Un service Web pour l'annotation sémantique de données biomédicales avec des ontologies

Clement Jonquet, Nigam Shah, et Mark A. Musen

Center for Biomedical Informatics Research, Stanford University School of Medicine, USA

Abstract

The range of publicly available biomedical data is enormous and is expanding fast. This expansion means that researchers now face a hurdle to extracting the data they need from the large numbers of data that are available. Biomedical researchers have turned to ontologies and terminologies to structure and annotate their data with ontology concepts for better search and retrieval. However, this annotation process cannot be easily automated and often requires expert curators. Plus, there is a lack of easy-to-use systems that facilitate the use of ontologies for annotation. This paper presents the Open Biomedical Annotator (OBA), an ontology-based Web service that annotates public datasets with biomedical ontology concepts based on their textual metadata. The biomedical community can use the annotator service to tag datasets automatically with ontology and terminology terms (from UMLS and NCBO). We have used the annotator service internally to index several online datasets (e.g., ArrayExpress, PubMed, ClinicalTrial.gov). The index is directly queriable in the NCBO BioPortal ontology repository (www.bioontology.org). Such semantic annotations facilitate translational discoveries by integrating annotated data.

Keywords

Medical Informatics Applications, Information Storage and Retrieval, Controlled Vocabularies, Unified Medical Language System, Terminology, Semantics, Knowledge (*from MeSH*) + Ontology, Annotation, Semantic Annotation (*not in MeSH*)

1 Introduction

La richesse et la diversité des données biomédicales aujourd'hui accessibles au public permettent l'intégration translationnelle de multiples études bioinformatiques (*translational bioinformatics research*) [1][2]. Cependant, les découvertes qui pourraient être réalisées par la fouille des données biomédicales sont limitées car la plupart des ressources publiques ne sont généralement pas décrites à l'aide de terminologies et d'ontologies. Un chercheur qui étudie les variations alléliques d'un gène voudrait connaître toutes les voies métaboliques qui sont touchées par ce gène, les médicaments dont les effets pourraient être modulés par les variations alléliques de ce gène, les maladies qui pourraient être causées par ce gène. Il peut également être intéressé par les essais cliniques qui ont étudié ces médicaments ou ces maladies. La connaissance nécessaire pour répondre à ces questions est aujourd'hui publique et disponible dans les ressources de données biomédicales en ligne, le problème est désormais de trouver cette information. La communauté biomédicale

reconnait d'ores et déjà l'importance des terminologies et des ontologies pour faciliter l'intégration de données et pour permettre de nouvelles découvertes [3]. Cependant, la variété des données est très importante et celles-ci sont rarement annotées à l'aide de concepts décrits dans des ontologies biomédicales.[1] Le plus souvent, les éléments d'une ressource (e.g., recueil de données expérimentales, diagnostics, maladies, échantillons, descriptions d'essais cliniques, publications, images) sont annotés avec des métadonnées textuelles qui décrivent cet élément. Le problème est que ces descriptions textuelles sont rarement structurées et que le plus souvent elles n'utilisent pas des termes définis dans des ontologies biomédicales. Il existe donc un challenge qui consiste à produire pour ces descriptions textuelles des annotations (ou labels, tags) qui utilisent des termes d'ontologies et faciliteront la recherche et l'indexation de ces données ainsi que leur intégration [4][5]. Par exemple, une recherche des essais cliniques pour le concept `carcinoma` doit considérer également les éléments indexés avec les termes `Malignant epithelial tumor` et `Epithelial Neoplasm` car l'ontologie SNOMED-CT nous précise que le premier terme est un synonyme du concept `carcinoma` et l'ontologie NCI Thesaurus nous précise que ce concept est un sous type (`is_a`) de `Epithelial Neoplasm`. Il existe quelques exemples pour lesquels *l'annotation sémantique* (i.e., l'annotation à l'aide de concepts définis dans des ontologies) des données s'est avérée très utile. Par exemple, Gene Ontology est largement utilisée pour décrire les fonctions moléculaires, les composants cellulaires et les processus biologiques des produits de gènes. Gene Ontology permet, entre autre, d'intégrer ces descriptions dans plusieurs bases de données, en gardant un formalisme commun. Autre exemple, quand une nouvelle citation PubMed est créée, le titre et le résumé de l'article correspondant sont indexés (grâce à des annotations manuelles) avec des termes de MeSH améliorant significativement la performance des recherches d'articles. Toutefois, en dehors de certains bons exemples, l'annotation sémantique de données biomédicales reste encore marginale. C'est pourquoi les succès mentionnés servent de motivation à notre travail.

