MLPubMed : une base de données bibliographique multi-lingue

Stéfan J. Darmoni[a,b], Lina F. Soualmia[a,b], Nicolas Griffon[a], Julien Grosjean[a], Gaétan Kerdelhué[a], Ivan Kergourlay[a], Benoit Thirion[a], Badisse Dahamna[a]

[a] *CISMeF & TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France*
[b] *INSERM, Unité Mixte de Recherche en Santé (UMR_S) 872, équipe 20, Paris, France*

Résumé

MEDLINE/PubMed contient de nombreux articles dans des langues autres que l'anglais, mais il est impossible d'y accéder sans connaître l'anglais (interface) et difficile de les trouver sans connaître la recherche avancée de PubMed. L'objectif de ce travail est de proposer un outil permettant d'accéder au sous-ensemble MEDLINE/PubMed d'une langue donnée en interrogeant dans cette même langue ; dans ce travail, nous nous focaliserons sur le français.

Pour le français, nous avons amélioré significativement la traduction du MeSH existante, en y ajoutant les synonymes des descripteurs MeSH et une traduction partielle des Supplementary Concepts et des Concepts. Pour 11 langues principalement européennes, les traductions du MeSH ont été intégrées dans notre portail terminologique inter-lingue (HeTOP). Les sous-ensembles de PubMed concernant cinq langues européennes ont été chargés dans une base de données interne, en utilisant un parseur dédié. Le moteur de recherche sémantique de CISMeF a été utilisé, grâce à : a) une généricité permettant d'interroger tout type de document ; b) un outil maintenant multi-lingue.

Pour le français, plus de 654 000 citations de PubMed ont été intégrées dans MLPubMed, une base de données bibliographiques interrogeables grâce à notre moteur multi-lingue dans cinq langues européennes : français, allemand, espagnol, portugais et norvégien. La preuve de concept a été évaluée pour le français, permettant aux professionnels de santé et aux patients ne lisant pas suffisamment bien l'anglais de rechercher dans leur langue natale des articles scientifiques.

Keywords: Databases, bibliographic; French language; Information storage and retrieval; PubMed; User-Computer interface;

Introduction

MEDLINE créée par la Bibliothèque de Médecine des Etats-Unis (NLM®) est la base de données bibliographiques la plus utilisée dans le monde. A ce jour (16/1/2013), elle contient 20 174 596 citations[1] de 14 555 journaux indexés de 81 pays. Chaque citation est indexée manuellement par le thésaurus Medical Subject Headings (MeSH®) [1]. MEDLINE est un sous-ensemble de PubMed (URL: http://pubmed.gov/), qui contient notamment des citations d'articles très récents pas encore indexés manuellement. À la même date, PubMed contenait 22 423 351 citations (+10% par rapport à MEDLINE)[2]. Dans sa version 2013, le thesaurus MeSH contenait 26 853 Descripteurs, 83 Qualificatifs (Q) ou sous-mots-clés, 215 043 Supplementary Concepts, 335 545 Concepts, and 744 909 Entry Terms ou synonymes. Les descripteurs MeSH sont traduits dans de nombreuses langues, dont en Europe, le français, l'allemand, l'espagnol, le portugais et le norvégien.

Plusieurs outils ont été développés et publiés pour aider les non-anglo-saxons a requêter MEDLINE/PubMed dans leur langue natale, en particulier BabelMeSH [3], [4] et PICO Linguist [4]. Notre équipe a également développé un outil permettant d'accéder à MEDLINE/PubMed en Français via un navigateur MeSH également en Français [5]. Cet outil est actuellement utilisé par 500 utilisateurs par jour ouvré, principalement des documentalistes, des étudiants en médecine et des médecins. Cet outil est enseigné dans la moitié des 31 facultés de médecine française.

L'objectif de ce travail est d'aller un pas plus loin et de développer une base de données bibliographiques multi-lingue (en cinq langues européennes en dehors de l'anglais) avec une interface, un langage de requêtes et des résultats dans la langue natale (ou choisie) par l'utilisateur : la cible de ce dernier est clairement le professionnel de santé ou le patient incapable de lire suffisamment bien l'anglais et qui restreindra sa recherche aux articles issus de PubMed dans sa langue natale.

Méthodes

Deux points méthodologiques ont du être résolus : la création d'un système d'information capable de gérer les données issues de PubMed et l'intégration des traductions des terminologies et des interface issues d'autres sources.

