Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical

Clement Jonquet, Adrien Coulet, Nigam Shah, Mark Musen

Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical

Clément Jonquet, Adrien Coulet, Nigam H. Shah, et Mark A. Musen

Center for Biomedical Informatics Research, Stanford University, CA 94305, USA
{jonquet, coulet, nigam, musen}@stanford.edu

Résumé : De nombreuses découvertes scientifiques sont contraintes aujourd'hui par la difficile intégration des données misent à disposition dans différentes ressources. 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. Dans cet article nous présentons un mécanisme d'indexation de ressources de données textuelles dirigé par les ontologies. Nous détaillons la création et l'utilisation d'un index de ressources de données biomédicales qui fournit un accès uniforme à plus d'une vingtaine de ressources indexées avec plus de 200 ontologies. Cet index est accessible via la plateforme Web BioPortal du Centre National pour les Ontologies Biomédicales (NCBO) : http://bioportal.bioontology.org/.

Mots-clés : ontologie, annotation/web sémantique, indexation sémantique, intégration de données, recherche d'information, ontologie biomédicale, Web sémantique.

Type de communication : appliquée.

1 Introduction

Un des objectifs du Web sémantique est de décrire le contenu du Web à l'aide de connaissances formalisées dans des ontologies. Cette description peut prendre la forme d'*annotations sémantiques* [Handschuh & Staab, 2003]. L'expression 'annotation sémantique' fait alors référence au processus d'identification de concepts, et de relations entre concepts, dans un document ou une donnée [Uren et al., 2006]. 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). Les ontologies sont utilisées pour représenter la connaissance de nombreux domaines. Cependant elles sont peu utilisées pour annoter les données des ressources électroniques de ces domaines pour plusieurs raisons :

- les annotations doivent souvent être crées manuellement par des experts,
- l'annotation est une tâche fastidieuse et sans retour immédiat pour l'expert,

• le nombre d'ontologies disponibles est important ce qui rend difficile leur considération par les experts. En outre, ces ontologies changent régulièrement, elles se chevauchent les unes les autres, et sont dans des formats différents.

Dans le domaine biomédical, la richesse et la diversité des ressources de données publiques permettent l'intégration translationnelle des résultats de multiples études bioinformatiques (*translational bioinformatics research*) [Butte & Chen, 2006]. Cependant, les découvertes qui pourraient être réalisées par la fouille des données biomédicales sont limitées car la plupart des ressources publiques ne sont pas décrites à l'aide de terminologies ou d'ontologies.[1] Un chercheur qui étudie les variations interindividuelles d'un gène voudrait connaître toutes les voies métaboliques dans lesquelles ce gène est impliqué, les médicaments dont les effets pourraient être modulés par les variations interindividuelles de ce gène. Il peut également être intéressé par les essais cliniques qui ont étudié ces médicaments. La connaissance nécessaire aux experts pour répondre à ces questions est aujourd'hui disponible (souvent sous la forme de texte, ou de relation valeur-attribut) dans des ressources de données biomédicales accessibles en ligne. Le problème est désormais de trouver cette information. La communauté biomédicale reconnait d'ores et déjà l'importance des ontologies pour faciliter l'intégration de données et pour faciliter la découverte de connaissance [Bodenreider & Stevens, 2006]. Cependant, la variété des données est très importante et celles-ci sont rarement annotées à l'aide de concepts décrits dans des ontologies biomédicales. Le plus souvent, les éléments d'une ressource[2] (e.g., recueil de données expérimentales, diagnostics, descriptions d'essais cliniques, publications, images) sont accompagnés par 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 qui utilisent des concepts d'ontologies pour faciliter l'indexation, l'intégration et donc la recherche, de ces données [Moskovitch et al., 2007]. Ce challenge a été relevé avec succès notamment dans le cadre de projets associés à la Gene Ontology et au vocabulaire MeSH [Trieschnigg et al., 2009]. Par exemple, Gene Ontology est largement utilisée pour décrire les fonctions moléculaires, les composants cellulaires et les processus biologiques des produits de gènes [Rhee et al., 2008]. Gene Ontology permet, entre autre, d'intégrer ces descriptions dans plusieurs bases de données, en se référant à un vocabulaire commun. Toutefois, en dehors de certains bons exemples, l'annotation sémantique des ressources biomédicales reste marginale. Cette lacune et les succès mentionnés servent de motivation à notre travail.

Dans le cadre du Centre National pour les Ontologies Biomédicales (*National Center for Biomedical Ontology* ou NCBO), nous avons conçu un workflow qui

[1] Dans le reste de l'article, nous utilisons seulement le mot « 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.

