DPI (données biologiques [LOINC], données d'actes [CCAM], des comptes-rendus médicaux de radiologie, d'anatomopathologie et des courriers de sortie) et d'un ensemble d'outils de RI combinant des méthodes de RI sémantique basées sur l'exploitation des métadonnées spécifiques aux documents (métadonnées issues du systèmes d'information clinique, métadonnées sur la structure logique du document), et des méthodes de RI plein texte exploitant le contenu textuel. A ce jour, Roogle permet une RI à la fois sur données non structurées et données structurées.

D'autres outils linguistiques permettent la recherche sur des données textuelles, par exemple : Currie et al. [22] proposent une approche linguistique (variation lexicale des termes, prise en compte du contexte) pour analyser les documents médicaux afin d'identifier les patients qui ont des problèmes de cœur et qui fument. Jain et al. [23] proposent une méthode d'expansion d'une requête en utilisant plusieurs sources de connaissances, incluant les relations sémantiques fondées sur des ontologies, des méthodes d'apprentissage supervisé des co-occurences d'un terme. Plaza and Díaz [24] utilisent l'outil MetaMap pour exploiter les relations sémantiques d'UMLS afin de rechercher des cas similaires de DPI [25] pour bénéficier de la puissance du métathésaurus de l'Unified Medical Language System (UMLS) [26].

3 Matériel

3.1 Le modèle de données

L'utilisation d'un DPI est devenue une pratique courante dans tous les hôpitaux. Les modèles de ces DPI varient d'un hôpital à l'autre, mais ils sont le plus souvent complexes : à titre d'exemple, le DPI du CHU de Rouen contient plus de 100 tables. Le modèle de données de notre moteur de recherche décrit dans [9] est volontairement compact (onze tables) pour minimiser les temps de réponse de la RI (voir Figure 1).

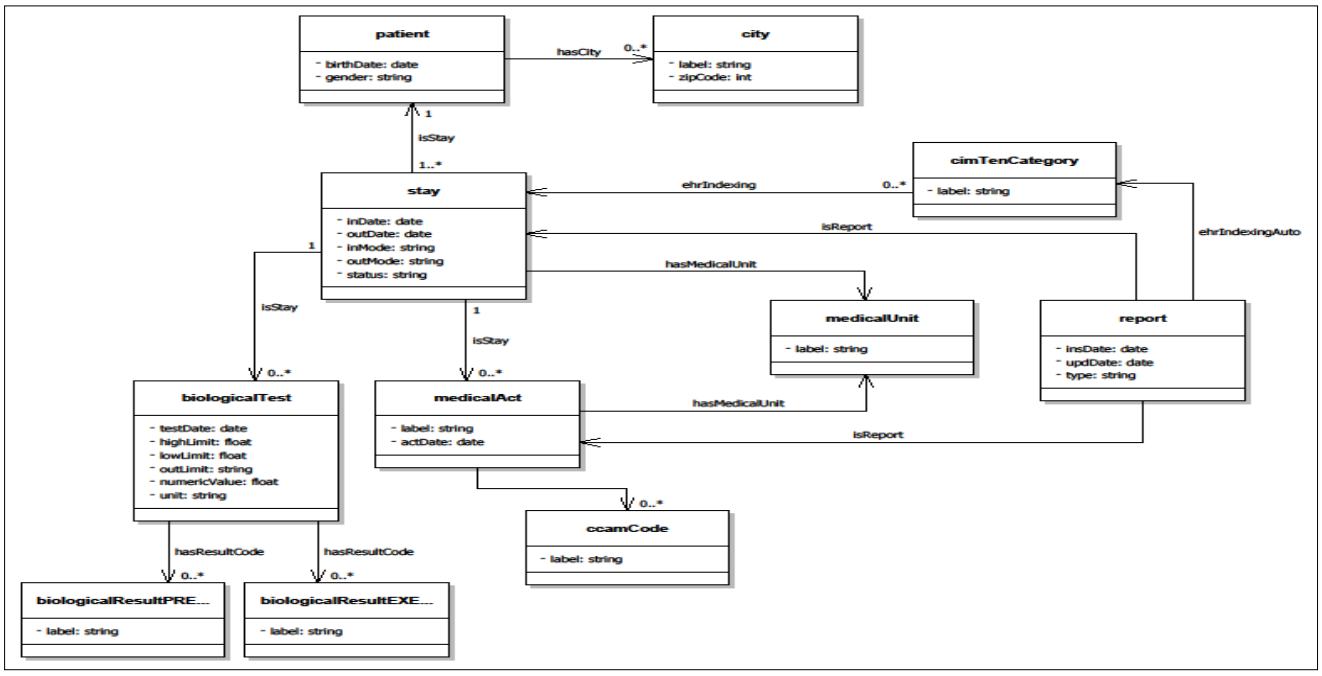

Figure 1 : Schéma du modèle de données du moteur de recherche

4 Méthode

4.1 Le langage de requête

4.1.1 Description :

L'intérêt de proposer un nouveau langage de requêtes est de faciliter la consultation et la recherche d'information dans le DPI et ainsi proposer une alternative aux langages de type SQL utilisés dans de nombreuses structures. Le principe est de rendre invisible la couche SQL afin de proposer une syntaxe la plus simple et intuitive et assez riche pour construire des requêtes complexes sans avoir besoin d'autres connaissances que la connaissance des « entités existantes » dans la base de données : patient, séjour…

