SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique

Clement Jonquet, Amina Annane, Khedidja Bouarech, Vincent Emonet, Soumia Melzi

SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique

SIFR BioPortal: French biomedical ontologies and terminologies available for semantic annotation

Clement Jonquet,*[1]Amina Annane,[2] Khedidja Bouarech,[1] Vincent Emonet,[1] Soumia Melzi[1]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), Université de Montpellier & CNRS, France
** Center for BioMedical Informatics Research (BMIR), Université de Stanford, USA*
[2] Ecole nationale Supérieure d'Informatique (ESI), Alger, Algérie

Résumé

Contexte – *En dépit d'une large adoption de l'anglais, une quantité significative des données en biomédecine est en français. Outre l'existence de nombreuses ressources en anglais, il y a beaucoup moins de terminologies et d'ontologies en français et il manque crucialement d'outils et de services pour les exploiter. Cette lacune contraste avec le montant considérable de données biomédicales produites en français, particulièrement dans le monde clinique (e.g., dossiers médicaux électroniques).* Methode & Résultats – *Nous présentons le SIFR BioPortal, une plateforme ouverte et générique pour l'hébergement d'ontologies et de terminologies biomédicales françaises, basée sur la technologie du* National Center for Biomedical Ontology. *Le portail facilite l'usage et la diffusion des ontologies du domaine en offrant un ensemble de services (recherche, alignements, métadonnées, versionnement, visualisation, recommandation) y inclus pour l'annotation sémantique. En effet, le SIFR Annotator est un outil pour traiter des données textuelles en français. Une évaluation préliminaire, montre que le service web obtient des résultats équivalents à ceux reportés précédemment, tout en étant public, fonctionnel et tourné vers les standards du web sémantique. Nous présentons également de nouvelles fonctionnalités de la plateforme.*

Abstract

Background – *Despite a large adoption of English in science, a significant quantity of biomedical data uses the French language. Besides the existence of various English tools, there are considerably less terminologies and ontologies available in French and there is a strong lack of related tools and services to exploit them. This lack does not match the huge amount of biomedical data produced in French, especially in the clinical world (e.g., electronic health records).* Method & Results – *We present the SIFR BioPortal, an open platform to host French biomedical ontologies and terminologies based on the technology developed by the National Center for Biomedical Ontology. The portal facilitates use and fostering of ontologies by offering a set of services (search, mappings, metadata, versioning, visualization, recommendation), including for annotation purposes. Indeed, the SIFR Annotator is a*

publicly accessible ontology-based annotation tool to process text data in French. A preliminary evaluation shows the web service matches previously reported work in French in performance, while being public, functional and turned toward the semantic web standards. We also present new improvements of the platform.

Mots clés : Ontologies/terminologies biomédicales, portail d'ontologies, annotation sémantique, web sémantique, fouille de texte, BioPortal.

Keywords: Biomedical ontologies/terminologies, ontology portal, semantic annotation, semantic web, text mining, BioPortal.

1 Introduction

Le volume de données en biomédecine ne cesse de croître [1]. En dépit d'une large adoption de l'anglais, une quantité significative de ces données est en français [2]. En général, le contenu textuel de ces ressources est indexé par mots-clés pour permettre une recherche efficace mais avec des limites évidentes : synonymie, polysémie, utilisation des connaissances du domaine. L'intégration de données biomédicales et l'interopérabilité sémantique sont indispensables pour permettre de nouvelles découvertes scientifiques qui pourraient émerger du rapprochement des différentes données disponibles. Les terminologies et les ontologies[1] jouent un rôle central en sciences de la vie pour structurer les données médicales et les rendre interopérables [3]. En particulier, la communauté les utilise pour créer des index sémantiques, destinés à améliorer la recherche et la fouille de données grâce aux connaissances médicales que ces ontologies formalisent. Cependant, outre l'existence de nombreuses ressources en anglais, il y a beaucoup moins d'ontologies en français et il manque crucialement d'outils et de services pour les exploiter. Cette lacune contraste avec le montant considérable de données biomédicales produites en français, particulièrement dans le monde clinique (e.g., dossiers médicaux électroniques). En outre, lorsqu'il s'agit d'utiliser des ontologies dans le domaine de la santé, plusieurs questions se posent, par exemple : (i) si j'ai développé une ontologie, comment je la met à disposition des autres à moindre coût ? (ii) si j'ai besoin d'une ontologie, où est-ce que je la récupère dans le format de mon choix ? (iii) s'il existe plusieurs possibilités, comment savoir qu'elle ontologie utiliser, laquelle est la plus appropriée pour ma tâche ? (iv) comment est-ce que je peux utiliser les ontologies pour lier/annoter mes données ? Globalement, nous pouvons dire, qu'il existe un véritable besoin d'utiliser des ontologies sans avoir à les gérer : il faut permettre aux experts du domaine de se focaliser sur la science biomédicale, et ne pas avoir à se soucier des questions d'ingénierie des connaissances liées à une utilisation avancée des ontologies. Dans cet article, nous proposons des solutions nouvelles pour ces deux enjeux majeurs (mise à disposition des ontologies et annotation sémantique) pour la gestion des données biomédicales.

Le projet SIFR (*Indexation sémantique de ressources biomédicales francophones* - www.lirmm.fr/sifr), a pour objectif de résoudre les défis scientifiques et techniques pour exploiter les ontologies dans la construction de services d'indexation, de fouille, et de recherche de données pour les ressources biomédicales françaises. En collaboration avec le NCBO (*National Center for Biomedical Ontology* - www.bioontology.org) de l'Université de Stanford qui développe un portail d'ontologies biomédicales (le NCBO BioPortal [4]), nous avons développé un service d'annotation sémantique basé sur les ontologies biomédicales françaises similaire à celui qui existe pour les ressources anglaises [5], mais spécialisé

[1] Ci-après, nous utilisons seulement le mot « ontologie » y inclus pour terminologie et vocabulaire.

pour le français. Ce workflow, toujours en phase de prototypage, permet de détecter des concepts d'ontologies dans des données textuelles et de valoriser la sémantique de ces ontologies pour étendre et exploiter ces annotations. Pour mettre en œuvre ce workflow, ainsi que pour répondre à un besoin de la communauté biomédicale française d'avoir à disposition une plateforme ouverte et générique pour les ontologies de leur domaine, nous offrons également un portail d'ontologies, le SIFR BioPortal (http://bioportal.lirmm.fr), qui permet à n'importe quel utilisateur : (i) de rendre public une ontologie et bénéficier automatiquement d'un ensemble de service pour celle-ci (recherche, annotation, alignement, versionnement, recommandation, visualisation, etc.) ; (ii) d'identifier une ontologie a réutiliser pour une tâche particulière impliquant des données en langue française. Comme nous le discutons, cette plateforme est complémentaire des solutions existantes, en particulier du portail HeTOP [6], développé par le CISMeF du CHU de Rouen et de l'infrastructure ITM développée par la société Mondeca [7].

La suite de l'article est composée de la façon suivante : dans la section 2 nous ferons un point sur l'historique des travaux dans ce domaine (portail d'ontologie et annotation) pour positionner nos contributions. Dans la section 3, nous présenterons le contexte du projet SIFR avant d'en présenter les résultats, dans la section 4. La section 5 est consacrée à une évaluation préliminaire du service d'annotation. Finalement, la section 6 conclue et indique nos perspectives pour la l'extension et la pérennité des solutions proposées dans cet article.