[2] Dans cet article, nous appelons *éléments* les entités identifiables (e.g., document, article, rapport d'expérimentation) qui forment une *ressource* de donnée (e.g., base de données, ensemble d'articles).

permet l'annotation sémantique de données biomédicales de façon automatique. Ce workflow permet d'utiliser plus de 200 ontologies biomédicales initialement dispersées sur le Web et définies dans différents formats e.g., Web Ontology Language (OWL) ou Open Biomedical Ontologies (OBO). Ce workflow est disponible sous la forme d'un service Web : le *NCBO Annotator* [Jonquet et al., 2009]. Il permet aux chercheurs d'utiliser les ontologies biomédicales pour annoter leurs données automatiquement. Les annotations sont ensuite renvoyées aux utilisateurs. Dans cet article, nous détaillons l'utilisation de ce workflow d'annotation pour construire un index de ressources biomédicales [Shah et al., 2009b]. Nous détaillons l'architecture de ce système ainsi que ces cas et scenarios d'utilisation.

2 Indexation de ressources avec des ontologies

2.1 Limites de l'indexation classique

De nombreuses ressources biomédicales sont publiques et peuvent être interrogées en ligne. Le contenu textuel des ressources est généralement indexé (par exemple à l'aide de l'interface de programmation Lucene) pour permettre l'interrogation de la ressource par mot clés. Les limites de l'indexation par mots clés se font cependant souvent ressentir. Par exemple, les mots sont bien souvent polysémiques. Ainsi en regardant le mot anglais *cell*, un système peut difficilement faire la distinction entre la cellule vivante ou le téléphone cellulaire. Toutefois, les ontologies formalisent les concepts au voisinage de *cell* (e.g., *Microanatomic Structure, Stem Cell*) et cette connaissance peut être utilisée pour supprimer l'ambigüité lorsque ce mot apparait dans un texte. Autre exemple, une recherche sur une ressource du mot clé *cancer* ne renvoie bien souvent pas les éléments de la ressource qui contiennent l'expression *malignant neoplasm* qui est pourtant un synonyme de *cancer*. Par ailleurs, une recherche du mot clé *pheochromocytoma* sur la base de données d'expressions génétiques GEO (Gene expression Omnibus) renvoie 19 résultats. Cependant, une recherche du mot clé *retroperitoneal neoplasm* ne renvoie aucun résultat tandis que *pheochromocytoma* est un type de *retroperitoneal neoplasm* et donc les 19 résultats précédents sont appropriés et devraient également être renvoyés. Ces exemples illustrent clairement l'échec des systèmes de recherche qui ne font pas appel à la connaissance médicale. En biomédecine, comme dans de nombreux domaine, les chercheurs se sont tournés vers les ontologies pour représenter cette connaissance et permettre à des programmes de l'utiliser. Ainsi, le NCI Thesaurus, une ontologie développée en OWL représente formellement les trois unités de connaissance mentionnées précédemment. Cette ontologie, si associée aux mécanismes de raisonnement adéquats, permettrait de dépasser les limites de l'indexation classique.

2.2 Indexation dirigée par les ontologies

Au sein du projet NCBO, nous avons développé un système d'indexation de données dont l'objectif est d'indexer les éléments de diverses ressources biomédicales publiques non pas par mot clé, mais par concept d'ontologie. Pour chaque élément

d'une ressource de données, le mécanisme d'indexation repose sur la génération d'annotations sémantiques et sur l'agrégation des ces annotations. Ci-après, nous détaillons ce mécanisme à l'aide d'un exemple d'indexation d'un élément de la ressource GEO i.e., GDS1965 (figure 1). Le processus d'indexation est composé de quatre étapes :

1. Récupération 'contextualisée' des données : Tout d'abord, chaque ressource de données est accédée à l'aide d'un *accesseur* spécifique qui récupère de façon incrémentale chaque élément de la ressource. Le plus souvent les ressources proposent un service Web pour accéder à leurs éléments. En outre, les accesseurs sont développés pour identifier et récupérer régulièrement les mises à jour des ressources. Pour chaque ressource, un expert statue sur les champs de métadonnées textuelles qui devront être récupérés (e.g., titre, description) et affecte à chaque champ un poids (entre 0 et 1) attestant de son importance ; ce poids sera utilisé pour scorer les annotations. L'impact de la 'contextualisation' a était démontré en recherche d'information [Moskovitch et al., 2007] et il semble évident qu'il ne faut pas donner à une annotation faite dans le titre d'un document la même importance qu'à une annotation faite dans sa description. L'accesseur récupère aussi, lorsque elles existent, les références à des ontologies ; c'est-à-dire des annotations sémantiques déjà reportées dans un champ spécifique de l'élément. C'est notamment le cas dans notre exemple où le champ *organism* réfère à un concept d'une ontologie qui informe sur l'organisme sur lequel l'expérimentation d'expression génétique a était pratiquée.