Depuis février 1995, notre équipe a développé un catalogue de ressources de santé de qualité (CISMeF) [6], dont l'objectif est toujours de recenser, de décrire et d'indexer les principales ressources institutionnelles en santé. Une première version d'un moteur de recherche sémantique a été développée en 2000, permettant d'interroger avec le thésaurus MeSH en bilingue (français/anglais). Depuis 2005, une indexation et une recherche d'information multi-terminologique a été ajoutée [7], grâce à un portail terminologique de santé cross-lingue (HeTOP ; URL: www.hetop.eu) [8]. Ce moteur de recherche est devenu en 2012 générique, capable d'indexer des

[1] http://www.ncbi.nlm.nih.gov/pubmed?cmd=PureSearch&db=pubmed&term=medline[sb]

[2] http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=PureSearch&db=pubmed&term=all[sb]

ressources Web, des dossiers médicaux et maintenant des citations PubMed, d'autre part, comme HeTOP et l'outil d'indexation, il est devenu multi-lingue.

Le système d'information générique de CISMeF a évolué afin de permettre l'intégration de métadonnées hétérogènes. Cinq couches constituent ce système :· "la base de données relationnelle" (assurant la persistance des données et le requêtage SQL optimisé), "le cache" (dévolu au clustering et à la répartition des objets), "les composants métiers" (bibliothèques Java CISMeF), "les services web" (accessibles notamment via les librairies clientes), et enfin "la couche de présentation" (interfaces Web, Rich Internet Application). Pour intégrer les informations de la NLM au format XML dans notre système d'information, un parser spécifique Java SAX a été développé. Actuellement, le parsing XML et l'alimentation de la base de données sont séparées en deux étapes. Dans un environnement de production, ces étapes seront réunies dans un processus batch concernant les nouvelles citations.

Dans la plupart des cas, les qualificatifs MeSH ne sont pas traduits en dehors de l'anglais et du français ; de ce fait, nous avons institué une collaboration avec les équipes internationales suivantes pour cette tâche de traduction : UMIT, Autriche (URL: http://www.umit.ac.in/) pour l'Allemand, Centre Tudor, Luxembourg (URL: www.tudor.lu/) pour l'Allemand et le Portugais, Hôpital Italien de Buenos Aires, Argentine (URL: http://www.hospitalitaliano.org.ar/) pour l'Espagnol et bientôt l'Italien, et la Bibliothèque Electronique de Santé Norvégienne (UL: http://www.helsebiblioteket.no/english) pour le Norvégien.

Pour le Français, tous les Descripteurs et les Qualificatifs ont été traduits par le Département Information Scientifique et Technique de l'INSERM [9]. L'équipe CISMeF a ajouté 24 563 synonymes et 689 acronymes ambigus aux Descripteurs et 163 synonymes aux Qualificatifs. Nous avons également traduit manuellement 20 887 MeSH Supplementary Concepts (10.17%) et ajouté 27 295 synonymes en Français.

Résultats

Comme preuve de concept, des milliers de citations ont été incluses dans notre outil MLPubMed dans cinq langues : français, allemand, espagnol, portugais et norvégien. L'utilisateur peut choisir sa langue et débuter une requête dans cette langue et obtenir des citations PubMed toujours dans cette langue dans une interface dédiée traduites par les équipes collaboratrices sus-nommées. Les mêmes métadonnées disponibles sur le site PubMed sont disponibles dans MLPubMed, en particulier le titre de l'article, le nom des auteurs et l'indexation MeSH mais dans la langue choisie par l'utilisateur, ainsi que le lien direct vers l'article en texte intégral sur le site de l'éditeur via le Digital Object Indentifier (DOI). De ce fait, l'objectif initial de créer une base de données bibliographiques disponibles en plusieurs langues et interrogeable totalement dans une langue donnée a été réalisée. L'URL transitoire dans cette phase de preuve de concept de MLPubMed est http://cispro.chu-rouen.fr/mlpubmed. Actuellement, l'ensemble des citations de PubMed disponibles en français ont été intégré dans l'outil MLPubMed (sans néanmoins de processus de mise à jour qui sera développé pour la phase de production). Nous avons validé dans un travail précédent [10] que les temps de réponse de cet outil est satisfaisant, à savoir inférieur à deux secondes pour les principales maladies, au sens nombre de citations PubMed (par exemple, l'asthme) .