L'annotation sémantique de données biomédicales n'est pas une pratique courante pour plusieurs raisons [6] :

- Les annotations ont le plus souvent besoin d'être créées manuellement par des experts ou directement par les auteurs des données,
- Le nombre d'ontologies biomédicales disponibles est important. En outre, ces ontologies changent régulièrement et se chevauchent les unes les autres. Elles sont dans des formats différents et ne sont pas toujours accessibles via des interfaces de programmation (API) qui permettent aux utilisateurs de les utiliser « programmatiquement »,
- Les utilisateurs ne connaissent pas toujours la structure des ontologies pour faire les annotations eux-mêmes. Parfois, ils ne connaissent pas l'existence des ontologies qu'ils pourraient utiliser,
- L'annotation est souvent une tâche supplémentaire ennuyeuse et sans retour immédiat pour l'utilisateur.

Dans le cadre du projet National Center for Biomedical Computing (NCBO), nous travaillons sur un système d'indexation de ressources biomédicales appelé *Open*

[1] Dans le reste de l'article, nous utilisons seulement l'expression « ontologie », pour terminologie et ontologie. Pour le travail présenté ici, les deux modèles peuvent être considérés équivalents puisque le niveau sémantique nécessaire (i.e., nom, synonyme, alignement, relations is_a) existe dans les deux modèles.

Biomedical Resources (OBR) [1][2]. Dans cet article, nous présentons brièvement les résultats de ce travail. Nous présentons principalement un service Web d'annotation, appelé *Open Biomedical Annotator (OBA)*, qui permet aux chercheurs d'utiliser les ontologies biomédicales pour annoter leurs données automatiquement. L'annotateur traite les métadonnées textuelles brutes pour les taguer avec des concepts définis dans des ontologies biomédicales et utilise la connaissance représentée dans les ontologies pour étendre ces annotations. Les annotations sont ensuite renvoyées aux utilisateurs.

2 Matériel et méthodes

Le processus d'annotation du service OBA est composé de deux étapes principales (Figure 1). Tout d'abord, le texte brut de l'utilisateur est traité par un outil de reconnaissance de concept qui détecte la présence d'un concept par reconnaissance syntaxique. Cet outil utilise un dictionnaire (ou lexique). Ce dictionnaire est une liste de termes qui identifient des concepts définis dans des ontologies.[2] Il est construit en récupérant à partir d'un ensemble d'ontologies biomédicales tous les noms ou synonymes qui identifient syntaxiquement les concepts. Le choix de l'ensemble d'ontologies à utiliser pour créer le dictionnaire dépend du type de données biomédicales que l'annotateur doit traiter. Par exemple, si un utilisateur veut annoter des expressions génétiques avec des noms de maladie, les ontologies SNOMED-CT et NCI Thesaurus peuvent être utilisées. Le résultat de cette première étape du processus est un ensemble d'*annotations directes*.

L'ensemble d'annotations directes est ensuite traité par un ou plusieurs composants d'expansion sémantique qui créent de nouvelles annotations à partir des annotations directes et en utilisant la connaissance représentée dans les ontologies. Par exemple:

- Le composant *fermeture transitive is_a* traverse la hiérarchie parent-enfant (ou type sous type) des ontologies afin de créer de nouvelles annotations avec les concepts parents d'un concept constituant une annotation directe. Par exemple, si un texte est annoté directement avec un concept de NCI Thesaurus, `melanoma`, ce composant d'expansion sémantique peut générer une nouvelle annotation avec les concepts `skin tumor` et `neoplasms` car NCI Thesaurus spécifie les relations : `melanoma is_a skin tumor` et `skin tumor is_a neoplasms`. Le niveau maximum dans la hiérarchie à considérer pour l'expansion est paramétrable.[3]
- Le composant *distance sémantique* utilise la notion de similarité sémantique[4] [7][8][9] afin de créer de nouvelles annotations avec les concepts similaires d'un concept constituant une annotation directe. Par exemple, si un texte est annoté directement avec un concept de MeSH, `melanoma`, ce composant d'expansion sémantique peut générer une nouvelle annotation avec les concepts `Apudoma` et `Neurilemmoma` car MesH spécifie ces trois concepts comme frères. La distance maximum à considérer pour l'expansion est paramétrable.
- Le composant *mapping* utilise les alignements (ou mappings) [10] entre ontologies

[2] Un concept est unique dans une ontologie (i.e., classe). Un terme est une chaîne de caractères possible qui identifie un concept. Habituellement, un concept a plusieurs termes (e.g., nom, synonymes, label).

[3] Ce composant d'expansion sémantique assume la transitivité de la relation is_a pour toutes les ontologies et terminologies utilisées. Cependant, si un utilisateur considère que cette transitivité n'est pas valide, le niveau maximal à considérer dans la hiérarchie peut alors être paramétrer à 1.