Le langage de requêtes proposé dans ce travail à trois caractéristiques importantes :

- Il s'agit d'un langage de requêtes orienté objet avec un motif de syntaxe sur un modèle conceptuel de données. En revanche, il est conçu uniquement pour la consultation de données.
- Ce langage de requêtes est flexible et évolutif : il détecte automatiquement les entités conceptuelles de la base de données et donc il permet de prendre en compte automatiquement de nouvelles « entités », « attribut d'entités » ou « relations à d'autres entités » sans modification préalable du langage de la requête. Cette fonctionnalité importante nous a permis une extension aisée vers les données omiques (génomiques, métabolomiques, protéomiques, méthylation…) [27].
- Ce langage de requêtes a des capacités d'interrogation complète, c'est-à-dire que toutes les données incluses dans la base de données sont interrogeables. Il permet l'interrogation de données : symboliques (présence ou absence d'un diagnostic), numériques (comme les examens biologiques), impliquant de nouveaux opérateurs (>, <, =) par rapport aux opérateurs booléens plus classiques (ET, OU, SAUF) dans la recherche d'informations documentaires, et chronologiques (dernier examen échographique par exemple).

Ce langage permet l'affichage de toutes les informations contenues dans le DPI aux différents niveaux de celui-ci : patient, séjour ou le niveau le plus bas (actes, diagnostics, examens biologiques, dates, etc… qui est le niveau classique d'une RI documentaire).

4.1.2 Syntaxe :

Les principaux composants du langage sont les unités imbriquées de type :

ENTITE (CLAUSE de CONTRAINTES)

- **ENTITE** peut correspondre à n'importe quel type d'entité du modèle conceptuel de données (par exemple : patient, séjour, unité médicale…).
- **CLAUSE de CONTRAINTES** est une expression booléenne qui utilise des opérateurs booléens en combinaison avec des parenthèses pour lier logiquement les contraintes entre elles.

Par exemple, l'expression : patient(dateNaissancePatient='01/01/1937' ET sexe='M'), montre l'utilisation des attributs « **dateNaissancePatient »** et « **sexe »** de l'entité **« patient »,** où les expressions symboliques comme **sexe='M'** et expressions temporelles **dateNaissancePatient ='01/01/1937'** représentent les contraintes reliées entre elles avec l'opérateur booléen **« ET »**.

L'avantage d'une telle représentation est que les opérateurs booléens, les parenthèses, les comparateurs logiques sont les seules variables définies dans la grammaire du langage contrairement aux entités qui sont générées dans la grammaire suivant la base de données utilisée.

4.1.3 Les contraintes :

D'une manière plus simple, les contraintes unitaires sont l'expression d'une restriction d'un attribut direct de l'objet. Ainsi, dans le langage proposé, trois types de données sont traitées : symboliques ou textuelles, numériques, et temporelles, permettant une gestion chronologique (voir Tableau 1).

Tableau 1 : Exemples des types de données traitées dans les contraintes simples

Type de données	**Exemple**	**Description**
Symbolique	patient(sexe="M")	Patient de sexe masculin
Numérique	analyse(valeurNumericAnalyse >6 **ET** valeurNumericAnalyse <=6.25)	Test biologique dont la valeur est comprise entre 6 et 6.25
Date	sejour(dateEntreeSejour=2010-03-10)	Le séjour du 10/03/2010

D'un autre côté la puissance de ce langage de requêtes vient du fait que nous pouvons combiner plusieurs contraintes imbriquées à l'intérieur d'une même contrainte (voir Tableau 2). Pour les valeurs numériques, il est également possible d'exprimer un examen biologique en fonction d'une référence aux bornes inférieures et supérieures à la normale, présentes pour chaque analyse biologique : par exemple, rechercher pour un patient donné toutes les glycémies supérieures à 1,5 fois la normale (sous-entendu supérieur à 1,5 la borne supérieure). Notons ici que cette notion de valeur de la borne supérieure (ou inférieure) peut évoluer au cours du temps. Elle sera prise en compte par notre modèle de données et notre outil de recherche.

