Intégration des données de prescription dans un entrepôt de données biomédicales

Integration of prescription data in a clinical data warehouse

Kevin Dalleau[1] **Denis Delamarre[1,2]**
Thibault Ledieu[1,2] **Marc Cuggia[1,2]**

[1]Laboratoire Traitement du Signal et de l'Image (LTSI) INSERM : U1099 – Université de Rennes 1
[2] CHU Rennes, Pôle Pharmacologie et Investigation cliniques, Unité « Fouille de données », F35033 Rennes, France

Résumé

L'intégration de données relatives au médicament au sein d'un entrepôt de données biomédicales et leur réutilisation est d'une importance majeure, les prescriptions étant au centre de la prise en charge du patient au cours de séjour. En utilisant une base de données médicamenteuse indépendante, Thériaque, et une définition standardisée des prescriptions et plus généralement du circuit de médicament, la CIOsp (Classification InterOpérable des spécialités), nous avons implémenté une nouvelle facette a notre moteur de recherche, Roogle. Nous avons permis à l'utilisateur de rechercher la prescription d'un médicament, soit via la sélection de code ATC, soit directement via l'utilisation d'un formulaire de recherche avec autocomplétion et gestion des doublons. Bien que beaucoup de travail reste à faire, ces développements ouvrent la voie à une toute nouvelle manière de constituer des cohortes en vue de réaliser des essais cliniques, ou encore de trouver de potentiels effets indésirables liés à une molécule.

Abstract

Integration of drugs-related data into a clinical data warehouse and its good use is a key issue, prescriptions being at the center of patient care in a hospital. Using an independent drug-database, Thériaque, and a standardized definition of prescriptions and medications - the CIO (InterOperable Classification), we were able to implement a brand new facet to Roogle, our search engine. We enabled the user to search for a drug prescription either by selecting corresponding ATC codes, or directly with a search form including autocompletion and duplicates management. Althought a lot of work is still to be done, these developments pave the way to whole new ways to create cohorts of patients for clinical studies, or to find potential adverse effects related to a specific drug.

Mots-clés : Informatique médicale ; Conférence ; Entrepôt de données ; Médicaments ;

Prescription : Big data

Keywords: *Medical Informatics ; Conference ; Data Warehouse ; Drugs, Prescriptions ; Big*

Articles longs des 15es Journées francophones d'informatique médicale, JFIM 2014, pages 59–70
Fès, Maroc, 12–13 juin 2014

data

1 Introduction

Le domaine médical connait à l'heure actuelle une massification des données en santé (notion de Big Data)[1]. Ce gisement de données peut être réutilisé pour soutenir la recherche. Dans ce but, les établissements, en particulier les CHU, sont en train de se doter d'entrepôts de données biomédicales, dont l'objectif est d'intégrer et d'exploiter les données issues du système d'information hospitalier dans un objectif d'interrogation populationnelle. Parmi les données importantes exploitées dans ces entrepôts, celles concernant le médicament représentent plusieurs défis tant sur les plans de leur intégration que de leur exploitation.

Premier défi: gérer l'hétérogénéité des sources contenant des informations sur le traitement des patients. En effet, dans un SIH, ces informations peuvent être contenues dans des documents textuels (compte rendus d'hospitalisation, courriers), mais aussi dans des documents structurés, depuis les modules de prescriptions des DPI où cette information est le plus souvent codée selon une norme sémantique, ou encore dans des formulaires structurés de dossiers de spécialité. Il s'agit donc de pouvoir intégrer ces données hétérogènes dans un même entrepôt de données, et de permettre leur exploitation.

Deuxième défi: valoriser l'exploitation sémantique de ces données hétérogènes. Pour cela, l'interrogation des données sur le médicament doit s'appuyer sur un modèle de représentation sémantiquement riche, qui permette par exemple de rechercher des patients ayant pour traitement une classe médicamenteuse spécifique (par exemple, les patients sous fluoroquinolone), ou ayant un traitement indiqué dans une pathologie particulière (les patients ayant un médicament indiqué dans la chimiothérapie du sein, par exemple).