[4] Les distances sémantiques sont généralement basées soit sur la structure de l'ontologie, soit sur un corpus de données, ou les deux.

afin de créer de nouvelles annotations avec les mappings d'un concept constituant une annotation directe. Par exemple, si un texte est annoté directement avec le concept `NCI/C0025202` (`melanoma` dans NCI Thesaurus), ce composant d'expansion sémantique peut générer une nouvelle annotation avec les concepts `SNOMEDCT/C0025202` (`melanoma` dans SNOMEDCT) et `38865/DOID:1909` (`melanoma` dans Human Disease) parce que l'UMLS Metathesaurus et NCBO BioPortal fournissent ces alignements. Le type d'alignements à considérer pour l'expansion est paramétrable.

L'annotateur est conçu de façon à pouvoir sélectionner et paramétrer les composants d'expansion sémantique à la demande. Les performances du service sont liées aux paramètres choisis par l'utilisateur. Par exemple, la fermeture transitive is_a prend beaucoup de temps à traiter, même avec une hiérarchie pré-calculée. À la suite de la deuxième étape, les annotations directes ainsi que les *annotations expansées* sont classées par ordre de pertinence et renvoyées à l'utilisateur. Les annotations créées par l'annotateur ont pour sémantique : *ces données concernent (ou traitent de) ces concepts*. Cette sémantique peut être exploitée pour diverses applications comme la recherche ou l'intégration de donnée.

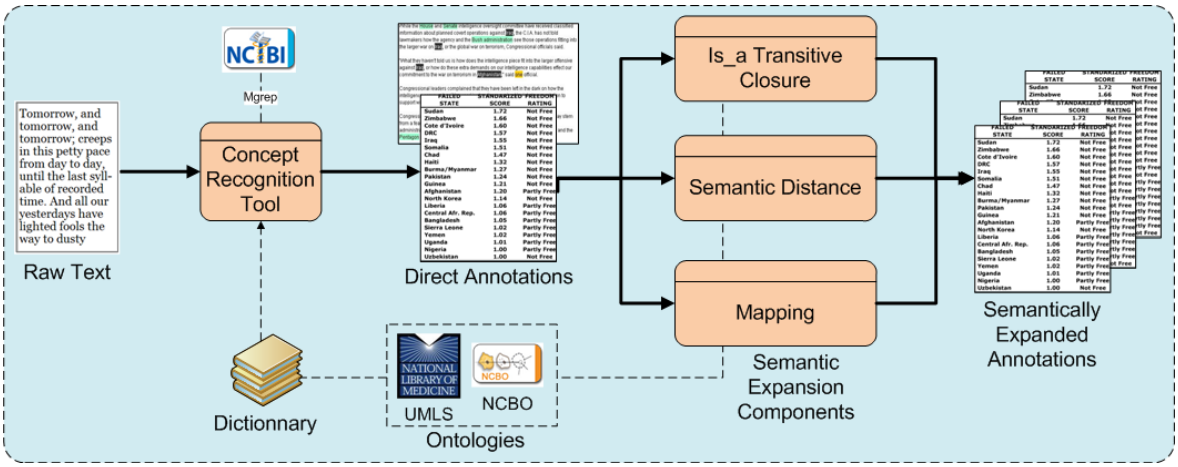

Figure 1 : Processus d'annotation du service OBA.

Les identifiants de concepts utilisés par l'annotateur proviennent soit de l'UMLS (Unified Medical Language System) Metathesaurus et sont formés par la concaténation des ABbreviated Source name (SAB) et des Concept Unique Identifier (CUI),[5] soit ils proviennent des Uniform Resource Identifier (URI) définis par le National Center for Biomedical Ontology (NCBO).

Chaque annotation possède un *contexte* qui précise si l'annotation est directe ou expansée et détaille l'origine de l'annotation. Par exemple, pour une annotation directe, le terme qui a était reconnu est précisé ainsi que sa position dans le texte. Dans le cas d'une annotation expansée le concept à partir du quel l'annotation a était créée est précisé. Les annotations peuvent être renvoyées à l'utilisateur dans différents formats (texte, tab-delimited, XML ou OWL). La description des résultats renvoyés par l'annotateur est disponible [11].

[5] Voir la documentation de UMLS : http://www.nlm.nih.gov/research/umls/meta2.html

3 Résultats et évaluation

3.1 Service web d'annotation (OBA)

Notre implémentation de l'annotateur utilise toutes les terminologies dans UMLS (anglaises seulement, soit 94) ainsi que toutes les ontologies (complètement utilisables et fonctionnelles) de BioPortal (132 ontologies au moment de l'écriture). Ces ontologies nous donnent un dictionnaire contenant 2.608.228 concepts et 5.186.484 termes, ce qui en fait un des plus gros dictionnaires disponibles aujourd'hui.