Tableau 2 : Exemples avancés sur l'utilisation des contraintes simples

Exemple	**Description**
Patient(analyse(codeEXEResultatBiologique (label="Phosphore") ET 0,81<valeurNumericAnalyse <= 1,17))	Les patients qui ont une analyse indexée par le terme EXE « Phospohore» et dont la valeur est comprise entre 0,81 et 1,17
analyse(codeEXEResultatBiologique(label="Sodium") ET valeurNumeriqueAnalyse<borneInfAnalyse ET patient(id="DM_PAT_125"))	Pour un patient donné (n° 125), affichez toutes les hyponatrémies

4.2 Le moteur de recherche

Par rapport à l'analyse de l'existant, nous avons imaginé notre outil de recherche à la fois comme le plus générique possible et avec une forte contrainte en termes de temps de réponse. Cet outil s'appuie sur nos travaux précédents sur la RI documentaire, où nous avons créé un moteur sémantique pour retrouver un type d'objets unique : les ressources Web[29]. La première phase de

la généricité avait permis la réalisation d'un moteur de recherche permettant de retrouver tout type d'objets (concept ontologique, article scientifique, ressource documentaire), mais toujours limité à un parcours de profondeur un. Dans le moteur actuel, cette limitation a été levée et nous pouvons désormais parcourir un arbre de profondeur N (voir Figure 3). Dans l'hypothèse où l'utilisateur n'aurait pas précisé le chemin à parcourir dans sa requête, un algorithme détermine le chemin optimal. Pour toutes ces raisons, nous avons élaboré un langage de requêtes complexe (niveaux 3 et 4 de Terry et al. [7]).

Ce moteur de recherche permet la recherche d'informations sur des données structurées (essentiellement numériques, mais aussi parfois symboliques, comme le sexe), mais aussi sur des données non structurées (issues essentiellement des différents comptes-rendus d'un DPI). Cette fonctionnalité importante a été rendue possible par l'intégration dans le moteur de recherche des outils de TAL développés par les équipes du Vidal (SP) et de Lille (NG). Pour chaque compte-rendu (CR), les outils TAL sont exécutés, aboutissant un ensemble de métadonnées détaillant les concepts médicaux reconnus en positif, en négatif ou en incertain ; ces outils TAL gérant la négation, souvent présente dans les CR (voir Figure 2). Les concepts peuvent être reconnus s'ils sont présents dans une ou plus des 55 terminologies et ontologies de santé du portail terminologique de santé [28]. Dans cet exemple (Figure 2), l'utilisateur choisit la recherche simple (niveau 2 de Terry et al. [7]) et saisit « gonalgie », qui est reconnu par trois terminologies (SNOMED International, SNOMED CT et MedDRA). Par défaut, la recherche s'effectue également sur le PMSI, mais comme ce concept n'existe ni dans la CIM10 ni dans CCAM, cette recherche s'effectue exclusivement dans les CR, indexées par les outils TAL. Le concept en rouge dans cet exemple « M65.9 - synovite et ténosynovite, sans précision » indique sa négation. A noter que l'outil RAVEL permet l'intégration de différents outils TAL : celui du projet RAVEL et celui développé au CHU de Rouen (ECMT dans le projet SYNODOS[30].