Pour cela, l'utilisation d'une base de connaissance sur le médicament devient incontournable.

Le travail présenté ici vise à exposer l'approche d'intégration et de recherche d'information concernant le médicament au sein de l'entrepôt de données biomédicales du CHU de Rennes (système Roogle).

2 État de l'art

À l'heure actuelle, le projet le plus avancé dans l'interrogation et l'exploitation d'entrepôts de données biomédicales est celui porté par le National Institude of Health (NIH) : I2B2 (Informatics for Integrating Biology and the Bedside). Il s'agit d'une plateforme open source dont l'objet est l'exploitation de données stockées au sein d'entrepôts de données biomédicales, initialement sur des données de génétique [2, p. 2].

I2B2 permet d'effectuer des recherches sur le médicament, et ce sous de nombreux angles (voie, spécialité commerciale, dose, etc.).

Cependant :

- I2B2 se base sur des données, référentiels et thésaurus propres aux Etats-Unis (RxNorm, nom de spécialités spécifiques), le rendant peu utilisable en France
- Les flux PN13 [PréNorme 13, majoritaires en France] ne sont pas pris en charge par la solution

Nous proposons, par l'ajout de cette fonctionnalité à Roogle, une exploitation des données liées

au médicament se basant sur un standard d'interopérabilité vivant et largement adopté en France, le standard PN13 – SIPH2, et s'appuyant sur une base de connaissances médicamenteuse puissante et exhaustive, Thériaque.

3 Matériel et méthodes

3.1 Matériel

3.1.1 Roogle

Projet débuté au sein de CHU de Rennes en 2007, Roogle était à ses débuts un système de recherche d'information au sein du dossier d'imagerie médicale, gérant à la fois la recherche sur des données structurées et sur des données plein-texte[3].

L'outil s'est au fil des années enrichi de nombreux autres flux, tels que les comptes-rendus hospitaliers, les actes réalisés, les comptes-rendus d'anatomopathologie, etc. Les champs d'applications sont nombreux, allant de la constitution de cohorte pour des fins de recherche à l'évaluation des pratiques professionnelles, en passant par la recherche de cas spécifiques.

À ce jour, le système exploité à Rennes contient plus d'un million de patients et près de 15 millions de documents.

3.1.2 Base Thériaque

Editée par le Centre National Hospitalier d'Information sur le médicament (CNHIM), Thériaque constitue l'une des plus grandes bases de données médicamenteuses en France. Elle regroupe de manière exhaustive toutes les informations sur les médicaments commercialisés en France, en puisant l'information de sources officielles (agence du médicament, journaux officiels, avis de la HAS, etc.), complétées d'informations issues d'ouvrages de référence. [4]

Ces données sont très structurées, permettant leur exploitation et leur intégration dans des applications tierces. Celles-ci sont par ailleurs mises à jour quotidiennement.

La base nous est fournie par le CNHIM sous de nombreux formats. L'utilisation des informations qu'elle contient permet de répondre à de nombreuses questions concernant les médicaments, allant de leur code ATC à leurs indications.

L'interopérabilité est assurée via l'intégration des fichiers UCD. Fournis par le CIP (Club Inter Pharmaceutique), le code UCD est un code numérique à 7 chiffres, correspondant à la plus petite unité de dispensation (gélule, comprimé, etc.).

3.1.3 Le standard PN13 – SIPH2

Il s'agit d'un modèle du circuit du médicament, partagé par les différentes applications du système d'information du circuit, qui définit les flux d'informations à émettre à chaque transmission. Le standard est constitué de 3 éléments indissociables :

- Le référentiel d'architecture technique, dont l'objectif est de décrire les messages qui transitent entre les systèmes : messages de prescriptions, de dispensation nominative, etc.