2. Génération des annotations directes : Chaque champ de métadonnées textuelles est traité par un *outil de reconnaissance de concept* qui détecte la mention d'un concept d'ontologie. Le système peut accepter différents outils de reconnaissance de concept allant de la simple reconnaissance syntaxique à des approches utilisant le traitement automatique des langues. Ces outils utilisent généralement un dictionnaire (ou lexique). Ce dictionnaire est une liste de termes qui identifient des concepts définis dans des ontologies.[3] Il est construit en récupérant à partir d'un ensemble d'ontologies biomédicales tous les noms ou synonymes qui identifient syntaxiquement les concepts. Sur notre exemple, les termes `melanoma`, `melanocyte`, et `cell` sont identifiés et permettent la génération d'un ensemble d'*annotations directes* avec les concepts correspondants dans les ontologies Human Disease (DOID), Cell type (CL) et BIRNLex (birnlex). Le terme reconnu, le champ textuel ou il est mentionné, ainsi que sa position dans le champ, sont conservés comme information de provenance de l'annotation. Les références à des ontologies sont également traitées lors de cette étape et l'identifiant adéquat est utilisé pour créer des *annotations directes reportées*.

3. Expansion sémantique des annotations : L'ensemble des annotations directes est ensuite traité par plusieurs composants d'expansion sémantique qui créent de nouvelles *annotations étendues* à partir des concepts formant les annotations directes. Ces annotations étendues sont dérivées à partir de la connaissance représentée dans les ontologies. Par exemple:

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

• La hiérarchie is-a permet de créer de nouvelles annotations avec les concepts parents d'un concept constituant une annotation directe. Ainsi, si un texte est annoté directement avec le concept `melanoma` de l'ontologie Human Disease, ce composant d'expansion sémantique génère de nouvelles annotations avec les concepts parents de melanoma (`cancer` et `cellular proliferation disease`) comme l'illustre la figure 1. Le niveau relatif dans la hiérarchie est conservé comme information de provenance de l'annotation. Il semble évident que plus le concept parent est distant, moins l'annotation sera précise (tout en restant valide) puisque le concept dans l'ontologie est supposé beaucoup plus général.

• Les mappings ou alignements (non représenté sur la figure 1) entre concepts permettent de créer de nouvelles annotations à partir des concepts constituants une annotation directe. Si un texte est annoté directement avec le concept `melanoma` du NCI Thesaurus (`NCI/C0025202`), ce composant d'expansion sémantique génère une nouvelle annotation avec le concept `melanoma` de Human Disease (`DOID:1909`) parce qu'il existe un alignement entre ces deux concepts. Dans ce travail, les alignements entre ontologies sont considérés comme déjà existants et utilisables. Nous ne considérons que des alignements un-à-un Le type d'alignement (e.g., manuel ou automatique) est conservé comme information de provenance de l'annotation.

4. Agrégation et score des annotations : A l'issue des deux étapes précédentes, chaque annotation provient d'un champ de métadonnées textuelles de la ressource (i.e., 'contexte'). En outre, l'information de provenance d'une annotation permet de lui affecter un *poids* représentant son importance (Table 1). Il peut exister pour un même élément de nombreuses annotations faites avec le même concept mais dans des contextes différents et/ou de provenances différentes. Un regroupement par paire unique « concept-élément » est alors effectué pour créer des *annotations agrégées* qui permettent de tenir compte : *(i)* du poids du champ annoté ; *(ii)* de la fréquence des annotations ; *(iii)* du type de chaque annotation. Un *score* final est attribué à chaque annotation agrégée. Ce score correspond à la somme des poids des annotations faites avec un même concept, pondérés par les poids des champs d'origine de ces annotations.

A l'issue de l'indexation d'un élément, deux niveaux de détail sont donc disponibles pour consulter et exploiter les annotations de l'index : *(1)* le niveau des annotations agrégées ; *(2)* le niveau des annotations directes et étendues.

Table 1. Poids (entre 1 et 10) d'une annotation en fonction de sa provenance.