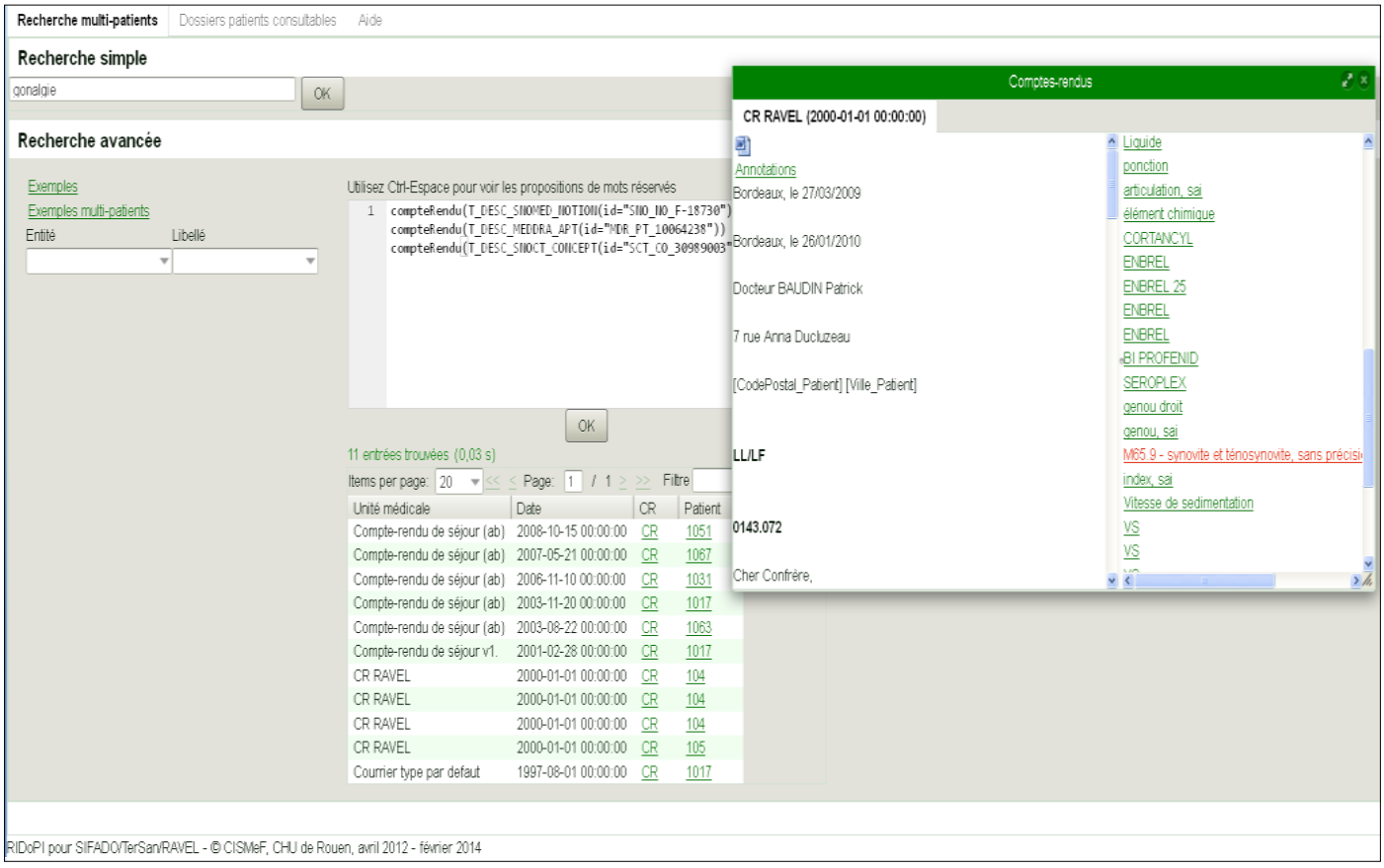

Figure 2: Copie d'écran de l'intégration des outils TAL dans le moteur de recherche RAVEL

A l'inverse, la recherche simple sur « infarctus du myocarde » s'effectuera à la fois sur les données structurées (ici, au sein du PMSI, dans les diagnostics de la CIM10 car l'outil reconnaît une maladie) et sur les données non structurées (au sein des différents CR du patient). Comme certaines expressions dans les CR ne sont pas reconnues comme des concepts médicaux de notre portail terminologique, nous avons également développé une recherche en texte intégral pour pallier ce manque : dans la recherche simple, une requête sera d'abord analysée pour rechercher les concepts médicaux issus du portail, ensuite seulement les expressions non reconnues seront recherchées en texte intégral. Dans notre langage de requêtes, la recherche en texte intégral s'écrit comme tel : compteRendu(FILE.f_html()="expression"),

Notre moteur est générique, car il permet une recherche d'information dans le DPI, au niveau d'un patient unique, essentiellement à visée « soin », de prise en charge effective de ce patient (le projet RAVEL se limite à ce cas d'usage), mais aussi au niveau de plusieurs patients (potentiellement tous les patients d'un établissement). Dans ce dernier cas, les objectifs sont plus variés : épidémiologie (création de cohortes par exemple), recherche clinique (détection de critères d'inclusion d'une étude), mais aussi calculs d'indicateurs de qualité (pour le circuit de biologie, temps global entre la prescription d'un examen et son retour au lit du malade).

Pour permettre la recherche à l'intérieur du DPI, nous avons développé un moteur de recherche permettant une interprétation des requêtes pour extraire les données correspondantes. Ainsi, le processus global du traitement d'une requête en langue naturelle par le moteur peut être divisé en trois phases :

4.2.1 Le parseur de requête

La première étape consiste à une analyse syntaxique de la requête en utilisant la grammaire définie par le langage de requêtes. Ainsi, cette étape permet de valider la requête (conforme ou non à la grammaire) et de deviser la requête en plusieurs « unités lexicales et syntaxiques » élémentaires.

4.2.2 La représentation en arbre de requête

L'objectif de cette étape est d'extraire les éléments significatifs de la requête. Dans cette phase, une représentation sous forme d'arbre de la requête est générée par l'analyse de sa structure hiérarchique (voir *Figure 3*). Pour cause, la structure de l'arbre est bien adaptée pour maintenir une bonne représentation de la structure logique de la requête surtout pour les éléments booléens. Cette structuration est également conforme au raisonnement du professionnel de santé.