- Le référentiel des nomenclatures, contenant l'ensemble des nomenclatures permettant de coder les attributs des éléments définis dans le référentiel d'architecture technique.
- La Classification InterOpérable des spécialités (CIOsp), composée d'un noyau contenant les informations de base (code UCD, nom commercial, classe ATC, etc.), et de composants additionnels.

Le standard garantit l'interopérabilité, permettant aux applications l'utilisant de communiquer et de se comprendre.

La distribution consiste en une archive contenant toutes les données de la CIOsp sous forme de fichiers texte, chaque fichier correspondant à une table du modèle physique de données.

3.1.4 Données de prescription issues de DxCare

Le CHU de Rennes a choisi DxCare ® (Medasys) comme solution au sein de son système d'information. Elle permet, via l'existence de plusieurs modules, de gérer de nombreux aspects du fonctionnement hospitalier : mouvement, séjour, dossier médical, planification des rendez-vous, prescriptions, etc.

DxCare ® est "CIO-inside", tout comme de nombreuses autres solutions du marché (Pharma ®, Génois ®, etc.). Ainsi, l'ensemble des prescriptions réalisées au sein de l'établissement transite au format PN13. L'architecture d’un élément de prescription dans ce format est présenté figure 1.

Figure 1 : Architecture d'un élément de prescription au format PN13

3.2 Méthodes

3.2.1 Cas d'usages adressés

L'intégration des prescriptions, et, plus largement, des médicaments à l'entrepôt de données répond à une demande croissante de la part des professionnels, dans plusieurs domaines (recherche clinique, épidémiologie, professionnels de la pharmacovigilance, etc.).

Afin d'identifier les besoins et la méthode d'intégration et de recherche d'information sur ce domaine, nous définissons les cas d'usages de Roogle suivants.

- Constitution de cohorte :

La recherche de patients ayant reçu un ou plusieurs traitement(s) spécifique(s) est l'un des éléments clés de la constitution de cohorte. L'intégration de la temporalité, rendue possible grâce à la haute structuration des flux de prescriptions, est par ailleurs un besoin très fort afin de mener de la manière la plus fine possible cette tâche, et de se rapprocher des contraintes du screening.

- Etude de faisabilité

Sur le même principe que la constitution d'une cohorte de patients, il est possible qu'un promoteur, industriel ou institutionnel, utilise Roogle pour déterminer les capacités d'inclusion de son site et d'un site investigateur. Contrairement à la phase de screening, les résultats de la requête ne porteront pas sur les informations individuelles de chaque patient mais sur les statistiques du site.

- Suivi des protocoles

Les protocoles ayant souvent des critères stricts devant être respectés, rendre possible la visualisation des différents évènements ayant eu lieu au cours de leur déroulement, notamment au niveau de la prise de médicaments, est aussi l'un des cas d'usage retenus.

- Recherche de potentiels effets indésirables

La prise en charge du médicament au sein de l'entrepôt de données permettra, en les croisant avec les données déjà présentes en son sein (courriers de sorties, résultats de laboratoire, etc.), de rechercher d'éventuels évènements consécutifs à la prise d'un médicament, et ce sur de grandes quantités de données.

3.2.2 Intégration

L'un des principes clés de Roogle est de conserver une structure de document la plus proche possible du document d'origine. Il s'agit de permettre à l'utilisateur qui fait une requête sur une information précise (la prescrition d'un médicament par exemple) de retrouver lors de l'affichage des résultats le contexte dans lequel a été prescrit le médicament (en l'occurrence l'ordonnance). La conservation du document le plus proche du format natif dans l'entrepôt de données est aussi un moyen d'assurer la meilleure intégrité de l'information, au cas où celle-ci deviendrait opposable. C'est pour cette raison que les données de prescription sont intégrées au format natif PN13 dans l'une des tables, contenant déjà les autres documents précédemment intégrés à l'entrepôt (compte-rendus, etc.).