Provenance	**Poids**
Annotation directe faite avec le terme « préféré » d'un concept	10
Annotation directe faite avec un terme synonyme d'un concept	8
Annotation reportée	10
Annotation étendue par l'utilisation des alignements	7
Annotation étendue par la hiérarchie is-a (pour un parent niveau n) (e.g., 9 for n=1; 7 for n=2; 4 for n=5; 3 for n=8; 1 for n>12)	$1+10*e^{-0.2*n}$

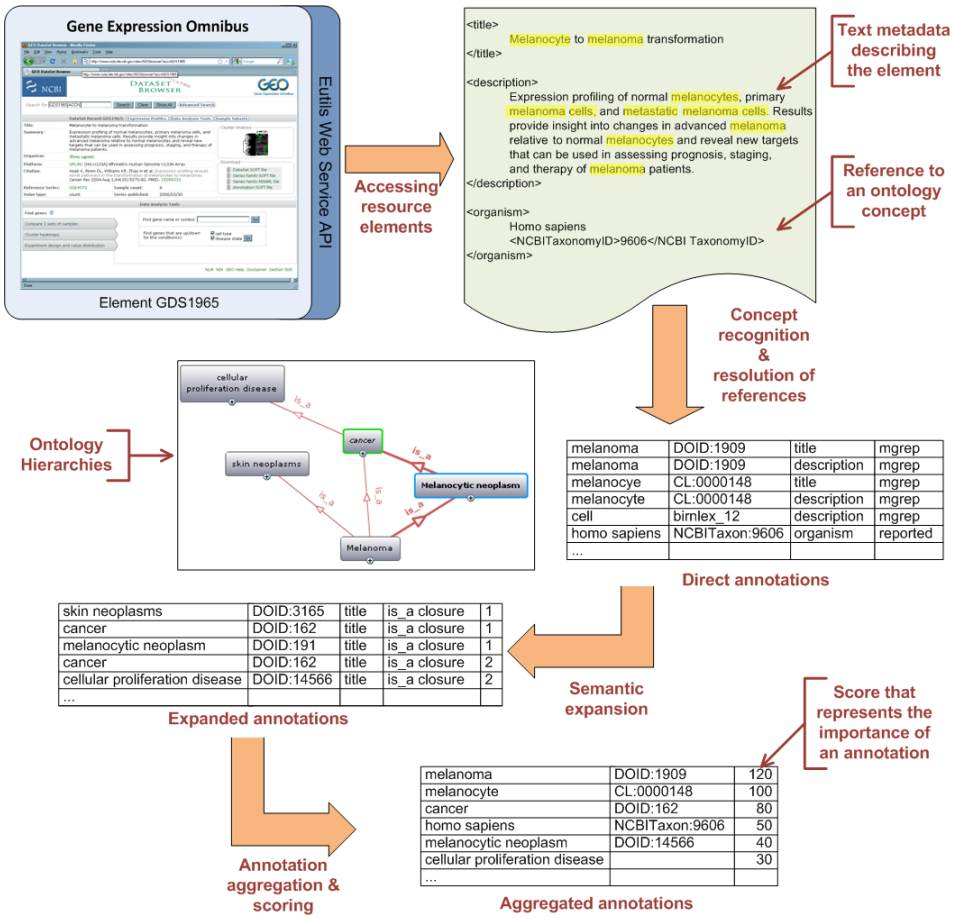

Fig. 1 – Génération des annotations pour un élément de la ressource GEO. Des *annotations directes* sont générées à partir des métadonnées textuelles et des références existantes de l'élément. Puis des *annotations étendues* sont créées en utilisant les hiérarchies des ontologies. Puis, les annotations sont agrégées et scorées en tenant en compte leur fréquence et leur contexte.

3 L'index de ressource biomédicale du NCBO

Le NCBO développe et maintient une application Web appelée *BioPortal* (http://bioportal.bioontology.org) qui permet d'accéder, visualiser, rechercher et commenter des ontologies biomédicales [Noy et al., 2009]. Le portail contient une grande collection d'ontologies, et permet de gérer leur évolution dans le temps. Les utilisateurs ont accès à la plateforme soit via l'interface graphique, soit via une interface de service Web.