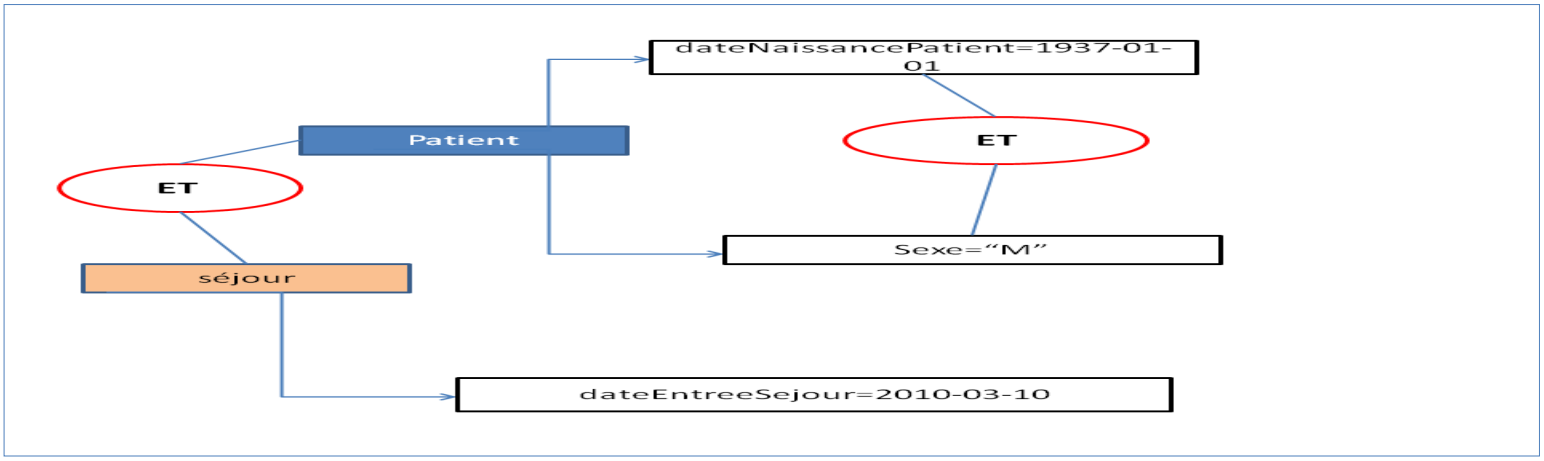

Figure 3 : Représentation en arbre de la requête « patient(dateNaissancePatient='01/01/1937' ET Sexe='M') AND sejour(dateEntreeSejour = 2010-03-10) »

4.2.3 La construction SQL de la requête

L'arbre construit dans l'étape précédente est une représentation complète de tous les aspects de la requête et, par conséquent, peut être utilisé pour générer de manière récursive la requête SQL appropriée qui correspond à la requête demandée.

5 Résultats

En amont du projet RAVEL, nous avons construit au CHU de Rouen une base de test de 2 000 patients anonymisés, contenant environ 200 000 CR pour le projet LERUDI, financé par l'ASIP Santé. Toutes les requêtes présentées dans ce travail ont été testées sur cette base, en attendant l'intégration des données du CHU de Bordeaux. Le Tableau 3 affiche les différentes requêtes gérées par le moteur de recherche du projet RAVEL.

Tableau 3 : Exemple de requêtes du moteur de recherche RAVEL

Exemple	**Description**
sejour(patient(id="DM_PAT_21") AND acte(label="PRELEVEMENT SANGUIN"))	Les séjours du patient DM_PAT_21 dans lesquels un acte de prélèvement sanguin à eu lieu
uniteMedicale(**sejour**(**patient**(id="DM_PAT_21") AND **acte**(label="PRELEVEMENT SANGUIN")))	Les unités médicales des séjours du patient DM_PAT_21 dans lesquels ont eu lieu les actes de prélèvement sanguin
analyse(**patient**(id="DM_PAT_1078") AND **prestationCodeResultatBiologique**(label="Plaquettes") AND 10*valeurNumeriqueAnalyse<borneInfAnalyse)	Les analyses de plaquettes du patient DM_PAT_1078 dont le résultat est plus de 10 fois inférieures à la normale
acte(**codeCCAM**(id="CCA_AM_EQQM006") AND dateActe="MAX")	Le dernier acte de code CCAM AM_EQQM006

Nous avons ensuite tenté de répondre au cas d'utilisation définis par notre collègue de Bordeaux (FT). Par exemple, nous sommes capables de répondre aux problèmes suivants des professionnels de santé :

- Fournir au cours du temps une visualisation globale du taux de polynucléaires d'un patient donné porteur d'une polyarthrite rhumatoïde ;
- Fournir tous les comptes-rendus d'imagerie qui affichent la « notion de métastases », avec la requête en texte intégral suivante : compteRendu(FILE.f_html()="notion de métastases"), puisque cette expression « notion de métastases » n'est pas un concept reconnu par notre portail terminologique, alors que le terme « métastase » aurait bien entendu été reconnu.

Nous avons signalé que notre outil de recherche savait gérer des requêtes pour un patient donné et aussi pour plusieurs patients (voir *Figure 4*).