Dans la première étape du processus d'annotation, le service utilise Mgrep [12], un outil de reconnaissance de concept développé par le National Center for Integrative Biomedical Informatics (NCIBI) à l'Université du Michigan qui rapporte un degré de précision élevé (>95%) pour la reconnaissance de noms de maladie [13]. Mgrep utilise une structure de radix-tree qui permet une reconnaissance rapide et efficace de texte à partir d'un dictionnaire. Nous avons réalisé une évaluation comparative de Mgrep avec MetaMap [14], l'outil de reconnaissance de concept qui sert de référence dans le domaine biomédical. Dans cette évaluation nous avons utilisé quatre ressources de données et quatre dictionnaires différents pour évaluer les deux outils en termes de précision, rapidité d'exécution, et de passage à l'échelle. Le tableau 1 reporte certains de ces résultats pour deux des dictionaires. Mgrep s'est montré très rapide avec une précision plus élevée pour la plupart des ressources de données traitées et pour les quatre dictionnaires. Par exemple, Mgrep obtient 88% de précision par rapport à 75% pour MetaMap dans le traitement d'expressions génétiques de GEO. En outre, Mgrep est ouvert à tout type de dictionnaire et n'est pas limité, comme MetaMap, à UMLS. Nous avons donc pu utiliser Mgrep pour traiter les ontologies NCBO qui sont généralement disponibles dans les formats OBO (Open Biomedical Ontology) et OWL (Web Ontology Language). Pour plus de détail sur cette évaluation comparative, voir [15].

Tableau 1 : Précision de Mgrep et MetaMap avec deux dictionnaires

Ressource de données biomédicale	Disease dictionary (Concepts d'UMLS avec comme type sematique « disease or syndrome »		Biological processes (Gene Ontology) dictionary	
	Mgrep	MetaMap	Mgrep	MetaMap
PubMed (subset) www.ncbi.nlm.nih.gov/PubMed	0,23	0,091	0,77	0,76
ClinicalTrials.gov http://clinicaltrials.gov	0,87	0,71	0,6	0,63
Gene Expression Omnibus www.ncbi.nlm.nih.gov/geo	0,88	0,755	0,93	0,73
ARRS GoldMiner (subset) http://goldminer.arrs.org	0,73	0,548	0,58	0,33

Dans la deuxième étape du processus d'annotation, le service fait actuellement appel à un composant fermeture transitive is_a et un composant mapping. Le service Web est déployé avec une interface SOAP (Simple Object Access Protocol) et une interface REST (REpresentational State Transfer) - http://obs.bioontology.org.

3.2 Index de ressources biomédicales (OBR)

Nous avons utilisé l'annotateur pour créer un système d'indexation de ressources biomédicales appelé *Open Biomedical Resources (OBR)* [1][2]. Nous avons traité : (1) les

données d'expression génétiques de ArrayExpress et Gene Expression Omnibus (GEO), (2) les descriptions d'essais cliniques de Clinicaltrials.gov, (3) des légendes et descriptions d'images de ARRS Goldminer, et (4) des titres et résumés de citation PubMed. Le tableau 2 présente à la fois le nombre actuel d'éléments annotés et le nombre d'annotations créé pour chaque ressource (ces chiffres ont été produits avec une version plus ancienne du dictionnaire). L'index fournit des annotations (au moins 1) pour 99% des éléments de PubMed et 100% des éléments des autres ressources traitées. Le nombre moyen d'annotations est situé entre 359 et 824 par élément, avec une moyenne de 27% d'annotations directes. L'index est interfacé par une API services Web (REST).

Tableau 2 : Resource annotées dans l'index OBR.

Ressource de données biomédicale	Nombre d'élément	Taille de la ressource (Mb)	Nombre d'annotations directes (Mgrep)	Nombre total d'annotations	Nombre moyen d'annotations/ élément.
PubMed (subset) www.ncbi.nlm.nih.gov/PubMed	1050000	146.1	30822190	174840027	160
ArrayExpress www.ebi.ac.uk/arrayexpress	3371	3.6	502122	1849224	525
ClinicalTrials.gov http://clinicaltrials.gov	50303	99	16108580	48796501	824
Gene Expression Omnibus www.ncbi.nlm.nih.gov/geo	2085	0.7	165539	772608	359
ARRS GoldMiner (subset) http://goldminer.arrs.org	1155	0.5	134229	662687	564
TOTAL	1106914	249.9	47732660	226921047	(moy)486.4

Cet index de ressources permet à un utilisateur de rechercher des données biomédicales annotées pour un concept donné. Il permet également, pour un identifiant d'un élément d'une ressource (e.g., PubMedID, GEO datasetID), d'obtenir toutes les annotations qui ont étaient créées pour cet élément. La création d'un tel index favorise l'intégration translationnelle de différentes ressources de données biomédicales.