Plusieurs améliorations effectuées dans le cadre du moteur Doc'CISMeF ont été appliquées à MLPubMed [11]:

1. Stratégies de recherche bilingues (Français & Anglais) (n=396) définies par les documentalistes de CISMeF et qui permettent d'améliorer la pertinence des résultats pour de nombreuses requêtes en langage naturel : par exemple la stratégie de recherche CISMeF "natrémie" est automatiquement transformée en requête MeSH "sodium/blood". Ces stratégies de recherche devront être traduites dans les quatre autres langues européennes pour être appliquées dans MLPubMed.
2. Super-concepts (ou métatermes; n=126) sont des (sous)-spécialités ou des sciences biologiques (comme la cardiologie ou la bactériologie) sélectionnés par le responsable de la bibliothèque de CISMeF (BT). Pour chaque super-concept, un lien sémantique a été créé manuellement avec au moins un Descripteur ou Qualificatif ; par exemple, le super-concept psychiatrie est associé aux Descripteurs *psychiatrie* et *hôpital psychiatrique*, qui se situent dans deux arborescences distinctes dans le MeSH.. Les super-concepts ont été créés pour optimiser la recherche d'information, du fait de l'étroitesse des Descripteurs MeSH concernant les spécialités médicales. Comme les stratégies de recherche, les super-concepts doivent être traduits dans les quatre autres langues européennes, alors que les liens sémantiques sont indépendants de la langue. La liste des stratégies de recherche et des métatermes sont disponibles dans le portail HeTOP.
3. Les facettes sont des critères permettant de filtrer les résultats d'une requête quelconque, en fonction des principales métadonnées choisies ; pour MLPubMed, celles-ci sont le pays et l'année de publication, les types de publications et les Descripteurs MeSH. Les facettes sont calculées par le moteur de recherche. Sur le plan technique, les facettes s'appuient sur une requête SQL calculant les cardinalités en temps réel. Toutes les facettes d'une requête donnée sont regroupées en un seul objet de type Clob. Pour chaque facette, les identifiants des ressources afférentes sont stockées dans des objets Java. Le processus s'opère en arrière-plan de sorte que la recherche principale ne soit pas impactée.
4. L'ordre d'affichage des résultats est différent de celui de PubMed et vise à améliorer la pertinence des premiers résultats proposés renvoyés pas notre outil. Dans MLPubMed, le moteur de recherche interprète la requête en tentant d'identifier les descripteurs des terminologies incluses dans notre système d'information, et au premier chef le MeSH qui est notre terminologie pivot. Plusieurs critères interviennent pour calculer l'ordre d'affichage : en premier lieu, les Descripteurs MeSH utilisés en majeur (correspondant aux thèmes principaux de l'article uniquement) et les termes du titre de la citation, mais aussi les dates de publication, et le type d'indexation (manuelle ou automatique). Ainsi, un score de 100% est attribué si la requête est entièrement alignée avec des Descripteurs MeSH utilisés en majeur ou des mots du titre. En cas d'égalité, l'affichage se fonde sur l'année de publication, en mode anti-chronologique.

D'autres développements du moteur de recherché Doc'CISMeF n'ont pas encore été implantés dans le nouvel outil MLPubMed (voir Discussion).

Discussion

Nous avons créé une base de données bibliographiques MLPubMed, disponible en cinq langues européennes et interrogeable par une version adaptée de notre moteur de recherche [6]. A part le français, nous pouvons extrapoler des résultats de l'étude précédente sur les temps de réponse [10] que ceux-ci également acceptable pour l'espagnol, le portugais et le norvégien car le nombre de citations PubMed dans ces trois langues est inférieur à celui du français avec respectivement 282 558, 72 700 et 34 854 ; le temps de réponse en allemand devrait être plus long d'un tiers, dans la mesure où le nombre de citations en allemand est supérieur d'un tiers au nombre de citations en français (792 861 vs. 654 096).

Dans un premier temps, la preuve de concept de MLPubMed sera finalisée pour le français, afin d'avoir une véritable base de données bibliographiques disponible gratuitement sur l'Internet. Nous prévoyons une mise en production à la fin du premier semestre 2013. Des contacts avec le principal éditeur de journaux en langue française ont été noués pour aller plus loin, et intégrer des journaux non indexés dans MEDLINE mais indexés dans Web of Science par exemple. L'aspect gratuit est fondamental, car contrairement à Al Gore qui permit l'accès gratuit de MEDLINE/PubMed en 1997, ce n'est toujours pas le cas en France : l'INIST du CNRS a une politique différente avec PASCAL et FRANCIS [12], qui ne sont toujours pas accessibles gratuitement sur l'Internet en janvier 2013.