2 Etat de l'art

2.1 Portails d'ontologies/terminologies biomédicales

En France, dans le domaine de la santé, le besoin de lister et intégrer les ontologies en langues françaises a été identifié depuis les années 2000, plus particulièrement au sein des initiatives UMLF (Unified Medical Language for French) [8], et VUMeF (Vocabulaire Unifié Médical Francophone) [9], qui avaient pour objectif de reproduire ou de se rapprocher des solutions de la US National Library of Medicine telle que la ressource UMLS [10]. Le projet InterSTIS (2007-2010) a donné lieu aux derniers résultats sur ce sujet. Le besoin d'offrir une base de terminologies unifiées et inter-reliées les unes les autres dans un format pivot avait été identifié par ce projet. Ce besoin devait servir la problématique d'annotation sémantique de données (également appelé indexation sémantique) [11]. Les principaux résultats de ce projet en termes de ressource multi-terminologiques ont été :

- Le portail SMTS basé entre autre sur la technologie ITM développée par la société Mondeca [7]. Si SMTS n'est plus maintenu aujourd'hui, la technologie ITM existe toujours et elle est déployée par la société pour ses clients, dans le domaine de la santé ou autre.
- Le portail HMTP [12] (Multiple Terminologies in a Health Portal) développé par le groupe CISMeF, qui est plus tard devenu HeTOP (Health Terminology/Ontology Portal - www.hetop.eu) [6]. HeTOP est un portail multi-terminologique et multilingue qui intègre près de 70 terminologies/ontologies telle que MeSH, la CIM-10, la CCAM, SNOMED-International, etc. Il permet de chercher des termes et d'accéder à leurs traductions, d'identifier les liens entre ontologies et tout particulièrement d'accéder aux données indexées par le CISMeF dans des plateformes telles que DocCISMeF [13]. La valeur ajoutée du portail vient clairement de l'expertise médicale de ses développeurs, qui intègrent les ontologies méthodiquement une par une, produisent des traductions des termes et indexent (souvent manuellement) les ressources de données du domaine.

Parallèlement à ces initiatives françaises, la communauté internationale a développé des plateformes similaires dont l'écho est parfois multiplié car la communauté adressée est bien plus grande (anglaise) et inclue bien souvent aussi les biologistes. En complément

d'UMLS [10], nous pouvons citer l'OBO Foundry [14], l'Ontology Lookup Service [15] et le NCBO BioPortal [4]. Ce dernier est devenu une plateforme de référence pour la publication et la mise à disposition d'ontologie en biomédecine. Le NCBO BioPortal permet d'accéder, visualiser, rechercher et commenter plus de 400 ontologies ou terminologies (principalement en anglais) de différent domaine en biologie ou médecine. Les ontologies peuvent être utilisées pour annoter automatiquement des données textuelles [5] et le portail offre également un index sémantique de plusieurs jeux de données biomédicales annotées avec les ontologies du portail [16]. La plateforme est tournée vers le web sémantique, et les utilisateurs ont accès soit via une application Web, une API REST, ou un SPARQL endpoint.

Les philosophies des deux portails HeTOP et NCBO BioPortal sont différentes même s'ils sont sur les mêmes créneaux. La vision d'HeTOP, semblable à celle d'UMLS, est de construire un « metathesaurus » où chaque ontologie source est intégrée dans un modèle spécifique (et propriétaire) et où elles sont manuellement inspectées, traduites, et alignées. Bien entendu, ce travail fastidieux a pour valeur ajoutée une grande richesse et confiance dans les données ci-intégrées, mais cela vient au prix d'un processus humain complexe et long qui ne passe pas à l'échelle du nombre d'ontologies produites pour le domaine de la santé aujourd'hui (même la NLM ne peut pas suivre le rythme de production d'ontologies biomédicales pour l'intégration dans UMLS). En outre, ce contenu est difficilement exportable du système d'information d'HeTOP, qui ne propose pas publiquement d'API et/ou de format standard et interopérable pour les récupérer facilement (bien que dans le cadre de ce travail plusieurs ontologies nous ont été données par CISMeF au format OWL). La vision de BioPortal est différente, elle consiste à offrir une plateforme ouverte, basée sur les standards du web sémantique sans intégrer les ontologies une à une dans un méta modèle. La plateforme supporte des mécanismes de production et stockage d'alignements et d'annotations, mais ne crée pas de nouveau contenu. Le portail n'est pas multilingue, mais il offre une variété de services aux utilisateurs qui veulent eux-mêmes déposer leurs ontologies ou seulement réutiliser certaines déjà stockées dans la plateforme. Pour une comparaison exhaustive des deux portails et des outils d'annotation, nous renvoyons le lecteur vers [17].

2.2 Outils d'annotation pour le biomédical français

L'utilisation d'ontologies pour indexer et intégrer les ressources de données est un moyen de valoriser la connaissance d'un domaine en facilitant la recherche et la fouille de données. L'expression 'annotation sémantique' [18] fait référence au processus d'identification de concepts, et de relations entre concepts, dans un document ou une ressource de donnée. Dans cet article, une annotation est vue comme une méta-information qui *associe une donnée à un concept*. L'association entre données et ontologies permet alors à des agents logiciels de profiter de la connaissance représentée dans les ontologies pour mieux exploiter les données (e.g., intégration, fouille). Cependant, produire pour les données textuelles des annotations qui utilisent des concepts d'ontologies pour faciliter l'indexation, l'intégration et donc la recherche de ces données est très difficile. Ce challenge a été relevé avec succès par exemple avec la Gene Ontology ou par des ressources comme PubMed qui est manuellement indexée avec le MeSH. Le choix d'annoter manuellement peut s'avérer très couteux, car cela nécessite des experts qui doivent avoir une connaissance approfondie du domaine, et ne passe pas à l'échelle de la quantité de données générées dans le domaine [19]. C'est pourquoi des approches automatiques ou semi-automatiques sont demandées par les chercheurs.

Ainsi, dans le domaine biomédical, l'annotation automatique et l'indexation de ressources textuelles est un sujet important. En anglais, il existe plusieurs outils de reconnaissance de concept qui permettent d'identifier des entités d'ontologies (concept ou relation) à partir de texte. Par exemple, IndexFinder [20], MetaMap [21], CONANN [22], SAPHIRE [23], et Mgrep [24], Terminizer [25], Semantator [26], cTakes [27], BioAnnote [28]. De manière

assez générale, la connaissance représentée dans les ontologies est rarement utilisée pour l'expansion d'annotations et les outils ne sont pas facilement accessibles (e.g., web service) et il est souvent difficile de les utiliser avec une nouvelle ontologie. La plupart de ces outils sont limités à UMLS ou à un petit nombre d'autres ontologies. Ne pas avoir ces limites a été rapporté comme un avantage du NCBO Annotator [5] que nous réutilisons dans nos travaux.