3.3 Intégration dans BioPortal

NCBO [15] développe et maintient une application Web appelée BioPortal qui permet d'accéder et d'utiliser des ontologies biomédicales (http://bioportal.bioontology.org) [17]. Le portail contient une grande collection d'ontologies, telles que Gene Ontology, National Cancer Institute Thesaurus, International Classification of Diseases, Foundational Model of Anatomy. Les utilisateurs peuvent consulter, rechercher et commenter la collection d'ontologies soit en ligne, soit via une API de services Web (REST).

L'index de ressources OBR est directement accessible dans BioPortal. Lorsqu'un utilisateur visite un concept donné, il a accès (lien Web) à l'ensemble des éléments annotés avec ce concept comme l'illustre la figure 2.

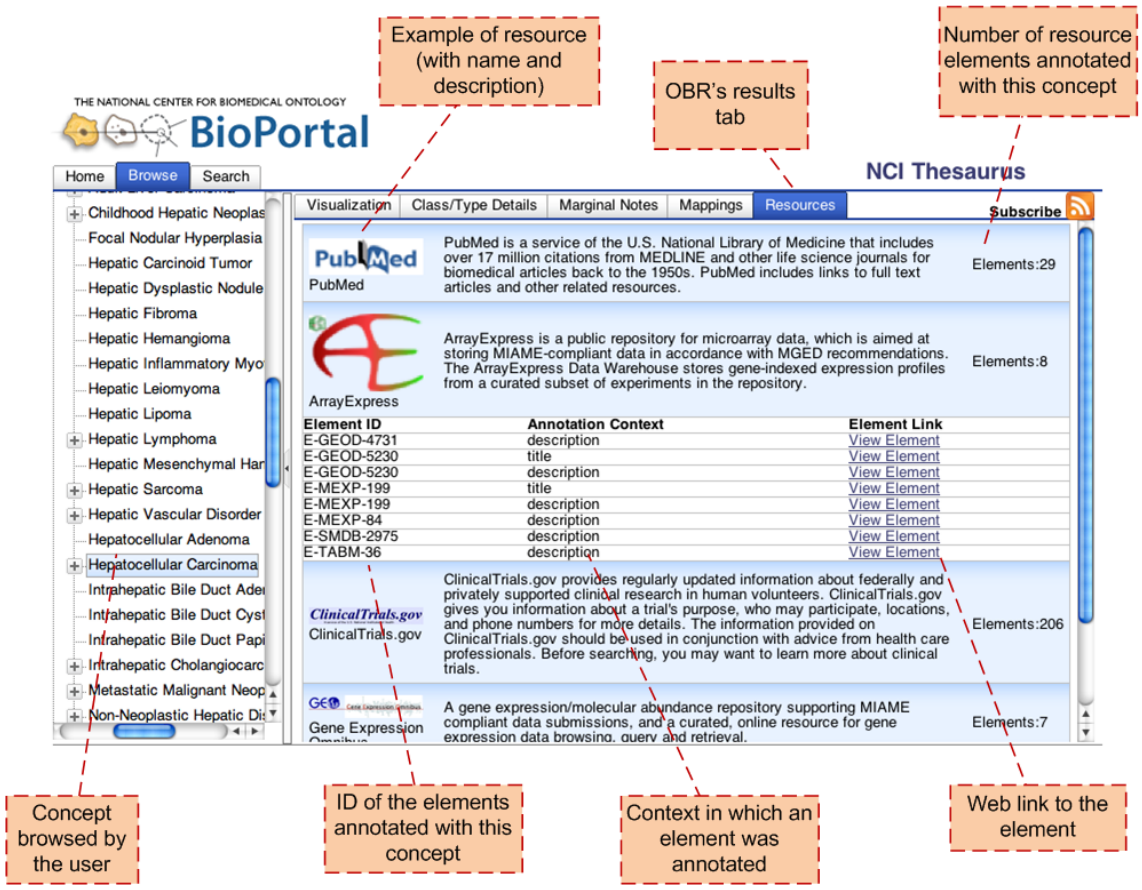

Figure 2 : Interface utilisateur dans BioPortal.

4 Discussion

4.1 Scenarios supportés par l'annotateur

Il existe de nombreux cas d'utilisation pour le service d'annotation autre que la création de l'index de ressources biomédicales présenté précédemment. Le service est actuellement en cours d'évaluation dans plusieurs projets extérieurs qui illustrent certains de ces cas d'utilisation :

(1) Les chercheurs travaillant sur Trialbank (www.trialbank.org) à l'Université de Californie, San Francisco, créent des annotations d'essais cliniques sur le VIH/sida afin de développer une application Web pour visualiser et comparer les essais. Ils évaluent notre service d'annotation pour traiter les champs 'health condition', 'intervention' et 'outcomes' d'essais cliniques de ClinicalTrials.gov.