Nous prévoyons ensuite de mettre à la disposition de nos collègues européens l'équivalent de MLPubMed en français pour quatre autres langues européennes : allemand, espagnol, portugais et norvégien. Un projet européen financé par l'Union Européenne pourrait faciliter cette montée en charge.

Si l'anglais est toujours de très loin la langue la plus utilisée dans le monde scientifique, sa proportion dans le temps diminue. Ainsi, les publications dans les autres langues conservent un intérêt notamment pour les personnes ne lisant pas (encore) correctement l'anglais. A notre connaissance, la construction de plusieurs bases de données bibliographiques dans plusieurs langues européennes fondées sur le même outil est une première.

La principale limitation de ce travail est le fait que le thésaurus MeSH n'est pas encore traduit complètement, notamment les MeSH Supplementary Concepts et les MeSH Concepts [13], sur lesquels notre équipe a commencé à travailler.

Perspectives

Plusieurs développements du moteur de recherche Doc'CISMeF n'ont pas encore été implantés dans le nouvel outil MLPubMed : en particulier l'indexation avec des terminologies de référence autres que le MeSH, l'affiliation des Qualificatifs aux MeSH Supplementary Concepts, l'indexation possible avec les MeSH Concepts et de même l'affiliation des Qualificatifs aux MeSH Concepts [13]. Il sera totalement illusoire d'envisager une indexation manuelles du corpus entier de PubMed en français (n=654 096). Nous utiliserons alors l'outil d'indexation automatique utilisé dans Doc'CISMeF ; cet outil devra être amélioré.

Conclusion

Nous avons réalisé la preuve de concept d'une base de données bibliographiques MLPubMed issues des citations PubMed et interrogeable totalement dans cinq langues européennes différentes, avec un chargement complet des citations pour le français. Cet effort sera poursuivi dans d'autres langues européennes.

Références

[1] Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc. 2000 Jul;88(3):265-6.

[2] Savage A. Changes in MeSH data structure. NLM Tech Bull. 2000 Mar-Apr;(313):e2. URL: http://www.nlm.nih.gov/pubs/techbull/ma00/ma00_mesh.html;

[3] Liu F, Ackerman M, Fontelo P. BabelMeSH: development of a cross-language tool for MEDLINE/PubMed. AMIA Annu Symp Proc. 2006:1012.

[4] Fontelo P, Liu F, Leon S, Anne A, Ackerman M. PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed. Stud Health Technol Inform. 2007;129(Pt 1):817-21.

[5] Thirion B, Pereira S, Névéol A, Dahamna B, Darmoni SJ. French MeSH Browser: a cross-language tool to access MEDLINE/PubMed. AMIA Annu Symp Proc. 2007 Oct 11:1132.

[6] Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med. 2000 Mar;39(1):30-5.

[7] Soualmia LF, Sakji S, Letord C, Rollin L, Massari P, Darmoni, SJ. Improving information retrieval with multiple health terminologies in a quality-controlled gateway. BMC Health Information Science and Systems , 2013 (in press).

[8] 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.

[9] Le MeSH bilingue anglais - français URL: http://mesh.inserm.fr/mesh/ (Accessed November, 21 2012).

[10] Darmoni SJ, Soualmia LF, Griffon N, Grosjean J, Kerdelhué G, Kergourlay I, Dahhamna B. Multi-lingual search engine to access PubMed monolingual subsets: a feasibility study. Submitted to MEDINFO 2013.

[11] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J. 2004 Dec;21(4):253-61.

[12] Base de données INIST URL: http://www.inist.fr/spip.php?rubrique9 (Accessed November, 21 2012).

[13] Darmoni SJ, Soualmia LF, Letord C, Jaulent MC, Griffon N, Thirion B, Névéol A. Improving information retrieval using MeSH Concepts: a test case on rare and chronic diseases. J Med Libr Assoc. 2012 Jul;100(3):176-83.

Adresse pour correspondance

SJ. Darmoni, CISMeF, Service d'Informatique Biomédicale, Cour Leschevin, porte 21, 3ème étage, 1 rue de Germont 76031 Rouen Cedex, France