Afin de tirer parti au maximum du fait que les flux de prescriptions soient structurés, les données sont par ailleurs stockées dans une autre table dédiée, où cette fois ci chaque ligne correspond à une ligne de la prescription. Cette méthode permet de conserver la plus fine granularité, chaque ligne étant indépendante, et pouvant être retrouvée via ses différents éléments. De plus, cette deuxième intégration permet de faciliter une éventuelle agrégation future des résultats. Plusieurs lignes de cette table peuvent donc correspondre à une seule et même prescription, stockée entièrement dans la première table.

L'ensemble de la base de données médicamenteuse Thériaque est par ailleurs intégrée, afin qu'elle puisse être utilisée comme base de connaissance pour la construction de requêtes sémantiquement riches.

Pour illustrer l'utilisation de la base Thériaque au sein de Roogle, nous avons développé un script permettant de charger la classification ATC, à partir de la base Thériaque, dans une table dédiée aux Thésaurus. Ce modèle de représentation sous forme de chemin parcourant un arbre hiérarchique a déjà été utilisé pour l'exploitation d'autres données structurées qui utilisent des terminologies dédiées (comme la CIM-10 pour les diagnostics, et la CCAM pour les actes médicaux et chirurgicaux).

La dernière distribution en date de la CIOsp a enfin été intégrée, afin de pouvoir assurer le la bonne marche de fonctions spécifiques de décodage des documents au format PN13.

3.2.3 Recherche d'information et visualisation

La problématique de visualisation des données lors d'une recherche d'information est double :

- l'utilisateur doit pouvoir construire sa requête de façon intuitive et exhaustive,
- la visualisation des résultats doit permettre à l'utilisateur d'extraire l'information de façon simple et non ambigüe.

Le constructeur de requêtes de Roogle propose actuellement à l'utilisateur 2 types de recherche :

- Les recherches sur les informations structurées comme les données issues du PMSI,
- Les recherches plein texte (full-text) portant sur toutes les informations du Dossier Patient Informatisé, que cette information soit structurée ou non.

Dans le cas particulier de la recherche sur des prescriptions, l'utilisateur s'orientera plus généralement vers une recherche sur des données structurées afin de profiter de la structuration des données des flux de prescriptions. Le formulaire de recherche dans les données structurées préexistant sera donc augmenté de la classification ATC, afin de pouvoir y effectuer des recherches simplement. La recherche de prescription peut également se faire en plein texte sur les courriers médicaux comme sur les flux de prescriptions.

Le pivot de toute recherche est le code UCD. Ce code, assurant l'interopérabilité au sein de tous les systèmes partageants les informations sur le médicament, permet à notre système de rechercher directement au sein des documents de prescriptions, au sein desquels chaque ligne est identifiée par son code UCD et son code ATC.

La CIOsp est enfin très largement utilisée afin de traduire en clair les informations présentes au sein des documents. En effet, de nombreux éléments sont codifiés (type de substance, voie, forme, etc.), et il est nécessaire de les convertir en langage humain afin de simplifier la lecture. Des fonctions se chargeant du décodage sont donc utilisées, s'appuyant directement sur la distribution la plus à jour.

4 Résultats

4.1 Recherche et visualisation

4.1.1 Recherche via code ATC

L'intégration de la classification ATC au thésaurus permet de naviguer dans les branches de la classification, notamment dans le cadre de la recherche et de la sélection de substances, comme vu figure 2.

Devant chaque nœud de l'arborescence figure le nombre de patients ayant une prescription d'un médicament appartenant à ce nœud. L'utilisateur peut choisir un nœud comme critère de sélection (critère d'inclusion ou d'exclusion) et les combiner à d'autres critères.

