Web sémantique pour la mémoire d'expériences d'une communauté scientifique : le projet MEAT

Khaled Khelif*, Rose Dieng-Kuntz*, Pascal Barbry**

* INRIA Sophia Antipolis 2004, route des Lucioles 06902
Sophia Antipolis - FRANCE
{Khaled.Khelif, Rose.Dieng}@sophia.inria.fr
** IPMC 660, route des Lucioles 06560
Sophia Antipolis - FRANCE
Barbry@ipmc.fr

Résumé. Cet article décrit le projet MEAT (Mémoire d'Expériences pour l'Analyse du Transcriptome) dont le but est d'assister les biologistes travaillant dans le domaine des puces à ADN, pour l'interprétation et la validation de leurs résultats. Nous proposons une aide méthodologique et logicielle pour construire une mémoire d'expériences pour ce domaine. Notre approche, basée sur les technologies du web sémantique, repose sur l'utilisation des ontologies et des annotations sémantiques sur des articles scientifiques et d'autres sources de connaissances du domaine. Notre approche peut être généralisée à d'autres domaines requérant des expérimentations et traitant un grand flux de données (protéomique, chimie,etc.).

1 Introduction

De plus en plus de connaissances scientifiques sont accessibles soit grâce à des documents publiés sur le web, soit dans des bases de données. Certaines de ces connaissances reposent sur des interprétations humaines de résultats d'expériences. Ces connaissances sont, entre autres, indispensables pour la vérification, la validation ou l'enrichissement du travail des chercheurs du domaine considéré. Mais la quantité énorme de données provenant de sources internes ou externes aux organisations rend très difficile la détection, le stockage et l'exploitation de ces connaissances. Ceci est le cas de la recherche dans le domaine de la biologie moléculaire et plus particulièrement dans le domaine des puces à ADN.

Les biologistes travaillant dans ce domaine manipulent de grandes quantités de données dans différentes conditions expérimentales et doivent se référer à des milliers de publications scientifiques liées à leurs expériences. Ces biologistes ont donc sollicité un support méthodologique et logiciel qui les aiderait dans la validation et/ou l'interprétation de leurs résultats et qui leur faciliterait la planification de nouvelle expérimentation.

C'est dans ce contexte que le projet MEAT a été proposé en fournissant des solutions permettant de remédier à ces problèmes.

Après la présentation du contexte général et de la problématique de ce travail, nous détaillons notre approche adoptée pour MEAT (Khelif et al, 2005) ainsi que les différentes

composantes de notre architecture et nous concluons avec une comparaison avec des travaux similaires.

1.1 Contexte

La technologie des puces à ADN a été développée après le séquençage dans le but de découvrir les fonctions des gènes dans différents contextes biologiques. Ces expériences permettent l'accès à des milliers de gènes simultanément et fournissent une masse énorme de données, ce qui engendre des difficultés pour les biologistes particulièrement dans la validation et l'interprétation des résultats obtenus.

Les besoins exprimés par les biologistes travaillant dans ce domaine sont:

- Une vue sur les expériences connexes : essayer d'identifier des relations entre les expériences (bds locales ou en ligne) et de découvrir des nouvelles pistes à explorer.
- Aide à la validation des résultats expérimentaux : en recherchant dans les articles traitant le phénomène étudié des informations qui argumentent, confirment ou infirment leurs hypothèses de départ, ce qui nécessite une richesse dans les annotations.
- Aide à l'interprétation des résultats : en identifiant de nouvelles relations entre gènes et/ou les interactions pouvant exister entre eux, avec des composants cellulaires ou des processus biologiques.

Ces besoins nous ont conduits à réaliser le projet MEAT en collaboration avec les biologistes de la plate-forme puce à ADN de Sophia Antipolis (localisée à l'IPMC[1]) qui distribue des puces pour les autres laboratoires français. Ce projet nous permet ainsi d'explorer l'intérêt d'un « web sémantique organisationnel » sur l'échelle d'une communauté.

1.2 Mémoire d'entreprise et web sémantique

Actuellement, les techniques du web sémantique peuvent jouer un rôle très important dans la gestion des connaissances et la construction des mémoires d'entreprise. En effet, les ontologies peuvent être utilisées dans la représentation des connaissances en fournissant un cadre formel pour décrire les différentes sources de connaissances et en guidant la création d'annotations sémantiques facilitant la description, le partage et l'accès à ces sources.