(2) Les chercheurs de l'Université de l'Indiana évaluent l'intégration de l'annotateur dans leur système de gestion de recherche scientifique appelé Laboratree (http://laboratree.org), de sorte que toute annotation textuelle créée dans Laboratree puisse être complétée d'annotations sémantiques.

(3) Les développeurs de Collabrx (http://collabrx.com) utilisent l'annotateur dans leur plate-forme Rex pour traiter le contenu généré par les utilisateurs, de façon à déterminer quels dictionnaires (i.e., quelles ontologies) sont les plus appropriés pour leur contenu.

(4) Les chercheurs du Jackson Lab (www.jax.org) évaluent l'utilité de l'annotateur pour

trier des articles à l'aide de concepts reconnus dans les titres et les résumés de ces articles.

Chacun de ces groupes bénéficient d'une meilleure interopérabilité de leurs données à l'aide des annotations créées avec l'annotateur.

4.2 Etat de l'art

Dans le domaine biomédical, l'annotation automatique et l'indexation de ressources est un sujet important. 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 [18], MetaMap [14], CONANN [19], SAPHIRE [20], Baud et al. [21] et Mgrep [12][13]. MetaMap, qui identifie des concepts de UMLS Metathesaurus, est généralement utilisé comme référence en terme d'évaluation. Notre choix pour Mgrep a été fait sur la base de critères tels que la flexibilité, la rapidité, l'ouverture et le passage à l'échelle comme décrit précédemment. CONANN est très similaire à l'annotateur que nous proposons et est également disponible en ligne en tant que service. CONNAN identifie les meilleures correspondances possibles entre un texte et un ensemble de concepts, tandis que Mgrep identifie le plus grand nombre de concepts (de façon à générer le plus grand nombre d'annotations possible). Toutefois, CONNAN est limité à UMLS et n'effectue pas l'étape cruciale d'expansion sémantique. De manière assez générale, la connaissance représentée dans les ontologies est rarement utilisée pour l'expansion d'annotations ce qui donne à notre service un avantage important.

Nous pouvons également noter que l'usage de telles pratiques d'expansion est plus fréquent dans le domaine de la recherche d'information. De nombreux projets utilisent les annotations sémantiques pour améliorer les performances de moteur de recherche. Par exemple, MedicoPort [22] utilise la sémantique de UMLS pour faire de l'expansion de requête. Moskovitch et al. [5] utilisent des annotations sémantiques (concept-based search) et démontrent l'importance du contexte (context-sensitive search) lors de l'annotation de documents structurés. Ils utilisent les relations is_a lors de leur étape d'expansion sémantique. De façon similaire, Essie [23] montre qu'une combinaison judicieuse de la structure des documents et de l'expansion sémantique de concept est une approche utile pour la recherche d'information. HealthCyberMap [24] utilise des ontologies et des distances sémantiques pour visualiser le contenu de ressources biomédicales. La plupart de ces outils sont limités à UMLS ou à un petit nombre d'autres ontologies. Cette limitation, donne à notre service un avantage important car il utilise, en plus de UMLS, toutes les ontologies (utilisables) de NCBO BioPortal.

Khelif et al. [25] présentent un travail similaire à l'index de ressources OBR. Les auteurs ont annoté la GeneRIF en utilisant la plateforme GATE [26] qui leur permet d'extraire non seulement des concepts, mais des relations. Ils utilisent une application à base de graphes conceptuels appelée Corese [27] pour faire l'expansion sémantique. Névéol et al. [28] présentent également un travail sur l'indexation de ressources biomédicales avec des termes de MeSH.

4.3 Perspectives

Les travaux futurs se concentreront sur trois grands axes : (1) l'amélioration de l'étape de reconnaissance de concepts grâce à des techniques de traitement automatique des langues. En outre, il serait intéressant de reconnaitre également des relations, (2) l'amélioration de la personnalisation du service (paramètres utilisés pour les composants d'expansion sémantiques et sélection des ontologies à utiliser), (3) l'amélioration de l'étape d'expansion sémantique par la composition de composants existants et par le développement de

nouveaux composants.[6] Par exemple, une distance sémantique peut être extraite de l'index OBR et utilisée par l'annotateur.

La plupart des ontologies disponibles étant en anglais, nous ne travaillons pour le moment qu'avec des ressources en anglais. Cependant, la méthodologie décrite précédemment est parfaitement valide pour d'autres langues.