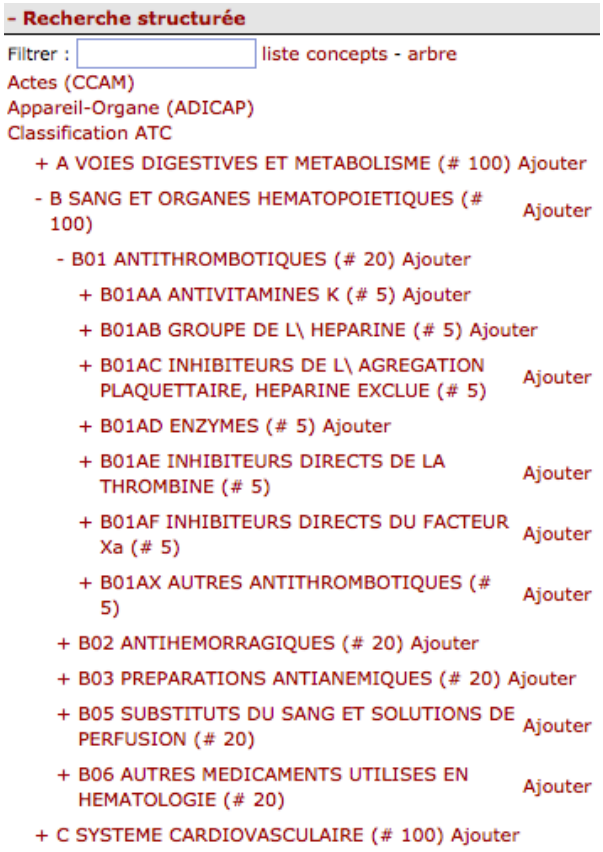

Figure 2 : Recherche et inclusion de médicaments par navigation dans la classification ATC

Le formulaire de recherche est composé :

- D'un arbre de concepts où chaque embranchement représente une sous-catégorie de l'ATC. Cet arbre peut être exploré manuellement par l'intermédiaire d'un menu dépliable.
- D'un champ texte permettant de filtrer les différents concepts présents dans les thésaurus intégrés, et de retourner uniquement ceux qui contiennent des patients, comme le montre la figure 3.

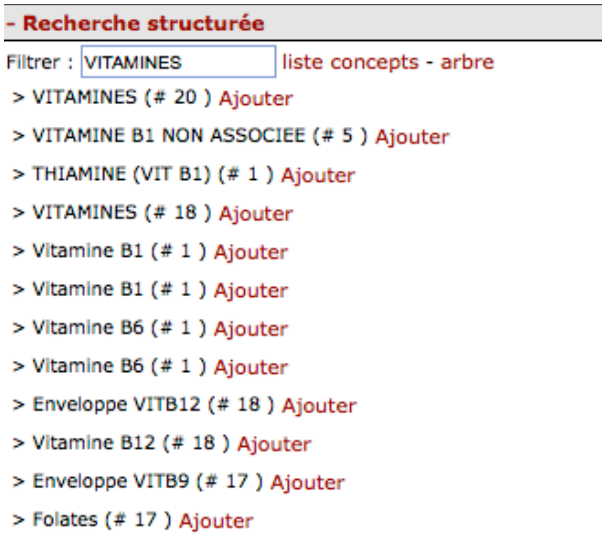

Figure 3 : Recherche de code par un moteur de recherche terminologique dans tous les thésaurus indexés.

4.1.2 Recherche via UCD

Roogle permet également de rechercher un médicament directement par son code UCD, comme le montre la figure suivante.

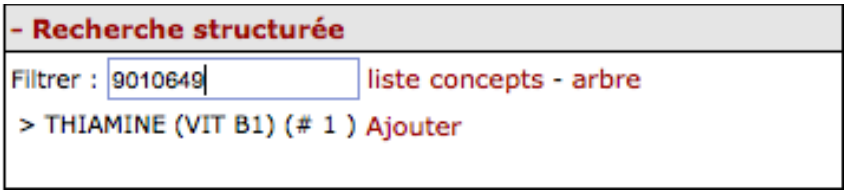

Figure 4 : Recherche via un code UCD. Ici, il est retourné la Thiamine, l'UCD recherché correspondant à celle d'une unité de dispensation de Bevitine®

4.1.3 Recherche par substance active

Une autre voie de recherche est explorée, permettant une recherche plus spécifique au médicament (figure 5).

