Un modèle sémantique d'identification du médicament en France

J. Grosjean, C Letord, Jean Charlet, X Aimé, L Danès, J. Rio, I Zana, Stéfan J. Darmoni, C Duclos

▶ **To cite this version:**

J. Grosjean, C Letord, Jean Charlet, X Aimé, L Danès, et al.. Un modèle sémantique d'identification du médicament en France. Atelier IA & Santé 2019, Fleur Mougin; Brigitte Séroussi, Jul 2019, Toulouse, France. hal-02917172

HAL Id: hal-02917172
https://hal.science/hal-02917172

Submitted on 18 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un modèle sémantique d'identification du médicament en France

J. Grosjean[1,2], C. Letord[1,2], J. Charlet[2,3], X. Aimé[2], L. Danès[2], J. Rio[1], I. Zana[2], SJ. Darmoni[1,2], C. Duclos[2]

[1] Department of Biomedical Informatics, Rouen University Hospital, 76031 Rouen Cedex, France;
{Julien.Grosjean,Catherine.Letord,Julien.Rio,Stefan.Darmoni}@chu-rouen.fr
[2] Sorbonne Université, INSERM, Université Paris 13, LIMICS, Paris, France;
xavier.aime@cogsonomy.fr,loane.danes@agroparistech.fr,ilan.zana26@gmail.com
[3] Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France
{Jean.Charlet,Catherine.Duclos}@aphp.fr}

Résumé : il n'existe pas de standard universellement accepté pour nommer les médicaments. L'identification du médicament a fait l'objet de nombreux travaux de normalisation. Notre objectif est de définir un modèle formel du médicament en français pour lier les différentes entités manipulables autour du médicament. Ce modèle formel vise un double sous-objectif : (a) créer et instancier une ontologie formelle du médicament ; (b) créer une terminologie du médicament, intégrable dans un serveur de terminologies.

Mots-clés : ontologie ; terminologie ; serveur de terminologie ; modèle formel ; médicament.

1 Introduction

Bien que le médicament soit fini, identifiable, il n'existe pas de standard universellement accepté pour les nommer (Cimino 1999). Selon le point de vue auquel on se place, on peut le définir à un niveau moléculaire comme une substance active, à un niveau clinique comme un produit capable de traiter une pathologie, à un niveau physique comme une présentation destinée à satisfaire la prescription et délivrable au patient. L'identification du médicament peut donc se concevoir à divers niveaux ayant des degrés d'abstraction plus ou moins grand (Sperzel 1998) (i) une présentation, un produit manufacturé ou un ingrédient correspondent à des objets physiques, (ii) un produit clinique ou une fraction thérapeutique sont de pures abstractions.

L'identification du médicament a fait l'objet de nombreux travaux de normalisation dont les plus récents définissent l'identification du produit médicinal et du produit pharmaceutique (ISO 11615, ISO 11616, ISO 20443, ISO 11238, ISO 11239, ISO11240) afin de rendre le partage international de l'information sur le médicament possible. Le référentiel résultant sera disponible en 2020 au niveau de l'agence européenne. D'autres modèles ont été adoptés pour représenter le médicament dans les bases de données sur le médicament (Broverman 1998, DMD 2008...), ou pour servir de pivot entre des bases des données sur le médicament (RxNorm – https://www.nlm.nih.gov/research/umls/rxnorm/). Dans ces modèles, se retrouve cette dualité entre virtuel et réalité et les relations de composition entre ingrédients actifs, dosage et forme.

Ces référentiels « normalisés », lorsqu'ils sont disponibles, n'incluent pas de médicaments français. L'agence nationale de sécurité du médicament et des produits de santé (ANSM) met à disposition, via la Base de Données Publique du Médicament (BDPM – http://base-donnees-publique.medicaments.gouv.fr/), des fichiers décrivant les médicaments français et leur composition mais ne respectent pas toujours les normes d'identification du médicament qui en permettrait un usage simple. Une équipe bordelaise a proposé une transformation de ces fichiers

dans le format RxNorm afin de disposer de nombreux variants dénommant les médicaments pour rechercher ces derniers dans des comptes rendus textuels de passage aux urgences (Cossin 2018).