Cependant, là encore, les solutions pour traiter les données en langue française sont très limitées e.g., [29]. Nous pouvons mentionner : F-MTI (*French Multi-Terminology Indexer*) [30, 31] développé au CISMeF et désormais la propriété de la société Vidal et son successeur, ECMT (*Extracteur de Concepts Multi-Terminologique* – http://ecmt.chu-rouen.fr), qui est un web service également proposé plus récemment par CISMeF et qui est lui aussi désormais transféré au domaine privé auprès de la société Alicante. L'avantage principal de ces outils est qu'ils utilisaient partiellement les techniques du traitement du langage naturel telles qu'une approche sac de mots et un outil de racinisation spécifique pour le français. Cependant, F-MTI n'est pas accessible au public. Et depuis très récemment, ECMT non plus, même si le produit évolue toujours (il est désormais utilisé au sein de divers projets commerciaux pour indexer des comptes rendus médicaux). Dans leur version précédentes, ces outils n'étaient pas réellement « orientés service » (peu flexible et peu de paramètres) ni n'utilisaient d'URI ; ainsi leurs résultats étaient difficilement interopérables avec le web de données. Même si ces aspects pratiques sont moins importants que bien entendu la qualité des annotations générées, il s'avère que ce sont des critères de choix pour favoriser l'adoption des outils d'annotation sémantique au-delà du cercle restreint des experts du traitement de la langue. Par conséquent, le besoin d'un véritable service d'annotation sémantique pour la communauté biomédicale française est toujours bien d'actualité.

En outre, le travail sur les workflow d'annotation sémantique n'est jamais très éloigné de celui sur les plateformes d'ontologies. Chaque fois qu'un groupe développe une plateforme, il développe en général un outil d'annotation qui va avec : MetaMap [21] pour l'UMLS, Whatizit [32] pour l'OLS, le NCBO Annotator [5] pour BioPortal et ECMT/F-MTI [30] pour le CISMeF. Ainsi, notre choix de réutilisation de la technologie du NCBO nous permet d'adresser les deux besoins mentionnés en introduction et pour lesquels, comme nous venons de le voir, les travaux précédents sont insuffisants ou pas satisfaisants.

3 Projet SIFR : annotation, portail d'ontologies et multilinguisme

Les ontologies jouent un rôle central en sciences de la vie pour structurer les données biomédicales et les rendre interopérables. L'utilisation d'ontologies pour indexer et intégrer les ressources de données est un moyen de valoriser la connaissance en facilitant la recherche et la fouille de données. Cependant, les découvertes qui pourraient être réalisées sont souvent limitées par la disponibilité et le traitement des données dans une langue seulement, le plus souvent l'anglais, pour laquelle il existe le plus d'ontologies et d'outils. La communauté exprime régulièrement un besoin pour des méthodes d'annotation sémantique automatiques capables de mettre en valeur le traitement automatique du langage naturel français. Un des objectifs du projet SIFR est de développer un workflow d'annotation sémantique de données textuelles francophones et de l'offrir sous forme de service ouvert et public à la communauté biomédicale française. Pour réaliser un tel objectif, se posent les questions suivantes : (i) à quelle plateforme d'ontologie attacher un tel outil, en particulier si les ontologies sont d'accès restreint ? (ii) comment gérer la question du multilinguisme (qui apparait dès qu'on travaille avec une autre langue que l'anglais) ?

Les laboratoires français qui produisent une ontologie dans le cadre d'un projet ou d'une collaboration doivent pouvoir la rendre disponible de façon stable et pérenne à la fin du projet pour assurer sa visibilité et réutilisation. Une plateforme ouverte (n'importe qui peut

soumettre du contenu) et générique (n'importe quel type de contenu et de format) dont le focus est la communauté scientifique française est ainsi nécessaire. Toutes les ontologies ne peuvent pas être intégrées par CISMeF dans HeTOP qui n'a pas cette vocation et pour lequel le processus d'ajout d'une ontologie est assez fastidieux. Ainsi, le SIFR BioPortal peut combler ce besoin dans l'écosystème français en permettant de garder les ontologies dans une plateforme développée par la communauté française ouverte, flexible et transparente. Bien entendue, la pérennité de la plateforme au-delà du projet SIFR dépendra de son niveau d'acceptation et d'adoption par la communauté.

Le choix de la technologie du NCBO nous a paru évident pour venir compléter l'écosystème des solutions existantes en France. La technologie du NCBO est open source et depuis fin 2013, l'organisation mets à disposition une « machine virtuelle » qui permet de déployer en local sa propre instance de BioPortal. En outre, miser sur la réutilisation de cette technologie nous permet à la fois de valoriser l'existant (ne pas refaire ce qui existe) et nous assure une interopérabilité avec le BioPortal principal puisque les API sont les mêmes. Ce dernier aspect nous permet également de nous intéresser à la deuxième question sur le multilinguisme. L'enjeu de la gestion du multilinguisme dans le domaine des ontologies biomédicales est selon nous primordial et dépasse les aspects linguistiques. En effet, l'intégration multilingue de jeux de données annotés sémantiquement pourrait permettre des études translationnelles sur des données relatives à des populations différentes. Et donc de pouvoir affiner les recherches médicales sur des domaines tels que la pharmacogénomique, l'étude du rôle de l'environnement sur l'expression des gênes, ou les relations gènes-maladies. Gérer le multilinguisme dans un portail d'ontologies ne se limite bien sûr pas à offrir l'interface graphique dans plusieurs langues (ce qui n'est pas encore fait dans notre portail). Il faut se poser les questions de la représentation multilingue des données du portail (ontologies, alignements) et de leur valorisation dans les services offerts (recherche, indexation, annotation, recommandation). Il faut gérer des cas complexes d'ontologies « monolingues » traduites et maintenues par des organismes différents (comme c'est le cas du MeSH français de l'INSERM) ainsi que les cas d'ontologies multilingues (ou partiellement multilingues) qui sont lexicalisées dans différentes langues au moyen de solutions très variées (e.g., rdfs:label et xmlang ou une représentation lexicale riche comme LEMON [33]). Les premières impliquent l'existence de plusieurs ressources (i.e., plusieurs fichiers) et donc plusieurs ontologies dans le portail ; tandis que les deuxièmes seront uniques. Même si nous avons spécifié les étapes pour rendre BioPortal multilingue [34], ce qui rendra le besoin d'un portail « français » partiellement caduque, nous avons opté dans un premier temps pour une plateforme non-multilingue et dans laquelle nous pouvons expérimenter tout en offrant un premier service d'annotation et d'hébergement d'ontologies à la communauté. Sur le long terme, notre objectif n'est pas de multiplier les plateformes mais bien de consolider une solution homogène, pertinente et satisfaisante pour les utilisateurs. Ainsi, en collaboration avec le NCBO, nous choisirons soit de (i) rendre le NCBO BioPortal complétement multilingue (et y transférer nos ressources ainsi que notre workflow d'annotation) ; (ii) soit de mettre en place un architecture de portails interconnectés et interopérables qui offrira aux utilisateurs des services pour leurs ontologies, indépendamment de leur langues (i.e., un réseau de BioPortals).

