SmartCRF, un prototype pour visualiser, rechercher et annoter les informations d'un dossier patient dans I2B2

Cossin Sébastien[1,2], Jouhet Vianney[1,2], Lebrun Luc[2], Niamkey Aymerick[2], Mougin Fleur[1], Lambert Mathieu[3], Diallo Gayo[1],Thiessard Frantz[1,2]

[1] EQUIPE DE RECHERCHE EN INFORMATIQUE APPLIQUÉE À LA SANTÉ (ERIAS), INSERM BPH U1219, Université de Bordeaux, F-33000 Bordeaux, France
sebastien.cossin@u-bordeaux.fr

[2] Informatique et Archivistique Médicales (IAM), Service d'Information Médicale, Pôle de Santé Publique, CHU de Bordeaux. France

[3] Service de médecine post-urgences, CHU de Bordeaux.

Résumé : Dans cet article, nous présentons SmartCRF, un prototype développé dans le cadre de l'entrepôt de données du CHU de Bordeaux permettant de visualiser, rechercher et annoter des informations d'un dossier patient informatisé. Cet outil permet d'aider les chercheurs à recueillir plus rapidement les données dont ils ont besoin pour une étude clinique. Les annotations réalisées pourront servir à entraîner des algorithmes d'apprentissage automatique pour faciliter la réutilisation secondaire des données.

Mots-clés : Electronic Health Records, User-Computer Interface, Information Storage and Retrieval, Semantics, Natural Language Processing, Data Warehousing, Data Curation

1 Contexte

L'entrepôt de données biomédicales du CHU de Bordeaux a été mis en place pour permettre la réutilisation secondaire des données du système d'information hospitalier (SIH) à des fins de recherche. Il intègre de nombreuses sources d'information d'un dossier patient informatisé (Programme de Médicalisation des Systèmes d'Information (PMSI), examens de biologie, prescriptions médicamenteuses, formulaires de saisie, compte-rendus de consultation et d'hospitalisation...).

Dans le cadre d'études rétrospectives menées sur ces données, l'une des étapes essentielle et préalable à l'analyse statistique est de recueillir et de structurer les informations pour chaque patient inclus. Cette collecte de données est chronophage et complexe en l'absence d'interfaces et d'outils dédiés (Pan & Cimino, 2014).

Des outils d'indexation sémantique, de recherche et de visualisation peuvent aider les médecins et les chercheurs à trouver rapidement des informations dans un dossier patient informatisé (DPI) (Thiessard *et al.*, 2012).

Notre objectif a été de réaliser un prototype fonctionnel permettant de visualiser l'ensemble des données d'un DPI, mais également de rechercher et de collecter des données pour une étude.

2 Fonctionnement

L'entrepôt du CHU de Bordeaux utilise la solution open source I2B2 (Murphy *et al.*, 2006). I2B2 possède un modèle de base de données relationnelles en étoile. Dans ce modèle, chaque donnée est reliée à un patient, à une venue à l'hôpital (consultation ou hospitalisation) et à une source de données (PMSI, biologie...). Chaque donnée possède une date (quand la donnée a été enregistrée) et un type (textuel ou structuré).

Le prototype a été développé avec le paquet Shiny du langage de programmation R. Ce dernier permet de réaliser des interfaces web interactives avec des bibliothèques JavaScript.

Pour visualiser un dossier patient, les données sont d'abord chargées en mémoire côté serveur (RShiny) en interrogeant I2B2. Quatre fonctionnalités, présentées ci-dessous, permettent d'interagir avec les données : une timeline, un moteur de recherche, un nuage de mots et un module d'annotation des éléments d'intérêt.

2.1 Timeline

Les données d'un patient sont affichées sur une timeline interactive permettant d'obtenir une vue d'ensemble des données d'un patient (figure 1).