Dans ce travail, nous proposons et présentons un modèle formel du médicament en français pour lier les différentes entités manipulables autour du médicament. Ce modèle sera disponible gratuitement pour la communauté scientifique. Ce modèle formel vise un double sous-objectif : (a) créer et instancier une ontologie formelle du médicament, en s'appuyant sur des données librement accessibles, en particulier celles fournies par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM) via la BDPM et celles issues des bases de données bibliographiques, comme PubMed ou LiSSa ; (b) créer une terminologie du médicament, intégrable dans un serveur de terminologie. Ce modèle servira en première intention au projet PsyHAMM, dont le but est de détecter des prescriptions hors AMM (Autorisation de Mise sur le Marché) en psychiatrie (https://anr.fr/Projet-ANR-18-CE19-0017). Idéalement, ce modèle pourra également être utilisé pour rechercher une information de qualité sur les médicaments dans les différents entrepôts de données développés en France.

Dans la prochaine section nous décrivons toutes les actions de normalisations et les difficultés rencontrées. Dans la section 3, nous décrivons la méthodologie de constructions des modèles et nous donnons un certain nombre de résultats dans la section 4. Nous terminons par quelques perspectives en section 5.

2 Définitions, référentiels et informations sur le médicament

L'analyse des modèles existants permet de repérer les concepts retrouvés pour identifier le médicament, à savoir la substance active, l'excipient, la dose, la forme, la voie d'administration, le nom commercial, le conditionnement, et la combinaison de ces concepts permettant de définir des produit cliniques, pharmaceutiques, médicinaux, et présentés. Ces concepts peuvent être décrits dans des terminologies comme l'ATC, le MeSH (http://www.nlm.nih.gov/mesh), la SNOMED (https://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-snomed-ct) ou encore l'UMLS. Des référentiels normatifs existent aussi pour représenter les formes, voies d'administration (Standard Terms de l'EMA), les unités (UCUM). Par ailleurs une norme française d'interopérabilité (NF 97-555) définit en France des identifiants de médicaments (CIS, UCD, CIP), les liens entre eux et l'attendu quant à la construction de la dénomination d'une spécialité pharmaceutique.

La Banque Publique du Médicament met à disposition les fichiers (a) de spécialités pharmaceutiques associant un identifiant CIS à une dénomination (devant normalement être construite selon le modèle nom de marque, dosage, forme), (b) de composition permettant de connaître la composition de la spécialité pharmaceutique en quantité de substance active et en fraction thérapeutique, (c) des présentations associant un/des identifiant(s) CIP à une/plusieurs présentations elles même reliées à une spécialité pharmaceutique. Ces informations sont associables à un résumé des caractéristiques du produit (RCP). Le RCP est une annexe de la décision d'autorisation synthétisant les informations notamment sur les indications thérapeutiques, contre-indications, modalités d'utilisation et les effets indésirables d'un médicament. Par exemple, « PROZAC 20 mg, gélule » est le libellé d'une spécialité pharmaceutique qui a pour code CIS 61885224. Elle contient de la « fluoxétine » (substance) sous forme de « chlorhydrate de fluoxétine » (un sel possible de cette substance). Sa vente est autorisée sous la présentation « plaquette(s) thermoformée(s) PVC-Aluminium de 14 gélules ». Cette présentation est référencée par un code identifiant de présentation (CIP) à 13 chiffres figurant sur la boîte du médicament.

Enfin il existe un identifiant (code UCD) correspondant au plus petit élément commun à plusieurs présentations d'une même spécialité pharmaceutique. Le code UCD représente, pour chaque forme galénique, la plus petite unité de dispensation (comprimé, flacon, ...). Le code UCD et le nombre d'UCD par présentation sont administrés par le Club Inter Pharmaceutique.

(source : https://www.has-sante.fr/portail/jcms/c_671889/fr/certification-des-logiciels-d-aide-a-la-prescription-lap#U)

La description du médicament par la BDMP est limitée par rapport aux attendus des différents modèles d'identification du médicament. Par exemple, le terme « PROZAC » n'est pas présent dans cette base, alors que « PROZAC 20 mg, gélule » l'est. Nous avons créé des racines pharmaceutiques correspondant aux différents noms commerciaux (ici « PROZAC ») présents de la BDPM. Ces ajouts permettent d'améliorer le rappel quand on applique un annotateur sémantique pour la détection des médicaments dans un document de santé. Cela permet également de regrouper les CIS ayant la même racine.