BioPortal, nous fournit un accès uniforme à un grand nombre d'ontologies. Ainsi, nous avons mis en œuvre le système d'indexation présenté précédemment avec une des plus grandes collections d'ontologies biomédicales publiques disponibles (215 ontologies au 04/02/2010). Ces ontologies nous permettent de générer un dictionnaire contenant 3 653 128 concepts et 7 061 411 termes. Comme outil de reconnaissance de concept, le système utilise Mgrep [Dai et al., 2008], développé à l'Université du Michigan qui présente un degré de précision élevé pour la reconnaissance par exemple de noms de maladie [Xuan et al., 2007]. Nous avons réalisé une évaluation comparative de Mgrep avec MetaMap [Shah et al., 2009a], 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 dictionnaires. Mgrep s'est montré très rapide avec une précision plus élevée pour la plupart des ressources de données traitées et pour les quatre dictionnaires. En outre, Mgrep est ouvert à tout type de dictionnaire et n'est pas limité, comme MetaMap, au métathesaurus UMLS (Unified Medical Langage System). Nous avons donc pu utiliser Mgrep pour traiter les ontologies NCBO qui sont généralement disponibles dans les formats OBO et OWL. Pour plus de détail sur cette évaluation voir [Shah et al., 2009a].

Table 2. Précision de Mgrep et MetaMap avec deux dictionnaires.

Ressource de données biomédicale	Maladies (termes d'UMLS)		Biological processes (GO)	
	Mgrep	MetaMap	Mgrep	MetaMap
ClinicalTrials.gov	0,87	0,71	0,6	0,63
Gene Expression Omnibus	0,88	0,755	0,93	0,73
ARRS GoldMiner	0,73	0,548	0,58	0,33

Nous avons indexé 22 ressources biomédicales publiques (au 04/02/2010 cf. http://rest.bioontology.org/resource_index/resources/list/) de différentes tailles (e.g., <800k éléments et <180Mo). Les domaines couverts par ces ressources sont divers et comprennent des données d'expression génétique (e.g., ArrayExpress, GEO), des descriptions d'essais cliniques (Clinicaltrials.gov), ou des légendes et descriptions d'images radiologiques (e.g., ARRS Goldminer). L'index contient plus de 2 milliards d'annotations agrégées et 10 milliards d'annotations détaillés (figure 2). L'index fournit des annotations pour tous les éléments des 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 (REpresentational State Transfer) et les annotations peuvent être retournées aux utilisateurs dans différents formats: texte, tab-delimited, XML ou RDF/OWL. Le contenu de l'index est également accessible via l'interface graphique du BioPortal.

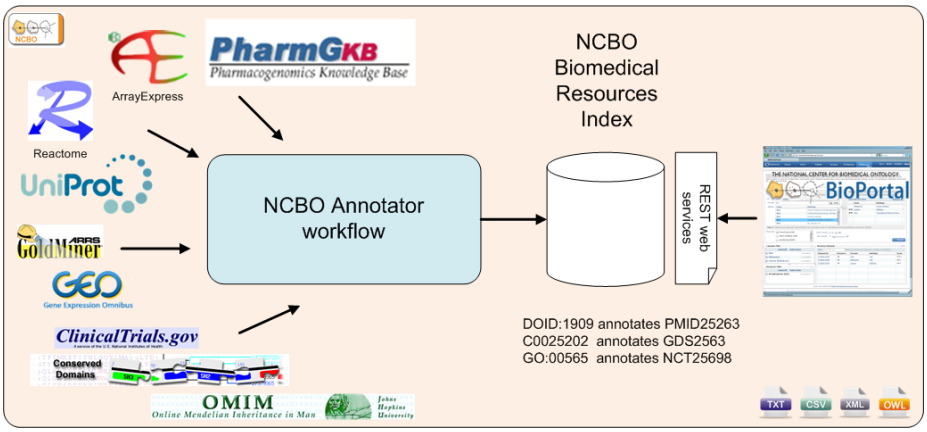

Fig. 2 – Index de ressources biomédicales du NCBO. Chaque ressource est traitée avec le mécanisme d'indexation (section 2). Les annotations générées sont uniformément accessibles via une API de services Web. Le *BioPortal* permet la recherche d'information.

4 Cas d'utilisation et scénarios supportés

Les annotations de l'index peuvent êtres accédées *par concept(s)* ou *par élément*. Ces deux modes permettent différents cas d'utilisation :

1. l'obtention, pour un concept donné, de l'ensemble des éléments d'une ou plusieurs ressources annotés avec ce concept e.g., les éléments de GEO et ArrayExpress annotés avec `DOID:1909` ;
2. l'obtention, pour un ensemble de concepts donnés, de l'intersection ou de l'union des ensembles d'éléments annotés avec ces concepts e.g., les éléments de GEO et ArrayExpress annotés à la fois par `DOID:1909` et `CL:0000148` ;
3. l'obtention, pour un élément d'une ressource, de l'ensemble des concepts d'une ou plusieurs ontologies qui annotent cet élément e.g., les concepts du NCI thesaurus qui annotent l'élément GDS1965 de GEO.