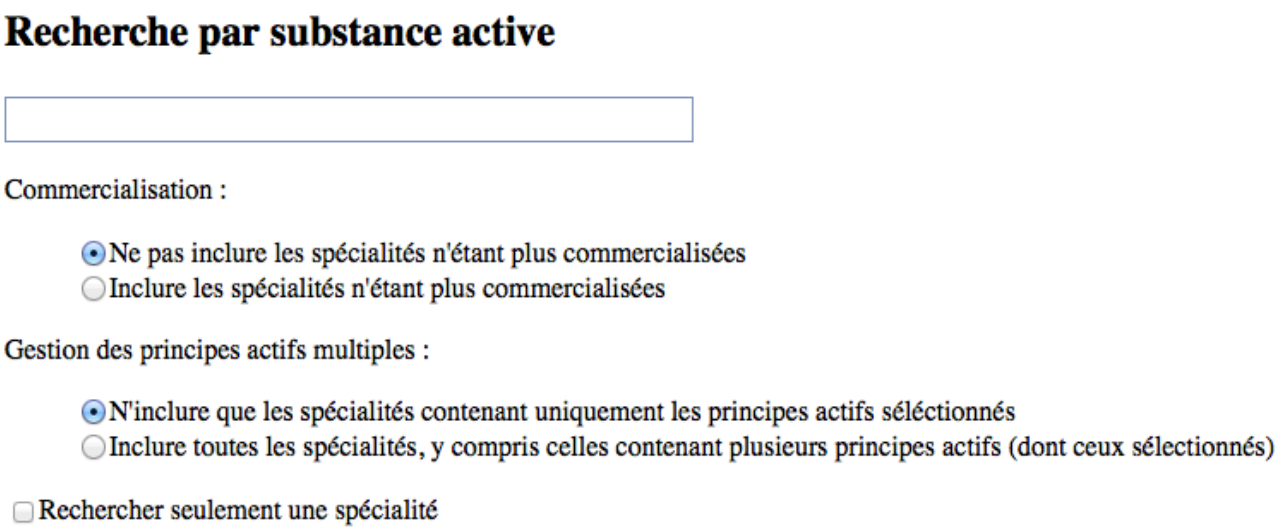

Figure 5 : Recherche de spécialités par entrée de substances actives

Ce mode de recherche gère l'autocomplétion, et permet une gestion des doublons. Des fonctionnalités supplémentaires telles que la recherche ou non de spécialités n'étant plus commercialisées sont par ailleurs implémentées.

L'affichage des spécialités correspondantes se fait à la volée dans un bloc de droite, permettant à l'utilisateur de vérifier les résultats correspondant à sa sélection, et éventuellement à exclure certaines spécialités renvoyées (figure 6).

Figure 6 : Ajout de substance active à la recherche. Si une même substance est sélectionnée plusieurs fois, les blocs fusionnent afin d'éviter les doublons. Une fois le terme ajouté, les spécialités correspondantes s'affichent dans un bloc dédié.

Les médicaments inclus dans la recherche sont *in fine* convertis en code UCD pour recherche.

La liste de l'ensemble des patients possédant un ou plusieurs documents contenant les termes recherchés sont retournés. L'affichage pourra se faire soit par prescription unique, soit par prescription agrégée sur l'ensemble du séjour du patient, par UF (unité fonctionnelle). Un exemple de résultats d'une recherche, avec affichage par prescription unique est présenté figure 7.