Plus récemment, quatre bases de données médicamenteuses labellisées par la Haute Autorité de Santé (HAS) se sont associées pour créer Medicabase, une base de données de médicaments virtuels. Les médicaments virtuels permettent de regrouper des spécialités qui comportent :

- le ou les même(s) principe(s) actif(s) ou des sels du ou des principe(s) actif(s) cliniquement équivalent(s) du point de vue des risques iatrogènes ;
- les mêmes dosages en base active des principes actifs ;
- une forme galénique considérée comme cliniquement équivalente du point de vue des risques iatrogènes.

Par exemple, l'« Abacavir 300 mg comprimé » regroupe l'« ABACAVIR MYLAN 300 mg, comprimé pelliculé sécable » et l'« ABACAVIR SANDOZ 300 mg, comprimé pelliculé sécable », qui sont des spécialités équivalentes.

Par ailleurs, au sein des fichiers de la BDPM, les voies d'administration semblent normalisées, au contraire des formes pharmaceutiques qui ne semblent pas respecter un modèle précis, encore moins les conditionnements (Cf. *infra*).

Le D2IM travaille depuis 10 ans (Pereira, 2008) sur le médicament afin d'intégrer le plus d'informations structurées concernant ce dernier au sein de son serveur terminologique HeTOP (Grosjean, 2012). Mais, il manquait à ce travail un modèle formel permettant par exemple, de retrouver tous les codes CIP ou UCD pour une substance active donnée ou pour un médicament virtuel donné. Pour finir, le LIMICS a initialisé en 2016 une ontologie sur le médicament mais le projet s'est heurté à de nombreux problèmes de cohérences et de qualité des bases utilisées pour construire cette ontologie (Steinberg, 2016).

3 Processus de construction de l'ontologie et de la terminologie des médicaments

A partir du travail des équipes impliquées et au regard des problèmes de qualité des fichiers d'origine, nous avons utilisé le travail ontologique comme un test de la cohérence des données issues de la BDPM et nous avons mis au point un processus de construction d'une base terminologique et d'une ontologie qui se déroule en plusieurs étapes avec un outil d'ETL (Talend) appliqué aux fichiers disponibles :

1. création d'une ontologie de départ avec un certain nombre de classes normalisées :
    a. intégration de l'ATC (https://www.whocc.no/atc_ddd_index/),
    b. hiérarchie des voies d'administration développées par un pharmacien,
    c. création des classes correspondant aux conditions d'utilisation des médicaments (liste à faire évoluer en fonction des évolutions (rares) de ces conditions),
    d. création des classes propres au modèle du médicament de l'ANSM (nom commercial, substance active, générique, …)

2. récupération des fichiers publics issus de l'ANSM, de Medicabase, et de l'ATC. Certains codes ATC sont corrigés en avance de phase par le D2IM pour les médicaments les plus récents. La table 1 résume l'ensemble des fichiers utilisés ;

3. vérifications de la cohérence des fichiers avec des tests. La table 2 liste les principaux tests de cohérence réalisés ;
4. création de l'ontologie de travail instanciant le modèle de l'ontologie de départ avec les fichiers décrits plus haut, en incluant la création de hiérarchies, d'une part pour les conditionnements et d'autre part pour les formes ; les deux étant issus des fichiers de l'ANSM ;
5. fourniture de l'ontologie de travail pour être intégrée et maintenue (processus manuel d'assurance qualité et processus automatique de contrôles d'intégrités) au sein du serveur terminologique d'HeTOP.
6. récupération de fichiers corrigés issus de HeTOP et génération de l'ontologie finale.