5 Conclusion

L'annotation sémantique de données biomédicales avec des ontologies joue un rôle crucial pour l'interopérabilité et l'intégration des données ainsi que pour favoriser les découvertes translationnelles. Cette situation est vraie de manière générale dans le domaine des e-sciences. La nécessité de passer du Web actuel à un Web sémantique pourvu d'un contenu riche annoté à l'aide d'ontologies a clairement était identifié [29]. Répondre à ce besoin exige la mise à disposition de services (utilisables aussi bien par des humains et que par des agents logiciels) qui peuvent être facilement intégrés dans les processus actuels de curation et d'annotation de données.

Nous avons présenté un service pour l'annotation sémantique de données biomédicales. Notre annotateur se distingue des précédents efforts pour plusieurs raisons :

- Il s'agit d'un service Web qui peut facilement être intégré dans des processus et des programmes (vision architecture orientée service),
- Il utilise des ontologies à la fois pour créer des annotations directes et pour l'étape d'expansion sémantique,
- Il utilise une des plus grandes collections d'ontologies biomédicales publiques disponibles (UMLS Metathesaurus et NCBO BioPortal).[7]

L'annotateur est actuellement utilisé au sein du projet NCBO afin d'annoter un grand nombre de ressources biomédicales. Le service d'annotation et l'index créé à partir de celui-ci sont tous les deux accessibles via une API de services Web (http://obs.bioontology.org). L'annotateur est disponible et est, d'ores et déjà, utilisé par la communauté. Des retours d'expérience intéressants sont à prévoir sous peu et motiveront les évolutions et améliorations futures.

Remerciements

Ce travail est supporté par le programme : National Center for Biomedical Computing (NCBC) National Institute of Health roadmap initiative; NIH grant U54 HG004028. Nous remercions également Manhong Dai and Fan Meng de Université du Michigan (NCIBI) ainsi que Chuck Kahn pour l'accès à la ressource Goldminer.

[6] Les composants utilises actuellement (relations is_a et alignements) sont les plus faciles et évidents a utiliser.

[7] L'avantage d'un tel dictionnaire n'est bien sur pas que quantitatif. Utiliser toutes ces ontologies dans la même application permet pour la première fois une large intégration dirigée par les données des ontologies de UMLS et BioPortal. Par exemple, nous travaillons sur l'extraction d'alignement entre concept de différentes ontologies dirigé par les annotations de l'index OBR. Si des milliers de données sont systématiquement annotés avec les mêmes concepts, il semble possible d'en extraire une relation entre ces concepts.

Références

[1] Clement Jonquet, Mark A. Musen, and Nigam H. Shah. A System for Ontology-Based Annotation of Biomedical Data. In A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, editors, *International Workshop on Data Integration in the Life Sciences, DILS'08*, volume 5109 of *Lecture Notes in BioInformatics*, pages 144–152, Evry, France, June 2008. Springer-Verlag.

[2] Nigam H. Shah, Clement Jonquet, Annie P. Chiang, Atul J. Butte, Rong Chen, and Mark A. Musen. Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. *BMC Bioinformatics*, 2008. Expected end of 2008.

[3] Olivier Bodenreider and Robert Stevens. Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, 7(3):256–274, August 2006.

[4] Nigam H. Shah, Daniel L. Rubin, Kaustubh S. Supekar, and Mark A. Musen. Ontology-based Annotation and Query of Tissue Microarray Data. In *American Medical Informatics Association Annual Symposium, AMIA'06*, pages 709–713, Washington DC., USA, November 2006.

[5] Robert Moskovitch, Susana B. Martins, Eytan Behiri, Aviram Weiss, and Yuval Shahar. A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. *American Medical Informatics Association*, 14(2):164–174, March-April 2007.

[6] Nigam H. Shah. *Encyclopedia of Database Systems*, chapter Biomedical Data/Content Acquisition, Curation, page In press. Springer-Verlag, New York, NY, USA, 2009.

[7] Jorge E. Caviedesa and James J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Biomedical Informatics*, 37(2):77–85, April 2004.

[8] Hisham Al-Mubaid and Hoa A. Nguyen. A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain. In *28th IEEE EMBS Annual International Conference*, pages 2713–2717, New York, NY, USA, September 2006.

[9] Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Biomedical Informatics*, 40(3):288–299, June 2007.

[10] Jerome Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Berlin Heidelberg, DE, 2007.

[11] Clement Jonquet, Mark A. Musen, and Nigam H. Shah. Help will be provided for this task: Ontology-Based Annotator Web Service. Research report BMIR-2008-1317, Stanford University, CA, USA, May 2008.

[12] Manhong Dai, Nigam H. Shah, Wei Xuan, Mark A. Musen, Stanley J. Watson, Brian D. Athey, and Fan Meng. An Efficient Solution for Mapping Free Text to Ontology Terms. In *American Medical Informatics Association Symposium on Translational Bioinformatics, AMIASTB'08*, San Francisco, CA, USA, March 2008.