4 Une plateforme ouverte et générique pour les ontologies françaises

4.1 SIFR BioPortal

Nous avons déployé une version spécifique du portail d'ontologie développé par le NCBO. Le SIFR BioPortal (http://bioportal.lirmm.fr) contient actuellement 10 (+1 privée) ontologies/terminologies médicales en français extraites de l'UMLS ou qui nous ont été fournies

par CISMeF, suite au développement d'un export OWL pour la plateforme HeTOP. La liste des ontologies actuellement dans le portail est donnée dans le Tableau 1. Les ontologies d'accès restreint peuvent être stockées en mode *private* (c'est-à-dire accessibles seulement aux ayant droits). Pour le moment, nous n'avons pas encore déterminé les restrictions d'accès pour toutes les ontologies avec leurs éditeurs, ainsi la plupart ne peuvent pas être téléchargées via le portail (mais peuvent être utilisée). Comme la plateforme d'origine, le SIFR BioPortal permet de :

- Stocker des ontologies et leurs métadonnées (pour plusieurs versions),
- Rechercher dans les ontologies,
- Gérer les versions de chaque ontologies et télécharger les fichiers,
- Stocker et rendre des alignements entre les ontologies,
- Générer des alignements automatiques simples (même identifiant ou lexicaux),
- Visualiser le contenu d'une ontologie,
- Laisser des commentaires sur une ontologie, une classe ou un alignement,
- S'abonner au flux de notifications pour une ontologie,
- Annoter des données textuelles avec des concepts (cf. section 4.2),
- Obtenir une recommandation d'ontologie pour un corpus de texte ou des mots clés,
- Stocker des projets qui utilisent des ontologies.

Tableau 1 – Ontologies disponibles dans la version 1.0 du SIFR BioPortal

Nom	Id BioPortal	Source	Classes
Dictionnaire médical pour les activités règlementaires en matière de médicaments	MDRFRE	UMLS	66378
Medical Subject Headings, version francaise	MSHFRE	UMLS	27455
Réseau sémantique UMLS	STY	UMLS	133
Terminologie minimale standardisée en endoscopie digestive	MTHMSTFRE	UMLS	1700
MedlinePlus	MEDLINEPLUS	CISMeF	849
Systematized Nomenclature of MEDicine, ver. française	SNMIFRE	CISMeF	106291
Biologie Hors Nomenclature	BHN	CISMeF	2534
Terminologie des effets indésirables	WHO-ARTFRE	CISMeF	3483
Classification Int. du Fonctionnement, du handicap et de la santé	CIF	CISMeF	1496
Classification Internationale des Maladies - 10ème rév	CIM-10	CISMeF	19853
Classification Internationale des Soins Primaires, 2è édition (*private*)	CISP2	CISMeF	745

La figure 1 présente des exemples de l'interface graphique du SIFR BioPortal :

(1) Visualisation des métadonnées et des informations sur une ontologie (nombre de visites, versions, métriques, projets, views, etc.),
(2) Interface de recherche dans les différentes ontologies du portail,
(3) Mise à disposition des alignements pour une ontologie donnée.

Les ontologies peuvent être directement chargées par les utilisateurs dans le portail, qui gère des formats différents tels qu'OWL, OBO, SKOS et UMLS. Des alignements peuvent également être chargés dans le portail et seront ainsi partagés avec la communauté dans la même

plateforme que les ontologies qu'ils relient. Un point fort de notre plateforme pour l'interopérabilité est qu'elle fournit des URIs pour les ontologies qui n'ont pas été développées dans un format standard web sémantique e.g., OWL ou SKOS. Par exemple, MSHFRE et MDRFRE qui sont exportées de l'UMLS ne possédaient pas d'URI. Grace à leur inclusion, chaque terme de ces ontologies francophones possède une URI par exemple :

- http://purl.lirmm.fr/ontology/MDRFRE/10007635 pour Cardiomyopathies

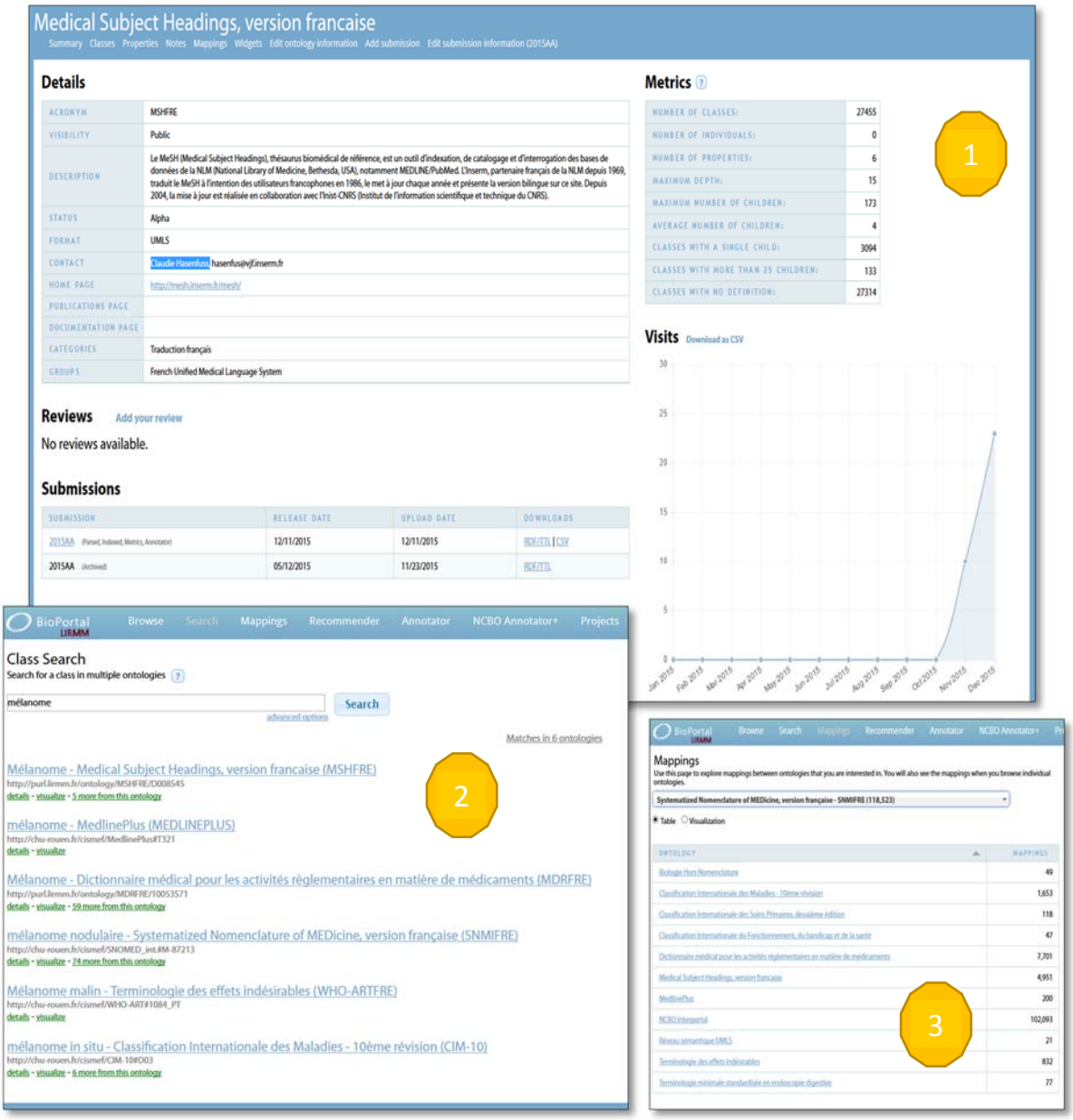

Figure 1 – *Interface graphique du SIFR BioPortal*

Dans le cas des ontologies fournies par CISMeF en OWL, les URIs utilisés sont ceux crées par le CISMeF, par exemple : http://chu-rouen.fr/cismef/MedlinePlus#T190 pour coloscopie. Cependant, les URIs de CISMeF ne sont pas pour le moment déréférencables, c'est-à-dire qu'elles ne renvoient pas automatiquement vers une page web contenant de l'information sur ce terme, tandis que les URI crées par le SIFR BioPortal le sont.