CIS.txt	Donne la liste des spécialités et de leurs détails.
COMPO.txt	Donne la liste des molécules, substance active ou fraction thérapeutique, et les spécialités qu'elles composent.
CIS_CIP.txt	Donne la liste des présentations (codes CIP) correspondant à chaque spécialité (code CIS) et leurs détails.
cis_cpd_bdpm.txt	Donne pour chaque spécialité la condition d'utilisation.
cis_gener_bdpm.txt	Donne pour chaque générique, la ou les catégories auxquelles il appartient, sa dénomination et sa/ses spécialités correspondantes.
Autorisations actives - ATC_mdts présents sur RSP.xls	Pour tous les médicaments possédant une AMM donne la relation entre le code ATC et chaque spécialité.
fic01den.txt	Donne la liste des groupes génériques et des DIC correspondantes.
fic02grp.txt	Donne la liste des génériques, le groupe auquel chacun appartient, sa voie d'administration et sa dénomination.
fic03spe.txt	Donne pour chaque spécialité, la catégorie à laquelle elle appartient, son générique correspondant et sa dénomination.
liste_medicamentVirtuels	Donne la liste des médicaments virtuels.
ATC_2018.owl	Donne l'arborescence ATC sous forme de fichier OWL avec les labels normalisés.

Table 1 – liste des principaux fichiers utilisés.

4 Résultats

Le modèle d'identification du médicament en français est détaillé dans la Figure 1. *N*ous avons travaillé à sa création, avec dès le départ, un test sur plusieurs cas d'usage, en démarrant par les plus simples (en évitant pour l'instant les associations médicamenteuses complexes et les médicaments biologiques). Ce modèle tente de minimiser les relations entre les différents concepts du médicament. Le reste des relations se calculent, comme par exemple la relation entre un médicament virtuel et un code ATC se déduit par les deux relations entre médicament virtuel et spécialité pharmaceutique d'une part, et spécialité pharmaceutique et code ATC d'autre part.

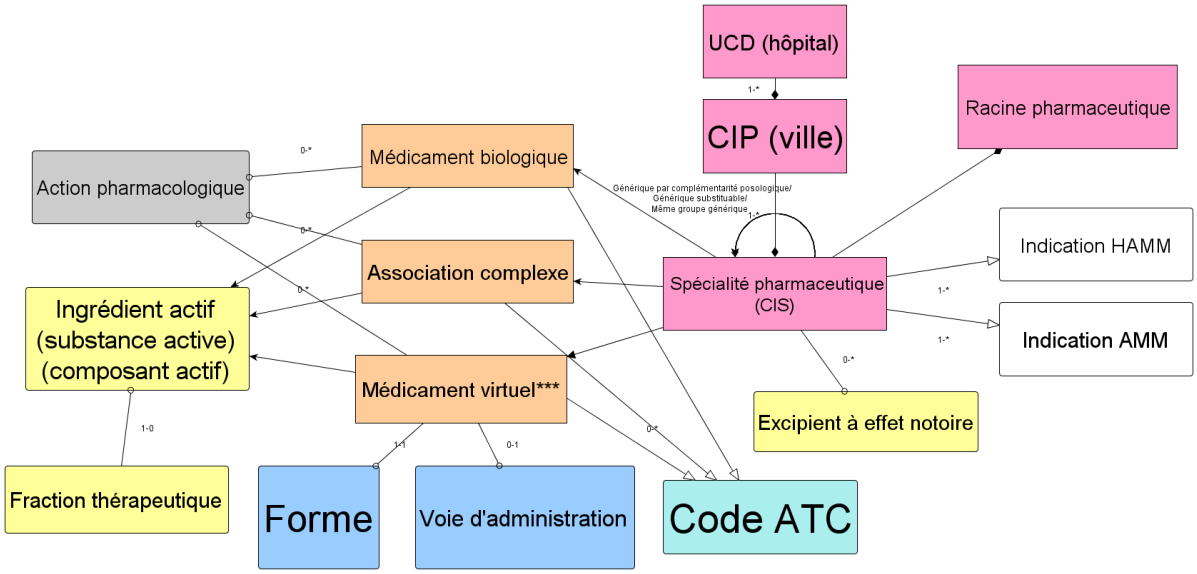

Figure 1 : modèle sémantique du médicament en France. La plupart de ces entités correspondent déjà à des ensembles de codes ou nomenclatures qu'il est nécessaire de mettre en correspondance. Les liens sémantiques sont définis explicitement ainsi que les cardinalités associées. Ces liens sont définis de la façon la "plus haute" possible afin de ne pas répéter l'information. Ce modèle est à la fois instancié sous forme de terminologie et d'ontologie pour des utilisations variées.