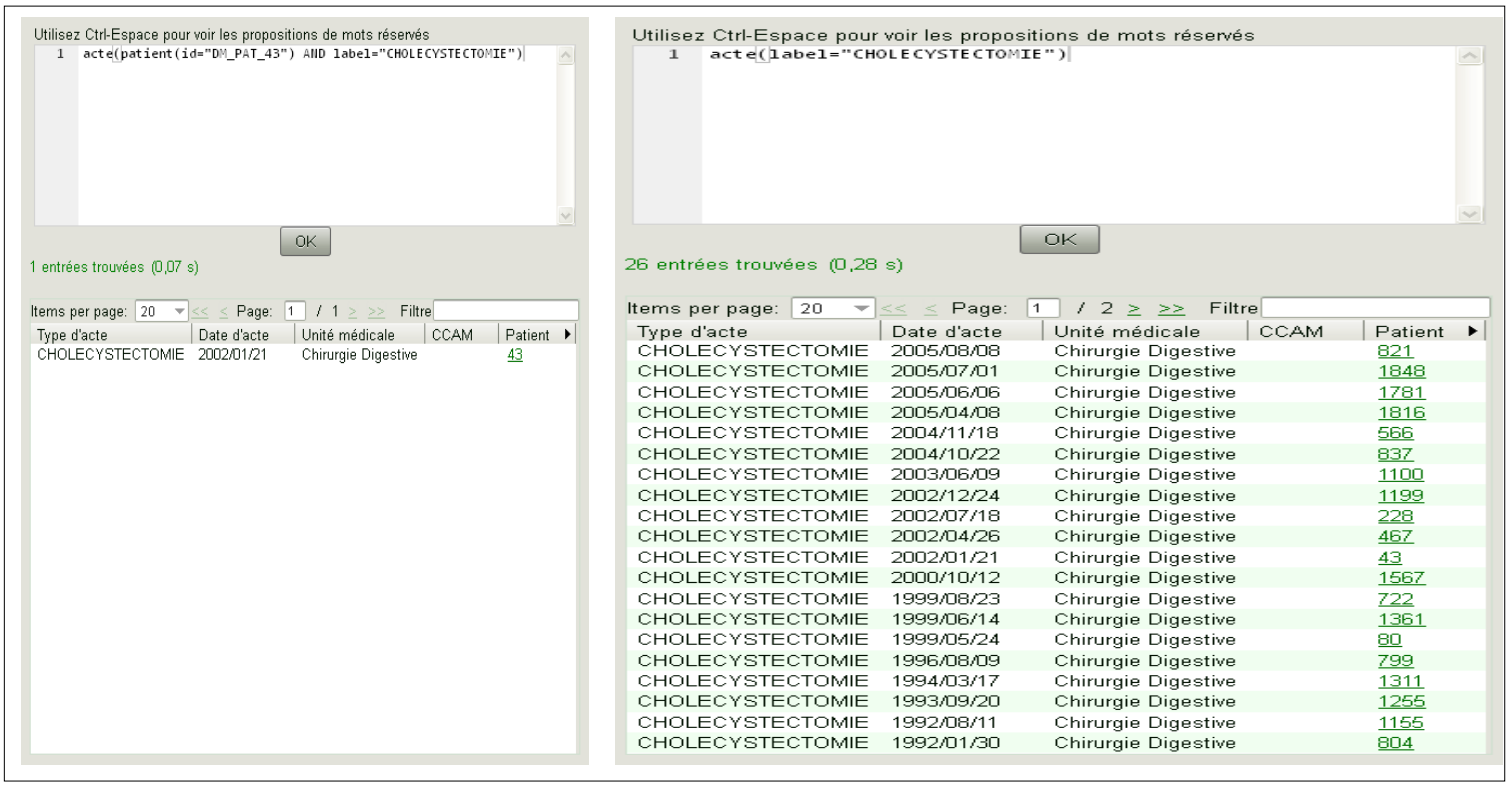

Figure 4 : Exemple d'une requête similaire pour les deux niveaux

En termes de temps de réponse, les deux contextes sont très différents. Pour un patient unique, nous avons fixé le maximum acceptable pour un professionnel de santé toujours pressé à deux secondes (il s'agit d'un temps serveur qui ne tient pas en compte le temps d'affichage qui varie selon la qualité de l'ordinateur et du réseau). Pour toutes les requêtes testées sur notre échantillon de 2 000 patients, ce temps a toujours été inférieur à ce seuil. En revanche, le temps d'affichage des examens biologiques reste encore trop long (de l'ordre de 10 secondes), mais il s'agit alors d'un problème de visualisation et non de recherche d'information proprement dit.