L'ensemble du contenu de la plateforme est accessible via une API de service web REST qui retourne du JSON-LD (http://data.bioportal.lirmm.fr/documentation) ou via un

SPARQL endpoint qui offre tout le contenu de la plateforme en RDF (http://sparql.bioportal.lirmm.fr/test). Ainsi, des applications tierces peuvent contribuer et consommer automatiquement tout le contenu du portail et développer des workflows qui valorisent les ontologies et/ou annotent des données médicales en RDF.

Nous avons développé plusieurs ajouts au SIFR BioPortal (section 4.4) et nous travaillons actuellement pour y inclure d'autres ontologies, qui parfois ne sont pas disponibles ailleurs telles que : OntoToxnuc, HRDO, OntoPneumo, Top-menelas, Ontoma, Phare, ONL. Par exemple, les ontologies développées par les laboratoires français LIMICS, MediCIS, ou LORIA sont des candidates pour le SIFR BioPortal. Huit autres ontologies produites par différents projets de recherche français, ont été identifiées pour la version 1.1 à venir. De plus, nous prévoyons de solliciter le CISMeF pour interconnecter les plateformes et permettre à une ontologie éditée, curée, et traduite dans HeTOP, d'être accessible dans le SIFR BioPortal. Et respectivement permettre aux URIs utilisés par le SIFR BioPortal d'être déréférencé vers HeTOP, quand c'est le CISMEF qui a créé la version OWL de l'ontologie.

4.2 SIFR Annotator

Grace à l'environnement BioPortal et au contenu français qu'il contient, nous avons pu déployer un service d'annotation spécifique pour le français : le SIFR Annotator (http://bioportal.lirmm.fr/annotator) qui peut être utilisé pour détecter et localiser des concepts d'ontologies dans des données textuelles et qui utilise la sémantique des ontologies pour étendre et exploiter ces annotations. Le fonctionnement du SIFR Annotator est principalement similaire au NCBO Annotator [5]. Notre annotateur se distingue des précédents efforts en français pour plusieurs raisons :

- Il s'agit d'un service web d'accès libre et public facilement interconnectable et qui renvoie des annotations dans des formats standards comme JSON-LD ou RDF ;
- La sémantique des ontologies est utilisée à la fois pour créer des annotations directes (basée sur la syntaxe des termes) et pour l'étape d'expansion sémantique qui valorise : les synonymes, les alignements entre ontologies, la hiérarchie is-a ;
- Le service est complétement paramétrable et l'utilisateur peut sélectionner les ontologies ainsi que le type d'annotation et leur classement ;
- Il utilise une grande collection d'ontologies biomédicales publiques disponibles et il peut quasi instantanément utiliser toute autre ontologie déposée sur le SIFR BioPortal.

La figure 2 illustre l'interface graphique du SIFR Annotator. Pour plus de détails sur le workflow d'annotation sémantique, nous renvoyons le lecteur à [5]. Nous nous focalisons ci-après sur les nouveautés mise en œuvre lors de sa mise en œuvre pour le français.

4.3 Mappings multilingues

Pour une ontologie monolingue l'exploitation pour l'annotation, la recherche ou l'indexation des données est fortement liée à la correspondance entre sa langue et la langue des données à annoter. D'où la nécessité de la gestion multilingue des ontologies en particulier lorsque elles ont été « traduites » (e.g., MeSH qui est produite à la fois par la NLM et l'INSERM) et non pas conçues de façon multilingue (e.g., Orphanet Rare Disease Ontologie). Dans le processus de traduction, une autre entité (fichier, ressource, document) est créé, et les alignements avec l'ontologie d'origine ne sont pas nécessairement conservés d'une manière qui permet d'interconnecter automatiquement les ontologies par la suite. Il faut souvent réconcilier ces alignements. C'est pourquoi, nous avons mené une étude sur la réconciliation des alignements entre 10 ontologies anglaises du NCBO BioPortal et les équivalentes françaises du SIFR BioPortal [35]. Ceci nous a permis de réconcilier près de 227K alignements et de formellement aligner 9 des ontologies concernées à plus de 96%. Il a fallu identifier pour

chaque couple d'ontologie les bonnes propriétés à utiliser pour faire les alignements (e.g., code, CUI, URI ou autre identifiant) et gérer tous les cas spécifiques car les versions françaises ne sont que rarement l'image exacte des versions anglaises. Nous avons explicitement représenté ces mappings à l'aide des propriétés de SKOS et GOLD et les avons rendus disponibles dans notre portail : http://bioportal.lirmm.fr/mappings. Pour cela nous avons dû modifier significativement les mécanismes de représentation et d'hébergement des mappings pour stocker des alignements qui n'ont qu'une partie dans le SIFR BioPortal.

Au cours de ce travail nous avons constaté quelques anomalies dans certains couples d'ontologies que nous avons transmis aux traducteurs afin de les vérifier et éventuellement les rectifier (par exemple, des codes dans la version française de MeSH inexistants dans la version anglaise, ou des erreurs de choix de CUI pour MedlinePlus). Ainsi l'interconnexion entre les deux portails est désormais assurée. Ces mappings vont aussi offrir la possibilité de faire de l'indexations, de la recherche et de l'intégration de données multilingues.