Ce modèle permet de retrouver un médicament dans un entrepôt de données, sous n'importe quelle de ses formes : qu'il s'agisse de son nom commercial ou de ses ingrédients, son code ATC ou ses codes CIS, CIP ou UCD. Par exemple, pour la substance active (ou ingrédient) modafinil, le modèle nous fournit les informations suivantes : nom commercial = MODAFINIL; code ATC = N06BA07 ; quatre codes UCD (correspondant à quatre génériques) = 3400893594582, 3400893615676, 3400893926680, 3400894100188 ; le seul médicament virtuel actuellement est : Modafinil 100 mg comprimé. De plus, en utilisant HeTOP, nous pouvons remarquer que cette substance active existe dans les termino-ontologies suivantes : MeSH, NCIT (US National Cancer Institute, SNOMED CT, LOINC (Logical Observations Identifiers Names and Codes), BNPC (Base Nationale (Française) des Produits et des Compositions).

À partir de l'ontologie de travail, nous avons réalisé huit principaux tests de cohérence, dont le détail est fourni dans la table1. Deux tests de cohérence ont détecté de nombreuses erreurs : 6 554 codes CIS n'ont pas de code CIP ; 509 codes CIS n'ont pas de codes ATC .

Unfound_CIS_Compo.xls	Recense les codes CIS présents dans le fichier CIS.txt mais absents du fichier COMPO.txt.	2
Unfound_CIS_CisCip.xls	Recense les codes CIS présents dans le fichier CIS.txt mais absents du fichier CIS_CIP.txt.	5
Without_Libelle_Cip13.xls	Recense les codes CIS du fichier CIS_CIP.txt qui n'ont pas de code CIP13 et/ou de libellé.	6554
Unfound_CIS_ATC.xls	Recense les médicaments associés à un code CIS dans le fichier CIS.txt qui ne sont pas présents dans le fichier ATC et qui donc n'ont pas de code ATC associé.	509
Unfound_ATC_Onto.xls	Recense les codes ATC associés à des codes CIS qui ne sont pas présents dans l'arborescence de l'ontologie (même après mise à jour).	15

Unfound_Fic3_Gener.xls	Recense la liste des codes CIS qui sont présents dans les fichiers des spécialités fic03spe.txt mais qui ne sont pas présents dans le fichier	118
Unfound_Fic3_CIS.xls	Recense la liste des codes CIS qui sont présents dans les fichiers des spécialités *fic03spe.txt*.	40
Voies_manquantes.txt (resp. voies_manquantes_fic.txt)	liste des voies qui sont présentes dans le fichier *CIS.txt* (resp. *fic02grp.txt*) mais qui sont absentes de l'ontologie.	1

Table 2 – principaux tests de cohérence et fichiers de contrôle générés, associés au nombre d'occurrences en janvier 2019.

Les conditionnements ne sont pas normalisés dans les fichiers de la BDPM. Le travail de l'ETL sur les conditionnements permet de normaliser les chaînes de caractère puis créer une hiérarchie. Ces conditionnements sont enregistrés sous formes de libellés associés aux classes en relations d'hyperonymie. Par exemple, on a une classe dont le libellé est « boîte », elle subsume un classe dont le libellé est « boite aluminium ». Cette dernière subsume une classe dont le libellé est « boite aluminium comprime ». Enfin, on trouve en dessous les classes correspondant à six conditionnements déclarés dans les fichiers, « 1 boite aluminium de 6/10/12/20/25/1000 comprimés ». En février 2019, il y a 25 661 chaînes de caractères différentes décrivant des conditionnements qui correspondent à autant de classes organisées en une hiérarchie dont le premier niveau comporte 99 classes et qui a une profondeur moyenne de 9.

Les formes ne sont pas non plus normalisées. Un travail du même type que les conditionnements a amené l'explicitation de 597 formes organisées en une hiérarchie dont le premier niveau comporte 100 classes et qui a une profondeur moyenne de 3,3. Il existe une liste anglo-saxonne de formes sur laquelle nous allons travailler pour améliorer et normaliser cette hiérarchie.