Les cas 1 et 2 sont des cas d'utilisation de recherche d'information dans un ensemble de ressources hétérogènes. L'index est dans ce cas utilisé comme un outil d'intégration de données qui propose une interface unique pour interroger plusieurs ressources dont les modèles de représentation sont différents [Lenzerini, 2002]. Nous avons pu remarquer qu'une recherche faite avec une conjonction de peu de concept (i.e., 2 ou 3) réduit suffisamment le nombre de résultats pour les rendre facilement consultable. Le troisième cas d'utilisation est plus atypique dans le sens où il informe sur le contenu d'un élément déjà identifié au sein d'une ressource.

Dans les trois cas d'utilisation, les résultats sont retournés aux utilisateurs sous formes d'un ensemble d'annotations. Les scores affectés aux annotations lors de l'indexation permettent de classer les résultats par ordre de pertinence. Il est possible d'obtenir directement de l'index des annotations détaillées, mais le plus souvent, il est

plus intéressant d'obtenir des annotations agrégées et si nécessaire, obtenir le détail des annotations à l'origine de cette agrégation. Un service Web est disponible pour chacun de ces trois cas d'utilisation. En conséquence, des utilisateurs peuvent composer ces services pour supporter des cas d'utilisation plus complexes. Par exemple, l'obtention, pour une ressource et un concept donné, des éléments annotés avec le meilleur score (cas 1), puis pour l'ensemble de ces éléments obtenir les concepts les plus représentatifs (cas 3). Une telle composition permet de découvrir les concepts qui sont fréquemment associés au concept initial et ainsi d'affiner la recherche sur la ressource. En outre, ces trois cas d'utilisation sont opérationnels pour l'index de ressources biomédicales du NCBO via trois interfaces graphiques différentes dans BioPortal. Par exemples, l'interface représentée par la figure 2 de [Shah et al., 2009b] correspond au cas d'utilisation 1. Pour les cas 2 et 3 voir http://bioportal.bioontology.org/resources et http://bioportal.bioontology.org/annotator.

Ces trois cas supportent différents scenarios de recherche, par exemples :

• Les chercheurs travaillant sur le projet Trialbank (http://www.trialbank.org) à l'Université de Californie à San Francisco, utilisent des annotations d'essais cliniques sur le VIH/sida de ClinicalTrials.gov afin de développer une application Web pour visualiser et comparer les essais. En particulier, ils exploitent l'origine des annotations (i.e., le champ de métadonnées) pour designer leur interface.

• Des chercheurs de l'université de Washington utilisent l'index pour connecter de nombreuses ressources de données sur les nanoparticules. Ils exploitent les ontologies pour la recherche d'information sur ces ressources.

• Des chercheurs du Medical College of Wisconsin utilisent les annotations d'expérimentation d'expression génétique de GEO pour fouiller cette ressource et accélérer significativement la construction d'une base de données sur le génome du rat qui supporte leur recherche sur les associations gène-maladies et gène-phénotype. Ils ont développés une plateforme Web de curation et d'exploration des annotations de l'index (http://gminer.mcw.edu).

• A l'Université Stanford, nous utilisons l'index pour fouiller des données relatives au financement de projets scientifiques et à leur impact en termes de publications. Nous utilisons les concepts d'ontologies comme dénominateur commun entre une ressource de données financières (http://www.researchcrossroads.org/) et une ressource de publications (PubMed). Nous évaluons alors si les tendances des publications reflètent, dans le temps, celles des financements.

5 Etat de l'art

L'annotation sémantique est un sujet de recherche largement étudié par la communauté du Web sémantique [Handschuh & Staab, 2003]. Une revue et un comparatif des outils du domaine ont été proposées par Uren *et al.* [Uren et al., 2006]. Dans le domaine biomédical, la perspective de la découverte de nouvelles connaissances fait de l'annotation automatique de ressources un sujet important. Il existe plusieurs outils de reconnaissance de concept qui permettent d'identifier des entités d'ontologies (concept ou relation) dans du texte. Par exemples, MetaMap

[Aronson, 2001], CONANN [Reeve & Han, 2007] et Mgrep [Xuan et al., 2007], ou [Baud et al., 2000].

Notons également que l'usage de connaissances représentées dans les ontologies est fréquent dans le domaine de la recherche d'information [Bhogal et al., 2007] et de l'intégration de données [Jalabert et al., 2006]. De nombreux projets utilisent les ontologies pour améliorer les performances de moteurs de recherche [Guelfi et al., 2007]. Par exemple, MedicoPort [Can & Baykal, 2007] utilise les relations de l'UMLS pour faire de l'expansion de requête. L'outil Essie [Ide et al., 2007] a permis de montrer qu'une combinaison de la structure des documents et de l'expansion sémantique de concept est utile pour la recherche d'information. La plupart de ces outils utilisent seulement l'UMLS ou un petit nombre d'ontologies. Cette limitation donne à l'index de ressources biomédicales du NCBO un avantage important car il utilise, en plus de l'UMLS, les ontologies disponibles sur le BioPortal.