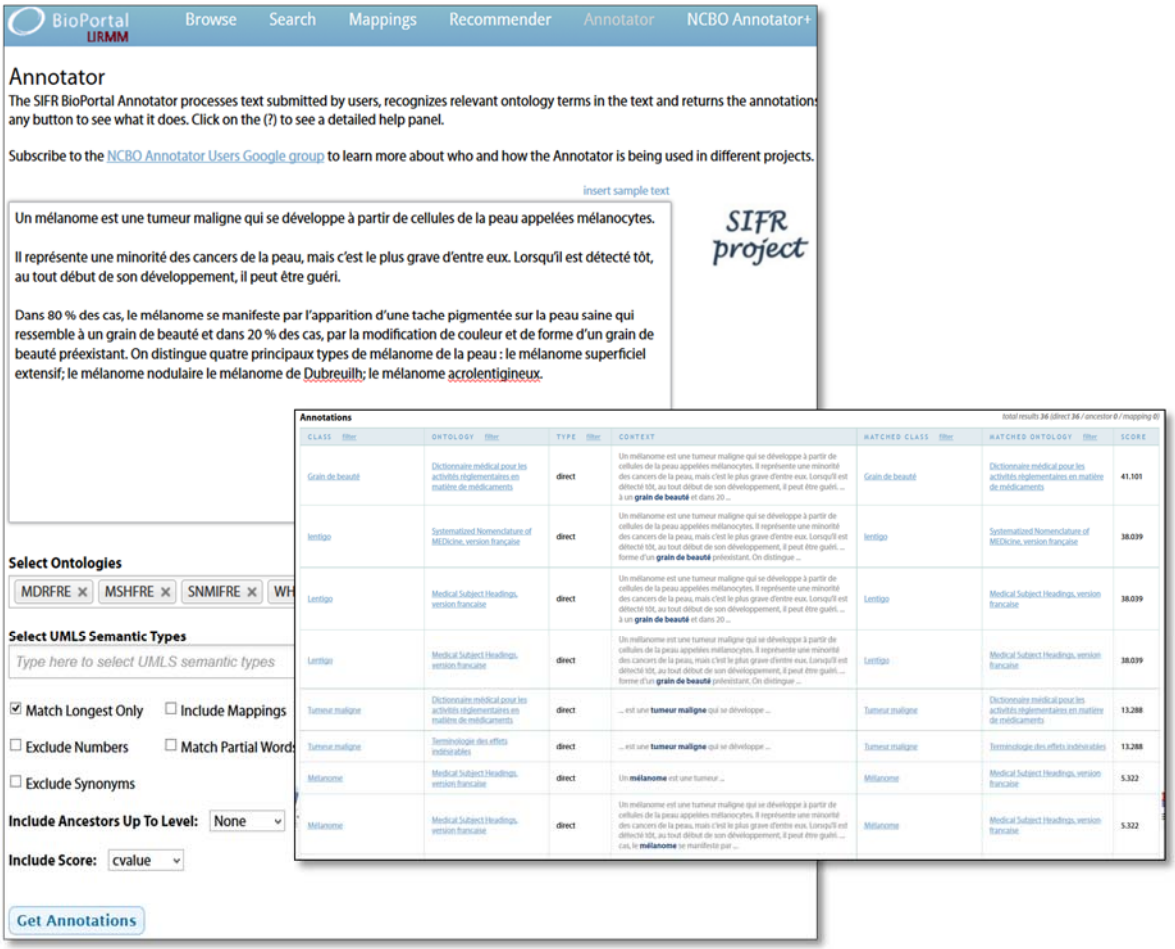

Figure 2 – *Interface graphique du SIFR Annotator*

4.4 Nouvelles fonctionnalités

En complément du travail réalisé sur les alignements multilingues, nous avons ajouté des fonctionnalités au SIFR Annotator telles que le score et le classement des résultats [36], ou leur transformation en RDF avec l'Annotation Ontology [37]. En particulier, le classement des annotations est depuis longtemps reconnu comme indispensable pour la tâche d'indexation. Cette fonctionnalité de classement implémente diverses méthodes basées sur la fréquence dont la plus performante, basée sur une méthode d'extraction de termes très utilisée en langage naturel (C-Value) [38]. Dans notre cas, ce n'est pas la méthode d'extraction qui

est utilisée, mais seulement la mesure sous-jacente qui permet de donner plus d'importance aux termes multi-mots (e.g., cancer du sein) en tenant compte dans le calcul des fréquences des termes imbriqués (e.g., cancer).

En outre, comme ces changements ne sont pas spécifiques au SIFR Annotator, nous avons développé un service *proxy* pour le NCBO Annotator (http://bioportal.lirmm.fr/ncbo_annotatorplus). Ainsi, nos améliorations ne sont pas limitées au SIFR Annotator mais sont aussi utilisables par le NCBO Annotator. Plus récemment, nous avons entamé des modifications au niveau du stockage des métadonnées sur les ontologies dans le SIFR BioPortal de façon à permettre le stockage de plus de métadonnées, décrites à l'aide des vocabulaires standards tels que le DublinCore, l'Ontology Metadata Vocabulary, VOID et d'autres.

5 Evaluation préliminaire du SIFR Annotator

Le workflow d'annotation sémantique sous-jacent aux NCBO et SIFR Annotators a été évalué à plusieurs reprises, incluant des évaluations extérieures. Sur l'anglais, par exemple, l'Annotator a été comparé à MetaMap [39], l'outil de reconnaissance de concept qui sert de référence dans le domaine biomédical et s'est montré très rapide avec une précision plus élevée pour différentes ressources et dictionnaires. Bien entendu la performance (précision) varie en fonction du type de données annotées. Par exemple, de 93% pour reconnaitre des processus biologiques dans des données d'expression de gênes à 60% dans des données d'essais cliniques, ou 88% pour reconnaitre des noms de maladies dans des données d'expression de gênes ou 23% dans des citations PubMed. Des évaluations extérieures ont rapportés des résultats équivalents [40, 41] : en moyenne la précision est de 73%, et le rappel de 78%.

Plus récemment, en termes d'évaluation du workflow d'annotation sémantique, nous avons mené différentes expérimentations en anglais et en français. Avant l'existence de la version française de l'Annotator, une évaluation réalisée sur un corpus de 2000 citations PubMed (titre seulement) en anglais que nous avons annoté avec MeSH nous a permis de comparer en terme de précision et rappel le comportement des trois outils ECMT, F-MTI et le NCBO Annotator. Le choix de MeSH s'est imposé car (i) c'est la seule terminologie que nous avons pu utiliser conjointement avec les trois outils ; (ii) nous avons pu alors utiliser les annotations manuelles (réalisées par les experts de la US National Library of Medicine pour indexer les citations PubMed avec MeSH) comme gold standard. Cette évaluation a montré qu'en terme de rappel, les trois outils étaient équivalent (0,291 (Annotator), 0,293 (ECMT), 0,264 (F-MTI)). Cependant, en terme de précision, l'Annotator obtenait sensiblement de meilleurs résultats (0,75) que ECMT (0,44) et F-MTI (0,67). Bien que ces résultats ne soient que partiels, ils ont permis de mesurer la performance des trois outils sur l'anglais donnant ainsi un point de repère pour le français, étant donné que les méthodes sous-jacentes des outils sont les mêmes. En complément, une évaluation comparative menée en 2014 avec la première version du SIFR Annotator, ECMT et F-MTI sur le français a également été réalisée. Pour 2000 citations PubMed (titre seulement) en français, nous avons pu là aussi mesurer le rappel et la précision des trois outils par rapport au gold standard des annotations manuelles. En termes de rappel, ECMT s'est montré un peu plus performant (7,63) devant l'Annotator (4,94) et F-MTI (1,67). Mais en termes de précision ECMT et l'Annotator étaient équivalents (respectivement 20,79 et 19,32) devant F-MTI (4,33). Ainsi, même si les résultats ne sont pas encore satisfaisants (des améliorations sont à prévoir et d'autres scénarios d'évaluation sont à mener), ils permettent d'établir que le SIFR Annotator peut rivaliser avec les outils précédents disponibles (en particulier ECMT). De plus, notre service bénéficie de nombreux autres avantages fonctionnels (inclusion dans BioPortal, web service, expansion sémantique, format RDF, ouverture et généricité) qui en font un outil intéressant pour la communauté biomédicale française.

Dans les deux évaluations, les chiffres relativement bas par rapport au gold standard, en particulier en terme de rappel, s'expliquent clairement par le fait que les indexeurs de la NLM utilisent non seulement le titre, mais aussi le résumé et parfois le papier complet pour identifier les annotations. Ainsi, il est parfaitement normal que des outils automatiques qui ne traitent seulement le titre ne rivalisent pas. Le biais étant le même pour les trois outils testés, l'objectif de ce gold standard était simplement de nous permettre de comparer les outils entre eux. La qualité effective des annotations générées par le SIFR Annotator fera l'objet d'une évaluation spécifique sous peu en utilisant un corpus de référence tel que par exemple *Quaero* (https://quaerofrenchmed.limsi.fr).