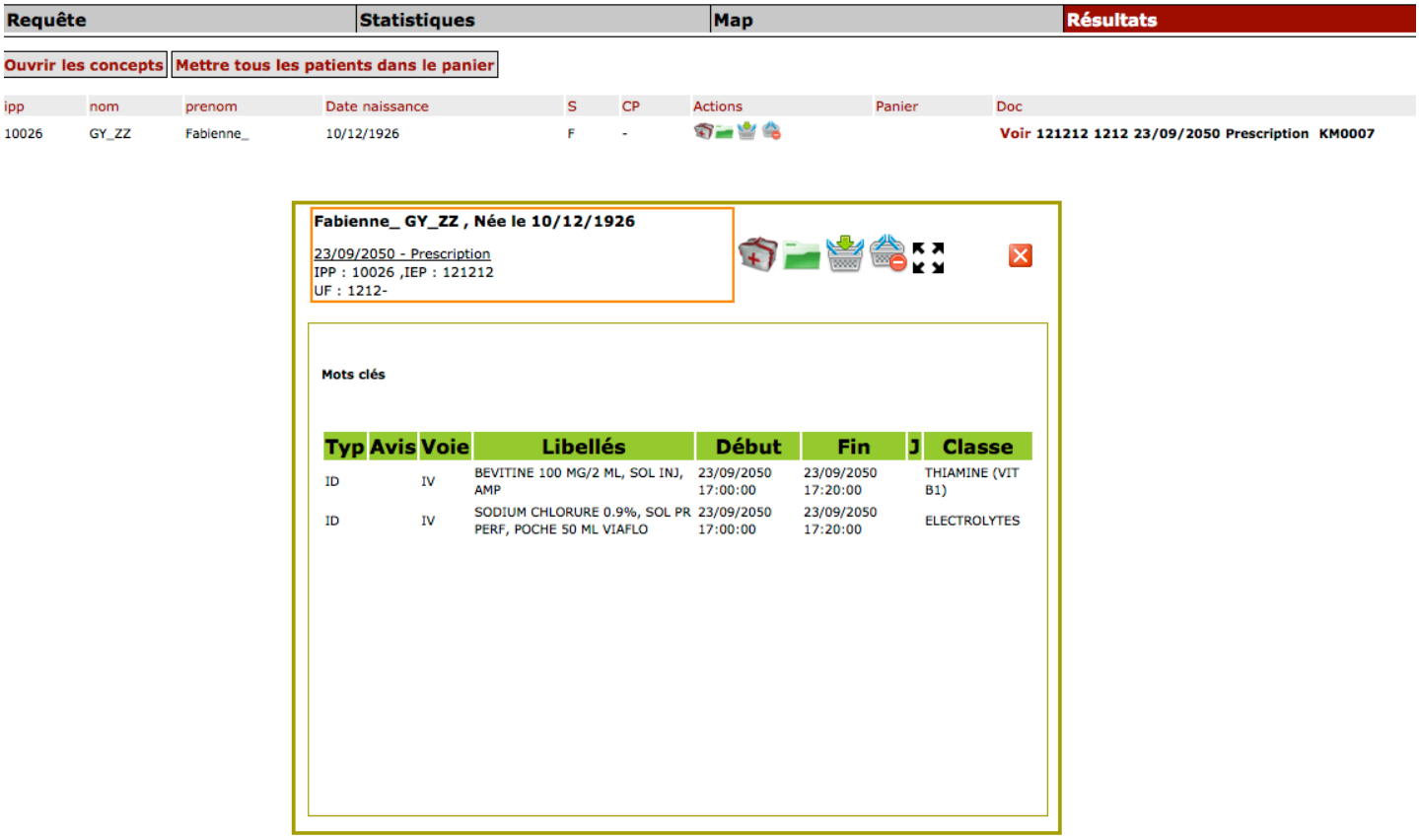

Figure 7 : Affichage des résultats d'une recherche (ici, sur « Thiamine »). Notons qu'il s'agit ici d'une prescription anonymisée.