Dans (Dieng-kuntz, 2005), il est proposé de matérialiser une mémoire d'entreprise à travers un « web sémantique d'entreprise » en utilisant les ontologies pour formaliser le vocabulaire partagé dans une communauté, et les annotations sémantiques basées sur ces ontologies pour décrire les sources de connaissances hétérogènes (corpus textuels, base de données, experts…) et faciliter leurs accès via Internet/Intranet.

Dans notre cas, ce web sémantique d'entreprise est constitué des composants suivants :

- Les ressources : les bases de données des expériences, les biologistes et les articles provenant de sources internes (base documentaire locale de chaque biologiste) ou de sources externes (articles provenant du web).
- Les ontologies : nous proposons MeatOnto, une ontologie modulaire composée de trois ontologies (cf. §3.1).
- Les annotations sémantiques : elles décrivent les expériences stockées dans les bases de données (résultats, interprétations), les connaissances extraites des articles scientifiques

[1] http://www.ipmc.cnrs.fr/

et les autres ressources du domaine. Ces annotations peuvent être générées manuellement ou automatiquement (grâce à techniques de fouille de textes).

2 Notre approche

Afin de tenir compte de tous les besoins des biologistes, nous avons proposé de construire une mémoire d'expérience pour le domaine des puces à ADN.

Nous avons donc commencé par faire l'inventaire de toutes les sources qui peuvent constituer cette mémoire :

- MEDIANTE[2]: un système d'information pour les expériences puces à ADN développé au sein de l'IPMC. Il permet la gestion des projets d'expériences en partant de la conception des puces jusqu'au stockage des résultats.
- Les journaux scientifiques: pour chaque projet, les biologistes constituent un corpus de documents concernant les gènes qui sont à priori intéressants pour l'expérimentation. Ces articles sont généralement sélectionnés à partir de journaux renommés (Annual Reviews, Physiological Reviews, Pharmacological Reviews…). L'hypothèse que nous avons posée consiste à dire que cette sélection garantit la qualité et l'authenticité du contenu et nous permet d'extraire des connaissances pertinentes à partir de ces articles, contrairement à d'autres approches qui font un scannage de bases de données documentaire tel que PubMed[3].
- Les points de vue des biologistes: l'interprétation des résultats, corrélation/connexion de phénomènes ou d'expériences…

La figure 1 résume l'architecture de MEAT et récapitule les différentes étapes de notre approche :

1. *La construction d'une ontologie* décrivant les différentes ressources utilisées par les biologistes (base de données d'expériences, articles scientifiques, entités biomédicales…) ; Cette ontologie a été baptisée *MeatOnto.*
2. *La structuration des connaissances* contenues dans les champs de la base de données MEDIANTE afin de faciliter la tâche de recherche d'informations pour les biologistes qui ont pour but de trouver des similarités/corrélations entre les expériences ; Dans cette étape nous avons fait intervenir *les annotations sémantiques basées sur MeatOnto*.
3. *La génération automatique d'annotations sémantiques* basées sur *MeatOnto* pour les articles jugés intéressants par les biologistes ; pour cela nous avons conçu et implémenté le module *MeatAnnot*.
4. *La fourniture d'interfaces* aux biologistes permettant l'ajout d'annotations sur les expériences ou les articles ; ce qui a mené au développement du module *MeatEditor.*
5. *L'aide à la validation des résultats expérimentaux* en proposant une recherche documentaire guidée par MeatOnto et utilisant les annotations sémantiques; nous avons donc développé un module nommé *MeatSearch*.
6. *L'aide à l'interprétation des résultats* en faisant des inférences plus avancées sur les annotations sémantiques pour expliquer un comportement particulier; ce qui mène à d'autres fonctionnalités du module *MeatSearch*.

[2] http://microarray.fr:8080/mediante/index
[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