6 Conclusion et perspectives

Dans cet article nous avons présenté une plateforme ouverte et générique pour l'hébergement d'ontologies en langue française (ou qui contiennent des labels français) ainsi qu'un service d'annotation. Ces produits sont encore nouveaux et nécessiteront d'être améliorés pour trouver complétement leur place dans l'écosystème français d'applications sémantiques pour le domaine de la santé. Par exemple, au sein d'un projet de pharmacogénomique (ANR Pratik-Pharma (2016-2020)), nous allons utiliser le SIFR Annotator pour annoter des dossiers patients de l'Hôpital Européen George Pompidou avec pour objectif de comparer les connaissances de l'état de l'art en pharmacogénomique (en anglais) et les connaissances cliniques (en français).[2] Nous souhaitons alors réaliser de nombreuses améliorations au SIFR Annotator spécifiques au traitement des données cliniques : gestion de la négation, du contexte, de la temporalité, et de la désambiguation. Nous avons plusieurs pistes pour cela établies suite à notre travail en traitement automatique de la langue sur l'extraction de termes [42] et plus récemment la désambiguation [43]. En ce qui concerne le portail d'ontologies, nous avons entreprit de compléter les métadonnées proposées par la plateforme, pour permettre à un utilisateur de choisir une ontologie avec le maximum d'information sur celle-ci. Nous avons entamé les démarches pour rendre le portail multilingue et bien entendu en traduire l'interface. Ensuite, nous prévoyons de solliciter la communauté pour dans un premier temps amorcer la plateforme en chargeant des ontologies existantes. Nous sommes optimistes sur le fait que le SIFR BioPortal offrira à la communauté biomédicale française (e.g., cliniciens, professionnels de santé, chercheurs) des services basés sur les ontologies performants, leur permettant d'améliorer leur processus de production et de consommation de données. En outre, les résultats du projet ne sont pas limités au français (mais inclus aussi l'anglais, l'espagnol) et nous sommes en train de les transférer dans le domaine de l'agronomie dans le cadre du projet AgroPortal (http://agroportal.lirmm.fr) [44]. Le code est disponible à l'adresse suivante : https://github.com/sifrproject.

Remerciements

Ce travail est réalisé au sein du projet SIFR (www.lirmm.fr/sifr) financé en partie par le programme JCJC de l'Agence Nationale de la Recherche (ANR-12-JS02-01001), l'Université de Montpellier, le CNRS et l'Institut de Biologie Computationnelle de Montpellier (ANR-11-BINF-0002). Nous remercions également le NCBO (Université de Stanford), le groupe CISMeF (CHU de Rouen), et Suzanne Pereira (VIDAL) pour la mise à disposition de leur ontologies et/ou technologie.

[2] Pour cette tâche, étant donné les restrictions d'accès aux données, une installation locale du SIFR Annotator sera réalisée en interne à l'HEGP. Cette situation a déjà été rencontrée par le NCBO.

Références

[1] T. B. Murdoch and A. S. Detsky, "The Inevitable Application of Big Data to Health Care," *Journal of the American Medical Association*, vol. 309, no. 13, pp. 1351–1352, 2013.

[2] A. Neveol, J. Grosjean, S. J. Darmoni, and P. Zweigenbaum, "Language Resources for French in the Biomedical Domain," in *9th International Conference on Language Resources and Evaluation, LREC'14* (N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Reykjavik, Iceland), pp. 2146–2151, European Language Resources Association, May 2014.

[3] D. L. Rubin, N. H. Shah, and N. F. Noy, "Biomedical ontologies: a functional perspective," *Briefings in Bioinformatics*, vol. 9, no. 1, pp. 75–90, 2008.

[4] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen, "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, pp. 170–173, May 2009.

[5] C. Jonquet, N. H. Shah, and M. A. Musen, "Un service Web pour l'annotation sémantique de données biomédicales avec des ontologies," in *13èmes Journées Francophones d'Informatique Médicale, JFIM'09* (M. Fieschi, P. Staccini, O. Bouhaddou, and C. Lovis, eds.), vol. 17 of *Informatique et Santé*, (Nice, France), April 2009.

[6] J. Grosjean, T. Merabti, N. Griffon, B. Dahamna, and S. Darmoni, "Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal," in *9th International Conference on Terminology and Artificial Intelligence, TIA'11*, (Paris, France), pp. 119–122, November 2011.

[7] G. Delaporte and F. Amardeilh, "ITM et intelligence économique : MONDECA," in *Congrès Veille stratégique, scientifique et technologique*, (Toulouse , France), p. 546, 2004.

[8] P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, Éric Jarrousse, N. Grabar, P. Ruch, F. L. Duff, B. Thirion, and S. J. Darmoni, "Towards a unified medical lexicon for French," *Studies in health technology and informatics*, vol. 95, pp. 415–420, 2003.

[9] S. Darmoni, E. Jarrousse, P. Zweigenbaum, P. L. Beux, F. Namer, R. Baud, M. Joubert, H. Vallée, A. Côté, A. Buemi, D. Bourigault, G. Recource, S. Jeanneau, and J.-M. Rodrigues, "VUMeF: extending the French involvement in the UMLS Metathesaurus," in *American Medical Informatics Association Annual Symposium, AMIA'03*, (Washington DC, USA), p. 884, November 2003.

[10] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. 267–270, 2004.

[11] M. Joubert, A.-L. Peretti, J. Gouvernet, and M. Fieschi, "Refinement of an automatic method for indexing medical literature - a preliminary study," in *19th International Conference of the European Federation for Medical Informatics, MIE'05* (R. Engelbrecht, A. Geissbuhler, C. Lovis, and G. Mihalas, eds.), vol. 116 of *Studies in Health Technology and Informatics*, (Geneva, Switzerland), pp. 683–688, IOS Press, 2005.

[12] S. J. Darmoni, S. Pereira, S. Sakji, T. Merabti, E. Prieur, M. Joubert, and B. Thirion, "Multiple Terminologies in a Health Portal: Automatic Indexing and Information Retrieval," in *12th Conference on Artificial Intelligence in Medicine, AIME'09* (C. Combi, Y. Shahar, and A. Abu-Hanna, eds.), no. 5651 in Lecture Notes in Computer Science, (Verona, Italy), pp. 255–259, Springer, June 2009.

[13] S. J. Darmoni, B. Thirion, J.-P. Leroy, M. Douyère, B. Lacoste, C. Godard, I. Rigolle, M. Brisou, S. Videau, E. Goupy, J. Piot, M. Quéré, S. Ouazir, and H. Abdulrab,

"Doc'CISMEF: a search tool based on "encapsulated" MeSH thesaurus," in *10th World Congress on Medical Informatics* (R. Rogers, R. Haux, and V. Patel, eds.), vol. 10, (London, UK), pp. 314–318, 2001.

[14] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, T. O. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. H. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, pp. 1251–1255, November 2007.

[15] R. G. Côté, P. Jones, R. Apweiler, and H. Hermjakob, "The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries," *BMC Bioinformatics*, vol. 7, p. 7, February 2006.

[16] C. Jonquet, A. Coulet, N. H. Shah, and M. A. Musen, "Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical," in *21èmes Journées Francophones d'Ingénierie des Connaissances, IC'10* (S. Despres, ed.), (Nimes, France), pp. 271–282, June 2010.

[17] J. Grosjean, L. F. Soualmia, K. Bouarech, C. Jonquet, and S. J. Darmoni, "Comparing BioPortal and HeTOP: towards a unique biomedical ontology portal?," in *(2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO'14*, (Granada, Spain), p. 11, April 2014.