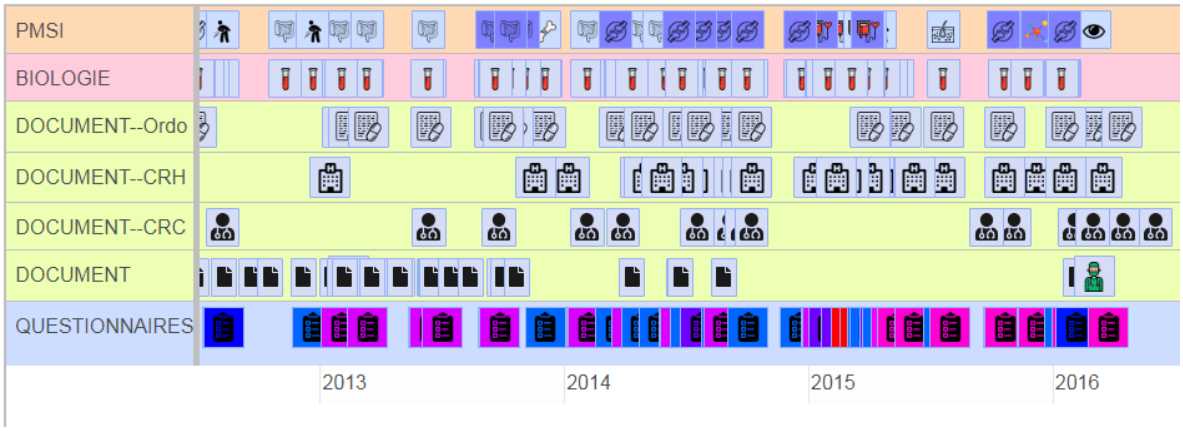

FIGURE *1 – Timeline interactive permettant de visualiser l'ensemble des données d'un patient. Chaque icône est cliquable afin d'afficher les données associées à l'événement sélectionné.*

La timeline contient plusieurs groupes (PMSI, biologie...) qui correspondent à différentes sources de données. Chaque icône de la timeline contient une à plusieurs données du patient. Par exemple, tous les résultats biologiques d'une prise de sang sont regroupés sous une même icône. Chaque icône est cliquable permettant d'afficher le détail de son contenu.

L'utilisateur peut ajouter dynamiquement des sous-groupes à la timeline. Il peut par exemple ajouter le sous-groupe "plaquettes" au groupe "Biologie" pour visualiser rapidement les résultats des plaquettes au cours du temps.

2.2 Moteur de recherche

Les syntagmes nominaux sont extraits par un programme Java. Ils correspondent aux principales informations d'un document textuel (Kathait *et al.*, 2017). Ceux-ci sont normalisés (transformation du texte en minuscule, retrait des accents et des caractères spéciaux) et lemmatisés avec TreeTagger (Schmid, 1997) avant d'être indexés avec ElasticSearch™. L'autocomplétion du moteur de recherche utilise ces syntagmes nominaux pour guider l'utilisateur dans sa recherche (figure 2). Les résultats d'une recherche sont affichés par ordre antéchronologique. Seule la phrase contenant le terme recherché est affichée et l'utilisateur peut accéder au document dans son intégralité s'il le souhaite.

2.3 Nuage de mots

Certains concepts comme les symptômes, les maladies et les médicaments sont détectés dans le texte libre en utilisant des terminologies issues de l'UMLS (Bodenreider, 2004). Les concepts détectés sont affichés à l'utilisateur dans un nuage de mots cliquable (figure 3). Un clic conduit à placer le terme dans la barre de recherche précédente et à trouver les documents le mentionnant.

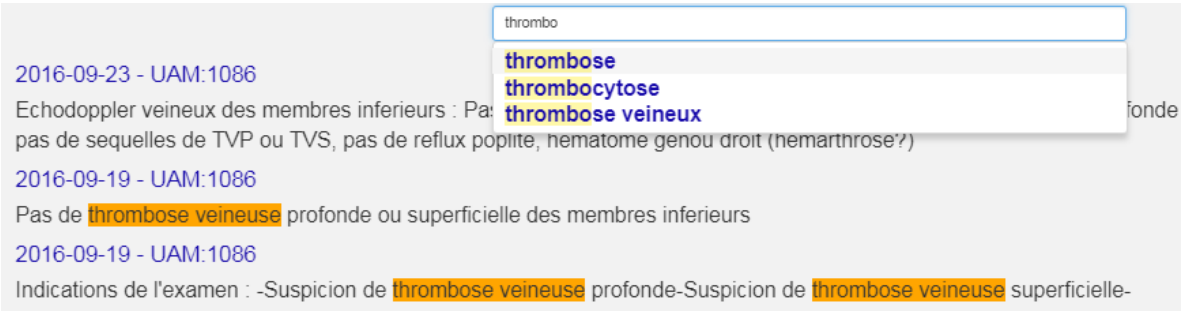

FIGURE 2 – *Moteur de recherche avec autocomplétion (syntagmes nominaux normalisés et lemmatisés) permettant de rechercher des informations dans les données en texte libre d'un dossier patient informatisé. Seule la phrase contenant le terme recherché est affichée et l'utilisateur peut accéder au document dans son intégralité. UAM : service hospitalier*