[13] Weijian Xuan, Manhong Dai, Barbara Mirel, Brian Athey, Stanley J. Watson, and Fan Meng. Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In *BioLINK SIG: Linking Literature, Information and Knowledge for Biology*, pages 55–58, Vienna, Austria, July 2007.

[14] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *American Medical Informatics Association Annual Symposium, AMIA'01*, pages 17–21, Washington, DC, USA, November 2001.

[15] Nipun Bhatia, Nigam H. Shah, Daniel L. Rubin, Annie P. Chiang, and Mark A. Musen. Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. Research report BMIR-2008-1332, Stanford University, CA, USA, March 2008.

[16] Daniel L. Rubin, Suzanna E. Lewis, Chris J. Mungall, Sima Misra, Monte Westerfield, Michael Ashburner, Ida Sim, Christopher G. Hute, Harold Solbrig, Margaret-Anne Storey, Barry Smith, John Day-Richter, Natalya F. Noy, and Mark A. Musen. National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. *OMICS A Journal of Integrative Biology*, 10(2):185–198, June 2006.

[17] Mark A. Musen, Nigam H. Shah, Natasha F. Noy, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, James Buntrock, Clement Jonquet, Michael Montegut, and Daniel L. Rubin. BioPortal: Ontologies and Data Resources with the Click of a Mouse. In *American Medical Informatics Association Annual Symposium, Demonstrations, AMIA'08*, pages 1223–1224, Washington DC, USA, November 2008.

[18] Qinghua Zou, Wesley W. Chu, Craig Morioka, Gregory H. Leazer, and Hooshang Kangarloo. IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. In *American Medical Informatics Association Annual Symposium, AMIA'03*, pages 763–767, Washington DC, USA, November 2003.

[19] Lawrence H. Reeve and Hyoil Han. CONANN: An Online Biomedical Concept Annotator. In S. Cohen-Boulakia and V. Tannen, editors, *4th International Workshop Data Integration in the Life Sciences, DILS'07*, volume 4544 of *Lecture Notes in Computer Science*, pages 264–279, Philadelphia, PA, USA, June 2007. Springer-Verlag.

[20] William R. Hersh and Robert A. Greenes. SAPHIRE - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and Biomedical Research*, 23(5):410–425, October 1990.

[21] Robert Baud, Patrick Ruch, Christian Lovis, and Anne-Marie Rassinoux. Recherche conceptuelle dans les textes m´edicaux. In M. Fieschi, O. Bouhaddou, R. Beuscart, and R Baud, editors, *8emes Journées Francophones d'informatique Médicale, JFIM'00*, volume 12 of *Informatique et Sant´e*, pages 205–216, Marseille, France, May 2000. Springer-Verlag.

[22] Aysu B. Can and Nazife Baykal. MedicoPort: A medical search engine for all. *Computer Methods and Programs in Biomedicine*, 86(1):73–86, April 2007.

[23] Nicholas C. Ide, Russell F. Loane, and Dina Demner-Fushman. Essie: A Concept-based Search Engine for Structured Biomedical Text. *American Medical Informatics Association*, 14(3):253–263, May-June 2007.

[24] Maged N. Kamel-Boulos, Abdul V. Roudsari, and Ewart R. Carson. HealthCyberMap: A Semantic Visual Browser of Medical Internet Resources Based on Clinical Codes and the Human Body Metaphor. *Health Information and Libraries*, 19(4):189–200, December 2002.

[25] Khaled Khelif, Rose Dieng-Kuntz, and Pascal Barbry. An ontology-based approach to support text mining and information retrieval in the biological domain. *Universal Computer Science, Special Issue on Ontologies and their Applications*, 13(12):1881–1907, 2007.

[26] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *40th Annual Meeting of the Association for Computational Linguistics, ACL'02*, Philadelphia, PA, USA, July 2002.

[27] Olivier Corby, Rose Dieng-Kuntz, Catherine Faron-Zucker, and Fabien Gandon. Searching the Semantic Web: Approximate Query Processing Based on Ontologies. *IEEE Intelligent Systems*, 21(1):20–27, January/February 2006.

[28] Aurélie Névéol, Alexandrina Rogozan, and St´efan Darmoni. Automatic indexing of online health resources for a French quality controlled gateway. *Information processing and Management*, 42(3):695–709, May 2006.

[29] Siegfried Handschuh and Stephen Staab, editors. *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2003.

Adresse de correspondance

Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine
Medical School Office Building, Room X-215
251 Campus Drive, Stanford, CA 94305-5479 USA

{jonquet, musen, nigam}@stanford.edu

http://www.bioontology.org
http://obs.bioontology.org