[18] S. Handschuh and S. Staab, eds., *Annotation for the Semantic Web*, vol. 96 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2003.

[19] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. A. Hunter, "Manual curation is not sufficient for annotation of genomic databases," *Bioinformatics*, vol. 23, no. 13, pp. 41–48, 2007.

[20] Q. Zou, W. W. Chu, C. Morioka, G. H. Leazer, and H. Kangarloo, "IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing," in *American Medical Informatics Association Annual Symposium, AMIA'03*, (Washington DC, USA), pp. 763–767, November 2003.

[21] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.," in *American Medical Informatics Association Annual Symposium, AMIA'01*, (Washington, DC, USA), pp. 17–21, November 2001.

[22] L. H. Reeve and H. Han, "CONANN: An Online Biomedical Concept Annotator," in *4th International Workshop Data Integration in the Life Sciences, DILS'07* (S. Cohen-Boulakia and V. Tannen, eds.), vol. 4544 of *Lecture Notes in Computer Science*, (Philadelphia, PA, USA), pp. 264–279, Springer-Verlag, June 2007.

[23] W. R. Hersh and R. A. Greenes, "SAPHIRE - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships," *Computers and Biomedical Research*, vol. 23, pp. 410–425, October 1990.

[24] M. Dai, N. H. Shah, W. Xuan, M. A. Musen, S. J. Watson, B. D. Athey, and F. Meng, "An Efficient Solution for Mapping Free Text to Ontology Terms," in *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'08*, (San Francisco, CA, USA), March 2008.

[25] D. Hancock, N. Morrison, G. Velarde, and D. Field, "Terminizer – Assisting Mark-Up of Text Using Ontological Terms," in *3rd International Biocuration Conference*, (Berlin, Germany), April 2009.

[26] D. Song, C. G. Chute, and C. Taoa, "Semantator: Annotating Clinical Narratives with Semantic Web Ontologies," in *AMIA Joint Summits on Translational Science*, (San

Francisco, USA), pp. 20–29, March 2012.

[27] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *American Medical Informtics Association*, vol. 17, pp. 507–513, June 2010.

[28] H. López-Fernández, M. Reboiro-Jato, D. Glez-Peña, F. Aparicio, D. Gachet, M. Buenaga, and F. Fdez-Riverola, "BioAnnote: A software platform for annotating biomedical documents with application in medical learning environments," *Computer methods and programs in biomedicine*, vol. 111, pp. 139–147, July 2013.

[29] R. Baud, P. Ruch, C. Lovis, and A.-M. Rassinoux, "Recherche conceptuelle dans les textes médicaux," in *8èmes Journées Francophones d'informatique Médicale, JFIM'00* (M. Fieschi, O. Bouhaddou, R. Beuscart, and R. Baud, eds.), vol. 12 of *Informatique et Santé*, (Marseille, France), pp. 205–216, Springer-Verlag, May 2000.

[30] S. Pereira, A. Névéol, G. Kerdelhué, E. Serrot, M. Joubert, and S. J. Darmoni, "Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue," in *American Medical Informatics Association Annual Symposium, AMIA'08*, (Washington DC, USA), pp. 586–590, November 2008.

[31] S. Sakji, Q. Gicquel, S. Pereira, I. Kergoulay, D. Proux, D. SJ, and M. Metzger, "Evaluation of a French Medical Multi-Terminology Indexer for the Manual Annotation of Natural Language Medical Reports of Healthcare-Associated Infections," in *13th World Congress on Medical Informatics, MedInfo'10* (C. S. et al., ed.), vol. 160 of *Studies in Health Technology and Informatics*, (Cape Town, South Africa), pp. 252–256, IOS Press, September 2010.

[32] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through Web services: Calling Whatizit," *Bioinformatics*, vol. 24, no. 2, pp. 296–298, 2008.

[33] J. McCrae, D. Spohr, and P. Cimiano, "Linking lexical resources and ontologies on the semantic web with lemon," in *8th Extended Semantic Web Conference, ESWC'11* (G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. DeLeenheer, and J. Pan, eds.), no. 6643 in Lecture Notes in Computer Science, (Heraklion, Crete, Greece), pp. 245–259, Springer, May 2011.

[34] C. Jonquet and M. A. Musen, "Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal," in *Terminology & Ontology : Theories and applications Workshop, TOTh'14* (C. Roche, R. Costa, and E. Coudyzer, eds.), (Brussels, Belgium), p. 2, December 2014.

[35] A. Annane, V. Emonet, F. Azouaou, and C. Jonquet, "Multilingual Mapping Reconciliation between English-French Biomedical Ontologies," in *UNDER SUBMISSION for 6th International Conference on Web Intelligence, Mining and Semantics, WIMS'16*, (Nimes, France), June 2016.

[36] S. Melzi and C. Jonquet, "Scoring semantic annotations returned by the NCBO Annotator," in *7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14* (A. Paschke, A. Burger, P. Romano, M. Marshall, and A. Splendiani, eds.), vol. 1320 of *CEUR Workshop Proceedings*, (Berlin, Germany), p. 15, CEUR-WS.org, December 2014.

[37] S. Melzi and C. Jonquet, "Representing NCBO Annotator results in standard RDF with the Annotation Ontology," in *7th International Semantic Web Applications and Tools for Life Sciences - poster session, SWAT4LS'14* (A. Paschke, A. Burger, P. Romano, M. Marshall, and A. Splendiani, eds.), vol. 1320 of *CEUR Workshop Proceedings*, (Berlin,

Germany), p. 5, CEUR-WS.org, December 2014.

[38] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the C-value/NC-value Method," *Digital Libraries*, vol. 3, pp. 115–130, August 2000.

[39] N. H. Shah, N. Bhatia, C. Jonquet, D. L. Rubin, A. P. Chiang, and M. A. Musen, "Comparison of concept recognizers for building the Open Biomedical Annotator," *BMC Bioinformatics*, vol. 10, September 2009.

[40] J. S. Simon N. Twigger, Joey Geiger, "Using the NCBO Web Services for Concept Recognition and Ontology Annotation of Expression Datasets," in *Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS'09* (M. S. Marshall, A. Burger, P. Romano, A. Paschke, and A. Splendiani, eds.), vol. 559 of *CEUR Workshop Proceedings*, (Amsterdam, The Netherlands), CEUR-WS.org, November 2009.

[41] I. N. Sarkar, "Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts," in *American Medical Informatics Association Annual Symposium, AMIA'10*, (Washington DC., USA), pp. 717–721, November 2010.

[42] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Extraction automatique de termes combinant différentes informations," in *21ème Traitement Automatique des Langues Naturelles, TALN'14* (B. Bigi, ed.), vol. 2, (Marseille, France), pp. 407–412, July 2014.

[43] J.-A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Prédiction de la polysémie pour un terme biomédical," in *12ème COnférence en Recherche d'Information et Applications, CORIA'15* (E. Gaussier, ed.), (March), pp. 437–452, Paris, France 2015.

[44] C. Jonquet, E. Dzalé-Yeumo, E. Arnaud, and P. Larmande, "AgroPortal: a proposition for ontology-based services in the agronomic domain," in *3ème atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement, IN-OVIVE'15*, (Rennes, France), p. 5, June 2015.

Adresse de correspondance

Clement Jonquet – LIRMM, 161 Rue Ada, 34090 Montpellier, France – jonquet@lirmm.fr