En outre, de nombreux travaux présentent les résultats de l'indexation d'une ressource unique avec une ou quelques ontologies (suivant une approche alors dite « verticale »). Par exemples, [Khelif et al., 2007] ont annoté la ressource GeneRIF en utilisant une plateforme qui leur permet de reconnaitre non seulement des concepts, mais aussi des relations. [Trieschnigg et al., 2009] ou [Névéol et al., 2006] présentent également un travail sur l'indexation de ressources biomédicales avec des termes du vocabulaire MeSH. Ces travaux suivent parfois des approches plus sophistiquées que celle décrite ici mais génèrent usuellement des résultats restreints à un domaine ou à un cas d'utilisation. L'index du NCBO adopte une approche plus « horizontale » en proposant publiquement, avec une interface unique, des annotations d'ores et déjà calculées (i.e., ce qui libère l'utilisateur de cette tache) pour un grand nombre de ressources et plus de 200 ontologies. Cette approche est en accord avec le principe d'intégration translationelle de résultats d'études bioinformatiques selon lequel les nouvelles connaissances peuvent provenir de croisement de données et de connaissances (ici ressources et ontologies) qui n'avaient pas été anticipés.

D'autres approches comme SemanticHacker (http://semantichacker.com) or OpenCalais (http://www.opencalais.com) utilisent des méthodes de fouilles de textes et s'appuient sur une sémantique précisément définie pour mettre en évidence de nouvelles connaissances. Cependant ces approches ne prennent pas en considération les connaissances déjà formalisées dans les ontologies. Dans le domaine biomédical, la quantité et la richesse du contenu des ontologies rendent leur utilisation particulièrement intéressante et obligatoire.

6 Discussion et 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. Nous avons présenté un système d'indexation données biomédicales à base d'ontologies qui s'inscrit dans cette vision. Ce système est transposable dans d'autres domaines où la dispersion des données textuelles empêche l'intégration de données.

Dans le domaine biomédical, la masse de ressources et d'ontologies à traiter rendent l'indexation de données avec des ontologies difficile. En outre, ces ressources et ces ontologies évoluent au fil du temps et maintenir un index est alors une tache fastidieuse. L'index du NCBO gère ces situations.

L'exactitude des annotations de l'index est liée d'une part à la qualité des ontologies utilisées (e.g., composition du dictionnaire) et d'autre part à la qualité de l'outil de reconnaissance de concept utilisé. Extraire des concepts à partir de texte est difficile, c'est pourquoi le mécanisme d'indexation a été conçu de façon à inter-changer des outils de reconnaissance de concept autres que Mgrep. Nous travaillons à améliorer cette étape grâce aux techniques de traitement automatique des langues.

Bien que notre objectif soit de fournir un index couvrant un grand nombre de ressources biomédicales, leur grand nombre et leur rapide évolution rend impossible de toutes les indexer. C'est pourquoi nous proposons à la communauté le workflow d'annotation sous forme de Web service, le NCBO Annotator [Jonquet et al., 2009]. En utilisant celui-ci, les responsables d'une ressource peuvent annoter leurs données dans un format compatible avec celui de l'index.

Références

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

[Baud *et al.*, 2000] BAUD R., RUCH P., LOVIS C. & RASSINOUX A.-M. (2000). Recherche conceptuelle dans les textes médicaux. In M. FIESCHI, O. BOUHADDOU, R. BEUSCART & R. BAUD, Eds., *8èmes Journées Francophones d'informatique Médicale, JFIM'00*, volume 12 of *Informatique et Santé*, p. 205–216, Marseille, France: Springer-Verlag.

[Bhogal *et al.*, 2007] BHOGAL J., MACFARLANE A. & SMITH P. (2007). A review of ontology based query expansion. *Information Processing and Management*, **43**, 866–886.

[Bodenreider & Stevens, 2006] BODENREIDER O. & STEVENS R. (2006). Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, **7**(3), 256–274.

[Butte & Chen, 2006] BUTTE A. J. & CHEN R. (2006). Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. In *AMIA Annual Symposium*, p. 106–110, Washington DC, USA.