5 Discussion

L'intégration et l'exploitation des données de prescription restent complexes. Ces données sont présentes dans le système d'information de façon hétérogène. L'exploitation des données structurées suppose l'utilisation d'un référentiel à la fois stable et mis à jour très régulièrement. Le choix de la CIOsp, lié intrinsèquement au format de flux PN13, nous permet d'avoir un système continuellement à jour et en parfaite adéquation avec le contenu des documents intégrés dans Roogle.

Actuellement, ce référentiel est proposé par PHAST sous la forme de fichier texte. Dans notre expérimentation, ce fichier a été parsé pour être intégré dans Roogle. L'évolution vers une approche webservice, comme le développe actuellement thériaque pour sa base, permettrait une mise à jour de ce référentiel plus efficiente.

L'intégration de données de prescription au format PN13 a été aisée. Ce format n'étant pas facilement lisible par un humain, nous avons développé un parseur permettant de visualiser les prescriptions de façon similaire à la façon dont elles étaient représentées dans le DPI DxCare.

Nous prévoyons d'intégrer également les données d'administration des médicaments, toujours au format PN13. Dans ce cas, la visualisation de ces informations sera plus complexe

puisqu'elle nécessitera de reconstituer l'équivalent d'une pancarte dans l'interface de Roogle.

L'intégration de la base Thériaque nous permet d'envisager l'exploitation des relations contenues dans la base. Dans notre système, Thériaque n'est pas utilisé comme composant d'intégration des données sources (pour cela, nous utilisons la CIOsp), mais comme une base de connaissance, utilisée pour la construction des requêtes et l'annotation de résultats.

Thériaque met à présent à disposition sa base sous forme de webservice. Nous étudions actuellement cette voie comme modèle de couplage entre Roogle et Thériaque, dont l'intérêt serait d'avoir un entrepôt de données exploitable à l'aide de connaissance sur le médicament à jour.

6 Conclusion

L'intégration des flux liés aux médicaments dans l'entrepôt de données ouvre le champ des possibles quant à ses utilisations, et apporte un outil supplémentaire aux différents métiers, notamment aux professionnels de la recherche clinique pour les études de faisabilité ainsi que pour la constitution de cohorte.

De nombreuses améliorations restent cependant à apporter, notamment au niveau performance et interface utilisateur. Il est enfin possible d'imaginer d'autres utilisations encore plus évoluées des données de prescriptions : recherche des coordonnées de patients en cas de rappels de lots, autocomplétion des e-CRF via un plugin connecté à Roogle, etc. Une évaluation de l'outil est prévue.

Remerciements

Nous remercions

- PHAST, pour nous avoir transmis la dernière distribution à jour,
- La CNHIM, pour nous avoir permis d'accéder à l'ensemble de la base Thériaque

Références

[1] F. F. Costa, « Big data in biomedicine », *Drug Discov. Today*.

[2] S. N. Murphy, M. E. Mendis, D. A. Berkowitz, I. Kohane, et H. C. Chueh, « Integration of Clinical and Genetic Data in the i2b2 Architecture », *AMIA. Annu. Symp. Proc.*, vol. 2006, p. 1040, 2006.

[3] M. Cuggia, N. Garcelon, B. Campillo-Gimenez, T. Bernicot, J.-F. Laurent, E. Garin, A. Happe, et R. Duvauferrier, « Roogle: an information retrieval engine for clinical data warehouse », *Stud Health Technol Inf.*, vol. 169, p. 584–588, 2011.

[4] M.-C. Husson, « Thériaque® : base de données indépendante sur le médicament, outil de bon usage pour les professionnels de santé », *Ann. Pharm. Fr.*, vol. 66, n° 5-6, p. 268-277, nov. 2008.

Adresse de correspondance

Dalleau Kevin, Laboratoire Traitement du Signal et de l'Image (LTSI) INSERM : U1099 – Université de Rennes 1, 2 avenue du Professeur Léon Bernard. 35043 *Rennes*

kevin.dalleau@gmail.com