Voici également comment s'écrivent deux requêtes pourtant sur la même notion, mais s'affichant à deux niveaux (patient et séjour). Dans notre exemple, il s'agit de rechercher

- soit le dernier séjour dans le département d'imagerie médicale ; la requête est alors : sejour(uniteMedicale(label="Imagerie Médicale") AND dateEntreeSejour="MAX") ;
- soit les patients ayant eu un dernier séjour dans le département d'imagerie médicale ; la requête est alors : patient(sejour(uniteMedicale(label="Imagerie Médicale") AND dateEntreeSejour="MAX"))

Le sens médical de ces deux requêtes est très différent, et sera sans doute utilisé plutôt dans le service d'imagerie ou plutôt dans les services de soins.

6 Discussion

Nous avons présenté dans ce travail les principales fonctionnalités du moteur de recherche du projet RAVEL, en insistant sur son langage de requêtes que nous jugeons générique et puissant. Notre approche est générique, permettant une recherche d'information sémantique multi-terminologique, à la fois sur un patient unique à visée soin et sur plusieurs patients à visée épidémiologiques. Cet outil de recherche permet également des requêtes à plusieurs niveaux du

DPI : patient, séjour, et niveau le plus bas, c'est-à-dire toutes les données du DPI (actes, prescription, PMSI, biologie, compte-rendus). Une recherche en texte intégrale a été ajoutée également. Enfin, nous avons étendu notre approche aux données omiques.

Néanmoins, nous envisageons, suite au projet, d'effectuer une évaluation comparative de l'expressivité et de précision du langage avec le standard de facto qu'est I2B2 [13], [14] et Transmart, sa surcouche omique. I2B2 étant installé dans plusieurs hôpitaux français, une comparaison de la puissance respective des langages de requêtes de ces deux outils pourra ainsi être menée. Dans tous les cas de figure, un parseur pourrait également être développé entre le modèle de notre outil [8] et celui d'I2B2, permettant aux données de basculer entre les deux outils pour unir si nécessaire les résultats des deux langages de requêtes.

Limites :

A ce jour, le langage de requêtes permet de décrire les niveaux de 2 à 5 de Terry et al [7]. Nous n'avons pas imaginé jusqu'à présent de niveau 1, avec une liste prédéfinie de requêtes, mais nous le discuterons avec les différents professionnels de santé du projet RAVEL. Le langage de requêtes est complexe à manipuler de notre point de vue. Néanmoins, nous l'avons démontré à plusieurs médecins de santé publique, qui disent être favorables à son usage après formation rudimentaire. Les documentalistes, les professionnels des sciences de l'information, les informaticiens sont également à même de l'utiliser dans les mêmes conditions. Quant au professionnel de santé, il est impératif de prévoir plusieurs interfaces, en dehors de la recherche simple : une des pistes d'inspiration pourrait être l'interface d'I2B2.

Cet outil de recherche ne permet pas de répondre à toutes les questions des cas d'usage établis dans le projet RAVEL. Par exemple, à ce jour, nous ne savons pas répondre à la question : « faire le comptage des petites articulations touchées lors d'une polyarthrite rhumatoïde (c'est-à-dire rouges ou gonflées) ».

Notre modèle d'information ne permet pas encore de gérer le niveau « établissement », dans la mesure où ne gérons pour l'instant que des données d'un seul établissement de santé.

En perspective, et en extension du projet RAVEL au CHU de Rouen, nous allons mener une étude de passage à l'échelle de notre outil de recherche sur les 65 000 patients de cet établissement passés dans le service de dermatologie depuis 1992. L'objectif sera de tester cet outil dans un contexte multi-patients pour créer une cohorte.

7 Conclusion

Nous avons présenté dans ce travail l'outil de recherche pour retrouver une information de santé au sein du dossier du patient informatisé, en insistant sur sa généricité et son langage de requêtes puissant mais difficile à manier pour un professionnel de santé.

Remerciements

Ce travail a été en partie financé par l'Agence Nationale de la Recherche (ANR) et la Direction Générale de l'Armement (DGA), sous le numéro Tecsan ANR-11-TECS-012.

Références

[1] Sewell J. *Thede: Informatics and Nursing: Opportunities and Challenges*. 3rd ed. Philadelphia, PA: Lippincott, Williams, & Wilkins; 2013.

[2] Hebda T, Czar P. *Handbook of Informatics for Nurses and Healthcare Professionals*. 5th ed.

Boston, MA: Pearson; 2012.

[3] International Organisation for Standardisation (ISO). 20514 Draft Technical Report: EHR Definition, Scope and Context.

[4] Garde T, Knaup T, Hovenga E, Herd S: Towards semantic interoperability for electronic health records. *Methods Inf. Med*. 2007; 46 (3): 332–343.

[5] Ondo K, Wagner J, Gale K. The electronic medical record: Hype or reality? Journal of Healthcare Information Management, 2002; 17(4):2.

[6] González-González AI, Dawes M, Sánchez-Mateos J, Riesgo-Fuertes R, Escortell-Mayor E, Sanz-Cuesta T, Hernández-Fernández T. Information needs and information-seeking behavior of primary care physicians. *Ann Fam Med*. 2007; 5: 345-352.