FIGURE 3 – *Nuage de mots permettant de visualiser les symptômes et les maladies détectés dans les documents en texte libre. Un clic sur un terme permet de lancer une recherche.*

2.4 Annotation

L'utilisateur peut surligner les informations pertinentes dans le dossier pour tracer et enregistrer les éléments lui ayant permis de structurer l'information. Par exemple, il enregistre le statut tabagique "non fumeur" et justifie son choix en sélectionnant le passage "pas de tabagisme" dans le dossier (figure 4). Ces annotations pourront être utilisées par la machine pour apprendre à trouver l'information du statut tabagique dans d'autres dossiers.

3 Discussion

Le prototype est en phase de tests avec des utilisateurs.

Nous réfléchissons à la manière d'implémenter un système de recommandation visant à présenter les éléments d'information susceptibles d'intéresser le chercheur.

Par exemple, si le chercheur a besoin de collecter l'information sur le statut tabagique, le

FIGURE *4 – L'utilisateur classifie le patient comme "Non Fumeur" et enregistre les éléments du dossier lui ayant permis de faire son choix ("Pas de tabagisme").*

nuage de mots devrait afficher les termes en lien avec ce concept (tabac, tabagisme, cigarette, paquet-année, nicotine...) trouvés dans le dossier et les éléments de la timeline concernés pourraient être mis en exergue.
Le système de recommandation devra apprendre à rechercher et à visualiser les informations pertinentes en fonction des annotations déjà réalisées. Le chercheur sera enclin à annoter les termes pertinents à sa recherche pour que la machine l'aide dans ses recherches futures, créant ainsi une interaction homme-machine vertueuse.
Un jeu de données annotées dans le cadre d'une étude pourra servir à entraîner des algorithmes de classification, comme par exemple prédire le statut tabagique d'un patient. Les annotations par un expert du domaine étant habituellement coûteuses à obtenir, le recueil de données dans le cadre d'études cliniques offre une formidable opportunité pour entraîner les algorithmes de machine learning. Ces algorithmes entraînés pourront être utilisés pour faciliter la réutilisation secondaire des données de l'entrepôt de données du CHU de Bordeaux.

Références

BODENREIDER O. (2004). The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–270.

KATHAIT S. S., TIWARI S., VARSHNEY A., SHARMA A., MESTROVIC A., GELBUKH A. F., RENDON E., GARCÍA-HERNÁNDEZ, DA CRUZ R. P., ARNULFO R., MONTIEL R., LEDENEVA Y., GUTWIN C., NEVILL-MANNING C. G., FRANK E., PAYNTER G. W., HASAN K. S., NG V., CARAGEA C., GOLLAPALLI S. D., ALHAKIM A., DUTTA B., GIUNCHIGLIA F., LOPER E., KLEIN E., GODEA A. & BULGAROV F. A. (2017). Unsupervised key-phrase extraction using noun phrases. *International Journal of Computer Applications*, **162**.

MURPHY S. N., MENDIS M. E., BERKOWITZ D. A., KOHANE I. & CHUEH H. C. (2006). Integration of clinical and genetic data in the i2b2 architecture. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, p. 1040.

PAN X. & CIMINO J. J. (2014). Locating Relevant Patient Information in Electronic Health Record Data Using Representations of Clinical Concepts and Database Structures. *AMIA Annual Symposium Proceedings*, **2014**, 969–975.

SCHMID H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. JONES & H. SOMERS, Eds., *New Methods in Language Processing*, Studies in Computational Linguistics, p. 154–164. London, GB : UCL Press.

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. (2012). RAVEL : retrieval and visualization in ELectronic health records. *Studies in Health Technology and Informatics*, **180**, 194–198.