Une instanciation d'HeTOP a été développée pour afficher et synthétiser l'ensemble des informations concernant le médicament en se fondant sur le modèle présenté dans ce travail. Cet « HeTOP médicament » est mise à disposition des utilisateurs qui ne sont pas experts du médicament dans une version compacte et simplifiée (https://www.hetop.eu/hetop/medicaments).

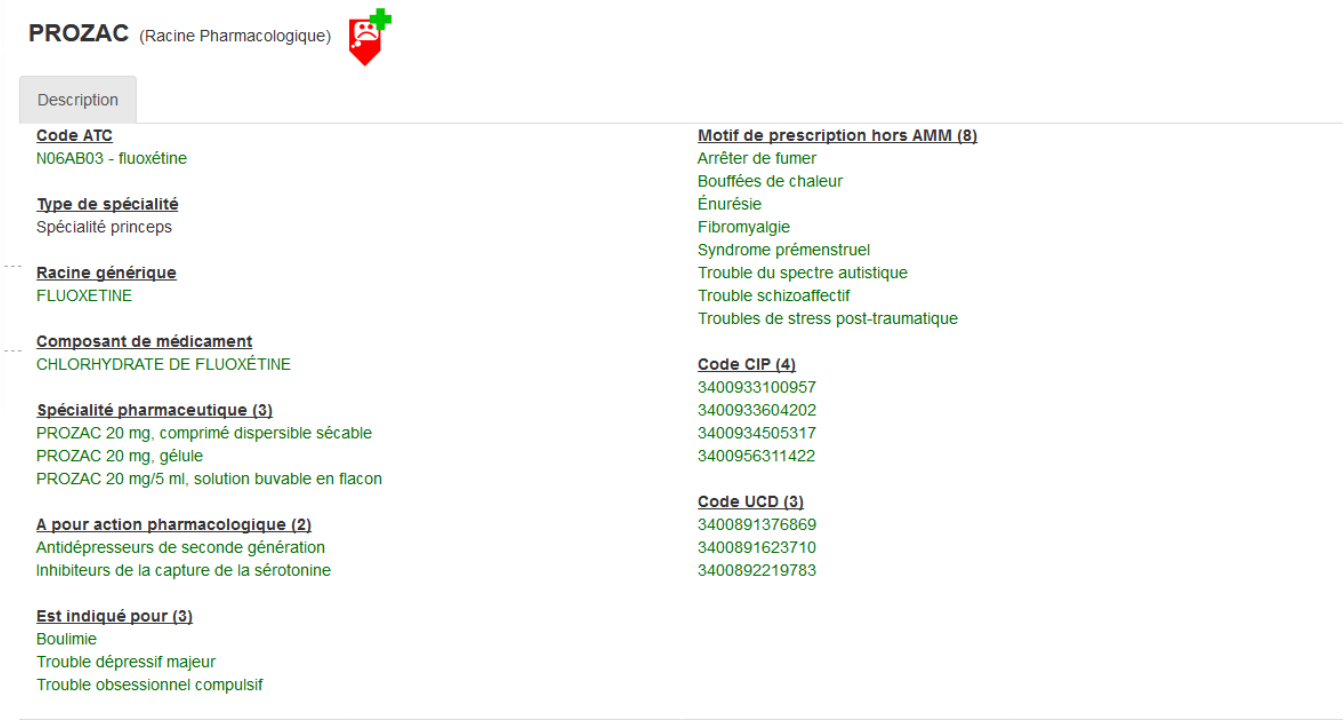

Figure 2 : exemple des informations sur un médicament fournies par le serveur terminologique « HeTOP médicaments » sur ECMT

5 Discussion

Notre travail a démontré qu'il était possible à partir des fichiers de l'ANSM de construire un référentiel ontologique et terminologique moyennant la correction de nombreuses erreurs ou oublis qui sont étonnantes puisque ces fichiers doivent se conformer à une norme et sont issus d'une autorité de référence.