[7] Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic C. Using your electronic medical record for research: a primer for avoiding pitfalls. *Fam Pract*. 2010;27(1):121-6.

[8] Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, Campillo B, Jouini W, Grosjean J, Massari P, Griffon N, Dupuch M, Tayalati F, Dugas E, Balvet A, Grabar N, Pereira S, Frandji B, Darmoni S, Cuggia M. RAVEL: retrieval and visualization in ELectronic health records. *Stud Health Technol Inform*. 2012;180:194-8.

[9] Dirieh Dibad AD. *Recherche d'Information Multi Terminologique au sein d'un Dossier Patient Informatisé*. Thèse de doctorat. Université de Rouen, 2012.

[10] Natarajan K, Stein D, Jain S, Elhadad N. An analysis of clinical queries in an electronic health record search utility. *Int J M Inform*. 2010 Jul;79(7):515-22.

[11] Spat S, Cadonna B, Rakovac I, Gütl C, Leitner H, Stark G, Beck P. Enhanced information retrieval from narrative German-language clinical text documents using automated document classification. *Stud Health Technol Inform*. 2008;136:473-8.

[12] Charlet J, Declerck G, Dhombres F, Gayet P, Miroux P, Vandenbussche PY. Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. 23èmes journées francophones d'Ingénierie des connaissances, Paris : France, 2012.

[13] Informatics for Integrating Biology and the Bedside. 2008 [https://www.i2b2.org].

[14] Deshmukh V, Meystre S, Mitchell J. Evaluating the i2b2 system for clinical research. *BMC Medical Research Methodology*. 2009;9(1):70.

[15] Kalra D, Beale T, Heard S. The openEHR foundation. *Studies in health technology and informatics*. 2005; *115*, 153-173.

[16] Lowe H, Ferris, T, Hernandez P, Weber S. Stride–an integrated standards-based translational research informatics platform. *In: Proceedings of the AMIA Annual Symposium*. 2009;391.

[17] Hanauer D.A. EMERSE: The Electronic Medical Record Search Engine. *AMIA Annu Symp Proc*. 2006; 2006: 941.

[18] Farfan F, Hristidis V, Ranganathan A, Weiner M. Xontorank: Ontology aware search of electronic medical records. In Proceedings of the 25th International Conference on Data Engineering, 2009; 820–831. IEEE.

[19] Cuggia M, Bayat S, Garcelon N, Sanders L, Rouget F, Coursin A, Pladys P. A full-text information retrieval system for an epidemiological registry. *Studies In Health Technology And Informatics*. 2010. 160(Pt 1):491–495.

[20] Chung J, Murphy S. Concept-Value Pair Extraction from Semi-Structured Clinical Narrative: A Case Study Using Echocardiogram Reports. *In Proceedings of the AMIA Annual*

Symposium. 2005; 131–135.

[21] Cuggia, M., Garcelon, N., Campillo-Gimenez, B., Bernicot, T., Laurent, J. F., Garin, E., Happe, A., Duvauferrier, R. (2011). Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform*, *169*, 584-588.

[22] Currie, A., Cohan, J., and Zlatic, L. Information retrieval of electronic medical records. Computational Linguistics and Intelligent Text Processing, 2001;460–471.

[23] Jain H, Thao C, Zhao H. Enhancing electronic medical record retrieval through semantic query expansion. *Information Systems and E-Business Management*. 2010; 1–17.

[24] Plaza L, Díaz A. Retrieval of similar electronic health records using UMLS concept graphs. *Natural Language Processing and Information Systems*. 2010; 296–303.

[25] Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp*. 2001 ;17-21.

[26] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inform Med*. 1993; 32:281–91.

[27] Cabot C, Grosjean J, Lelong R, Lefebvre A, Lecroq T, Soualmia LF, Darmoni, SJ. Omic Data Modelling for Information Retrieval. *Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO*. 2014 (in press).

[28] Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Teaching medicine with a terminology/ontology portal. *Stud Health Technol Inform*. 2012;180:949-53.

[29] Darmoni SJ, Thirion B, Leroy JP, Douyère M, Lacoste B, Godard C, Rigolle I, Brisou M, Videau S, Goupy E, Piot J, Quéré M, Ouazir S, Abdulrab H. Doc'CISMEF: a search tool based on "encapsulated" MeSH thesaurus. Stud Health Technol Inform. 2001;84(Pt 1):314-8.

[30] Dupuch M, Segond F, Bittar A, Dini L, Soualmia LF, Darmoni SJ, Gicquel Q, Metzger MH. Separate the grain from the chaff: make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records. *Proceedings of the LTC'13 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics.* 2013 (in press.).

Adresse de correspondance

Stéfan Darmoni, Service d'Informatique Biomédicale, Cour Leschevin, Porte 21, 3ème étage, 1 rue de Germont 76031 Rouen Cedex; Courriel : stefan.darmoni@chu-rouen.fr