[Can & Baykal, 2007] CAN A. B. & BAYKAL N. (2007). MedicoPort: A medical search engine for all. *Computer Methods and Programs in Biomedicine*, **86**(1), 73–86.

[Dai *et al.*, 2008] DAI M., SHAH N. H., XUAN W., MUSEN M. A., WATSON S. J., ATHEY B. D. & MENG F. (2008). An Efficient Solution for Mapping Free Text to Ontology Terms. In *AMIA Symposium on Translational Bioinformatics*, San Francisco, CA, USA.

[Guelfi *et al.*, 2007] GUELFI N., PRUSKI C. & REYNAUD C. (2007). Les ontologies pour la recherche ciblée d'information sur le Web : une utilisation et extension d'OWL pour l'expansion de requêtes. In F. TRICHET, Ed., *18èmes Journées francophones d'Ingénierie des Connaissances, IC'07*, p. 61–73, Grenoble, France: Cepadues.

[Handschuh & Staab, 2003] S. HANDSCHUH & S. STAAB, Eds. (2003). *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

[Ide *et al.*, 2007] IDE N. C., LOANE R. F. & DEMNER-FUSHMAN D. (2007). Essie: A Concept-based Search Engine for Structured Biomedical Text. *American Medical Informatics Association*, **14**(3), 253–263.

[Jalabert *et al.*, 2006] JALABERT F., RANWEZ S., DEROZIER V. & CRAMPES M. (2006). i^{2}dee: An Integrated and Interactive Data Exploration Environment Used for Ontology Design. In *15th International Conference on Knowledge Engineering and Knowledge Management, EKAW'06*, volume 4248 of *LNAI*, p. 256–271, Podebrady, Czech Republic: Springer-Verlag.

[Jonquet *et al.*, 2009] JONQUET C., SHAH N. H. & MUSEN M. A. (2009). The Open Biomedical Annotator. In *AMIA Summit on Translational Bioinformatics*, p. 56–60, San Francisco, CA, USA.

[Khelif *et al.*, 2007] KHELIF K., DIENG-KUNTZ R. & BARBRY P. (2007). An ontology-based approach to support text mining and information retrieval in the biological domain. *Universal Computer Science*, **13**(12), 1881–1907.

[Lenzerini, 2002] LENZERINI M. (2002). Data Integration: A Theoretical Perspective. In L. POPA, Ed., *21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS'02*, p. 233–246, Madison, WI, USA.

[Moskovitch *et al.*, 2007] MOSKOVITCH R., MARTINS S. B., BEHIRI E., WEISS A. & SHAHAR Y. (2007). A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. *American Medical Informatics Association*, **14**(2), 164–174.

[Noy *et al.*, 2009] NOY N. F., SHAH N. H., WHETZEL P. L., DAI B., DORF M., GRIFFITH N. B., JONQUET C., RUBIN D. L., STOREY M.-A., CHUTE C. G. & MUSEN M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, **37**, 170–173.

[Névéol *et al.*, 2006] NÉVÉOL A., ROGOZAN A. & DARMONI S. (2006). Automatic indexing of online health resources for a French quality controlled gateway. *Information Processing and Management*, **42**(3), 695–709.

[Reeve & Han, 2007] REEVE L. H. & HAN H. (2007). CONANN: An Online Biomedical Concept Annotator. In S. COHEN-BOULAKIA & V. TANNEN, Eds., *4th International Workshop Data Integration in the Life Sciences, DILS'07*, volume 4544 of *Lecture Notes in Computer Science*, p. 264–279, Philadelphia, PA, USA: Springer-Verlag.

[Rhee *et al.*, 2008] RHEE S. Y., WOOD V., DOLINSKI K. & DRAGHICI S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, **9**, 509–515.

[Shah *et al.*, 2009a] SHAH N. H., BHATIA N., JONQUET C., RUBIN D. L., CHIANG A. P. & MUSEN M. A. (2009a). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, **10**(9:S14).

[Shah *et al.*, 2009b] SHAH N. H., JONQUET C., CHIANG A. P., BUTTE A. J., CHEN R. & MUSEN M. A. (2009b). Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. *BMC Bioinformatics*, **10**(2:S1).

[Trieschnigg *et al.*, 2009] TRIESCHNIGG D., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval. *Bioinformatics*, **25**(11), 1412–1418.

[Uren *et al.*, 2006] UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, **4**(1), 14–28.

[Xuan *et al.*, 2007] XUAN W., DAI M., MIREL B., ATHEY B., WATSON S. J. & MENG F. (2007). Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In *BioLINK: Linking Literature, Information and Knowledge for Biology, SIG, ISMB'08*, p. 55–58, Vienna, Austria.