Le développement d'un modèle ontologique et terminologique de façon coordonnée permet d'avoir deux versions des mêmes connaissances. Le couplage termino-ontologique permet de créer un cycle de qualité, en détectant certaines incohérences. La version terminologique, incluse dans HeTOP pourra être immédiatement évaluée dans le contexte de l'annotateur sémantique ECMT (Cabot, 2016). La version ontologique sera utilisée dans un annotateur sémantique du LIMICS, d'abord dans le contexte du projet PARON (Cardoso, 2018). Le format ontologique permettra de mettre à disposition des versions réduites de l'ontologie en fonction des contextes d'usage : il suffira de préparer les versions nécessaires avec de requêtes SPARQL.

Un projet proche est le projet ROMEDI (Cossin, 2018), qui vise la détection de médicaments en texte libre. L'équipe de Bordeaux a produit un site Web (URL : http://www.romedi.fr/), qui fournit de nombreuses informations intéressantes sur le médicament et qui permet maintenant de récupérer une ontologie représentant un certain nombre d'informations du site. Dès qu'elles seront bien stabilisées, les ressources mises à disposition par les 2 projets pourraient être comparées.

Par la suite, un certain nombre d'étapes vont suivre :

- Notre modèle doit, en premier lieu, être validé par des médecins et des pharmaciens du consortium PSYHAMM n'ayant pas participé à sa mise au point. En particulier, les cas difficiles, comme les associations complexes et les médicaments biologiques, seront étudiés en priorité.

- À notre connaissance, les médicaments en Europe ne seront définis selon la nouvelle norme IDPM en 2022. De nouveaux identifiants seront fournis par IDPM et seront aisément intégrables dans notre modèle formel.
- Concernant les indications, il existe des informations en texte libre dans la BDPM et structurées dans les bases de données françaises labellisées par la HAS. L'intégration de ces informations au sein de la base terminologique puis de l'ontologie est au programme de l'équipe dans le cadre du projet PsyHAMM. On voit dans la figure 2, l'affichage d'informations importantes pour le projet, à savoir les motifs de prescriptions hors AMM. Ceux-ci seront retravaillés et complétés.

En conclusion, notre consortium a réalisé un modèle formel du médicament, avec un couplage termino-ontologique permettant un cycle de qualité. Les retombées de ce modèle sont nombreuses : en premier lieu, la possibilité de rechercher une information sur le médicament dans les entrepôts de données de santé en France, cette information pouvant être exprimée de multiple façons, grâce à une vision multi-terminologique.

Remerciements

Ce travail a été réalisé dans le cadre du projet PSYHAMM financé par l'Agence Nationale de la Recherche (https://anr.fr/Projet-ANR-18-CE19-0017)**.**

Références

CABOT C.; LELONG R, GROSJEAN J., SOUALMIA L. F. & DARMONI, S. J (2016) Retrieving Clinical and Omic Data from Electronic Health Records.. Stud Health Technol Inform School: 2016; 221, 115.

Cardoso S., Aimé X., Meininger V., Grabli D., Melo Mora L.F., Bretonnel C.K., and Charlet J. (2018), A Modular Ontology for Modeling Service Provision in a Communication Network for Coordination of Care, Stud. Health Technol. Inform. (2018) 890–894. doi:10.3233/978-1-61499-852-5-890.

Cossin S, Loustau R, Jouhet V, Létinier L, Mougin F, Evrard G, Gil-Jardiné C, Diallo G, Thiessard F (2018). ROMEDI, une terminologie médicale française pour la détection des médicaments en texte libre.

Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, Kilbourne J, Lincoln MJ (2008). Exchange of computable patient data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): terminology mediation strategy. J Am Med Inform Assoc. 2008 Mar-Apr;15(2):174-83.

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.

Pereira, S; Plaisantin, B; Korchi, M; Rozanes, N; Serrot, E; Joubert, M & Darmoni, SJ. Automatic construction of dictionnaries, application to product characteristics indexing. Stud Health Technol Inform 2009;150: 512-6.

Steinberg K. (2016). Qualité des données de santé disponibles en France et de leurs modèles - Comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ? Mémoire CNAM, Titre professionnel niveau 1 Chef de projet en ingénierie documentaire et gestion des connaissances. 2016. Disponible à http://portaildoc-intd.cnam.fr/Record.htm?idlist=1&record=19298817124910160999