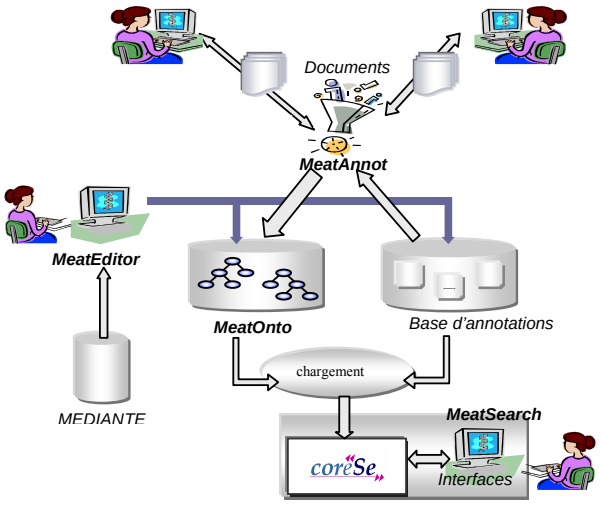

FIG. 1 - *Architecture de Meat*

3 Les composantes de MEAT

3.1 L'ontologie : MeatOnto

Comme nous l'avons décrit précédemment notre but était de construire une ontologie qui décrit toutes les ressources du domaine des puces à ADN, nous avons donc opté pour une ontologie modulaire composée de trois sous-ontologies dédiées à différentes parties :

- *UMLS* : ce projet élaboré par la NLM (National Library of Medecine de Bethesda) propose depuis 1986 de mettre au point un langage médical unifié (Humphreys et al, 1993). Pour ce faire, ce langage repose sur : (1) un métathesaurus qui énumère tout le vocabulaire médical existant et qui comprend des millions de termes ; (2) un réseau sémantique constitué d'une hiérarchie de 134 types sémantiques et de 54 relations ; il représente une classification de tous les concepts représentés dans le métathesaurus ainsi que les relations pouvant exister entre eux. Par analogie, nous avons considéré, le réseau sémantique de UMLS comme une ontologie : la hiérarchie des types sémantiques est la hiérarchie des concepts et les termes du métathesaurus sont des instances de ces concepts.
- *MGED*[4] : c'est une ontologie proposée pour décrire les expériences des puces à ADN afin de faciliter le partage des résultas (Stoeckert, 2003). Nous avons utilisé cette ontologie dans MEAT afin de décrire les expériences stockées dans MEDIANTE.
- *DocOnto* : Nous avons développé cette ontologie pour décrire des métadonnées sur les articles (auteurs, sources…) et sur les annotations (generated_by, validated_by…). Elle

[4] http://www.mged.org/

représente aussi la structure des articles (abstract, sentence, relation…) et fait le lien entre les articles et les concepts de UMLS (has_relation, speaks_about_genes…).

Nous avons effectué le codage de l'ontologie UMLS automatiquement à l'aide d'un script permettant de traduire le réseau sémantique de son format textuel vers une ontologie représentée dans le format RDFS (McBride, 2004). L'ontologie *DocOnto* a été construite progressivement pour couvrir tous nos besoins concernant la description des connaissances contenues dans les articles.

3.2 MeatAnnot

3.2.1 Génération des annotations

Malgré ses avantages, la création d'une annotation sémantique est un processus difficile et coûteux pour les biologistes. Ceci nous a amené à développer le système *MeatAnnot* qui à partir d'un texte (articles fournis par les biologistes) permet la génération d'une annotation structurée, basée sur *MeatOnto* et qui décrit le contenu sémantique de ce texte.

MeatAnnot repose sur des outils de TALN (Traitement Automatique de la Langue Naturelle) : GATE (Cunningham et al, 2002), TreeTagger (Helmut, 1994) et RASP (Briscoe et al, 2002) et sur nos propres extensions dédiées à la détection des relations sémantiques et l'extraction des instances des concepts de UMLS.

Dans chaque phrase où il détecte une instance d'une relation sémantique de UMLS, *MeatAnnot* essaie d'extraire les instances des concepts liés par cette relation et génère une annotation décrivant cette interaction.

La méthode de génération est composée de trois phases :

Phase 1 : Détection des relations

Dans cette phase nous avons utilisé JAPE (Cunningham et al, 2002), un langage basé sur les expressions régulières et qui offre la possibilité de créer des grammaires permettant l'extraction d'informations d'un texte traité par GATE. Pour chaque relation de UMLS (interacts_with, plays_role…), nous avons créé manuellement une grammaire permettant d'extraire toutes ses occurrences dans le texte. Nous nous sommes basés sur l'analyse de telles occurrences dans un corpus initial d'articles scientifiques fournis par les biologistes.

L'exemple ci-dessous montre une grammaire permettant la détection des instances de la relation sémantique « Has an effect » dans toutes ces formes lexicales (has an effect, had effects, have a positive effect…).

```
{Token.lemme == "have"}  | {SpaceToken} ({Token.string == "a"}| {Token.string == "an"})?
({SpaceToken})? ({Token.string == "additive"}   | ({Token.string == "synergistic"}|
{Token.string == "inhibitory"}| {Token.string == "greater"}  | {Token.string == "functional"} |
{Token.string == "protective"}| {Token.string == "monogenic"}| {Token.string == "positive"})?
({SpaceToken})? {Token.lemme == "effect"}
```

FIG. 2 - *Exemple de grammaire d'extraction de la relation "has an effect"*

Phase 2 : Extraction des termes

Pour extraire les termes, MeatAnnot utilise le module Tokeniser de GATE et l'outil TreeTagger. Le Tokeniser permet de découper le texte en unités élémentaires appelées tokens en distinguant les nombres, les ponctuations et les mots, et le TreeTagger permet d'affecter à chaque mot une catégorie grammaticale (nom, verbe…).

Après ces deux phases, vient l'extraction des termes candidats. MeatAnnot utilise une fenêtre de taille quatre (quatre mots successifs peuvent représenter un terme) et pour chaque terme candidat: s'il appartient à UMLS, il passe au mot suivant, sinon, la fenêtre d'extraction est diminuée jusqu'à ce qu'elle devienne nulle.

L'interrogation de UMLS se fait à travers le serveur de connaissances UMLSKS qui offre l'accès à toutes les ressources de UMLS et qui permet de les interroger et d'y naviguer à distance. UMLSKS nous renvoie une réponse en XML (si le terme existe dans le métathesaurus). Ce résultat est analysé par MeatAnnot qui en extrait toutes les informations nécessaires sur le terme (type sémantique, définition…). L'utilisation de ce serveur nous a offert une analyse linguistique plus fine car ce dernier traite quelques variations linguistiques simples (« development of lung » est reconnu comme « lung development ») et qui complètent le traitement fait par MeatAnnot à savoir la lemmatisation (récupérer la racine des mots).

Phase 3 : Génération de l'annotation

Dans cette phase, MeatAnnot utilise le module RASP, qui affecte à chaque mot son rôle linguistique dans la phrase (sujet, objet…), ce qui permet d'identifier les concepts liés par la relation : Pour chaque relation détectée, MeatAnnot vérifie si les sujets et les objets sont des instances de concepts de UMLS et génère une annotation décrivant l'instance de la relation.

L'exemple ci-dessous résume les différentes étapes. Considérons la phrase suivante :

"In vitro assays demonstrated that only p38alpha and p38beta are inhibited by csaids."

Étape 1: En appliquant les grammaires d'extraction de relations sur cette phrase, MeatAnnot détecte la présence de la relation « inhibits » (appartenant à l'ontologie UMLS).

Étape 2 : MeatAnnot exécute le processus d'extraction de termes UMLS sur cette phrase. Le résultat est décrit dans le tableau suivant :

Terme	Type sémantique	Synonymes
in vitro	Qualitative Concept	N/C
P38alpha	Gene or Genome	MAPK14 gene, CSBP1…
P38beta	Gene or Genome	MAPK11 gene, SAPK2…
Csaids	Pharmacologic Substance	Cytokine-Suppressant Anti-Inflammatory Drug

TAB. 1 - *Termes extraits par MeatAnnot*

Étape 3 : l'application de RASP sur cette phrase donne comme résultat:

(|ncsubj| |demonstrate+ed:4_VVN| |assay+s:3_NN2| _)
(|clausal| |demonstrate+ed:4_VVN| |inhibit+ed:11_VVN|)
(|ncsubj| |**inhibit+ed**:11_VVN| |**p38alpaha**:7_NN2| |**obj**|)
(|ncsubj| |**inhibit+ed**:11_VVN| |**p38beta**:9_NN2| |**obj**|)
(|arg_mod| |by:12_II| |**inhibit+ed**:11_VVN| |**csaid.+s**:13_NN2| |**subj**|)

FIG. 3 - *Résultat de RASP*

p38alpha et *p38beta* sont détectés comme étant les objets de la relation *inhibits. csaids* est détecté comme étant le sujet.

MeatAnnot génère ensuite une annotation RDF[5] pour chacune des instances.

```
<m: Pharmacologic_Substance rdf:about='csaids#'>
      <m:inhibits>
            <m:Gene_or_Genome rdf:about='p38alpha#'/>
      </m:inhibits >
      <m:inhibits >
            <m: Gene_or_Genome rdf:about='p38beta#'/>
      </m:inhibits >
</m:Pharmacologic_Substance>
```

FIG. 4 - *Exemple d'annotation RDF générée par MeatAnnot*

Toutes les annotations décrivant les interactions dans un article sont stockées dans un répertoire contenant les articles fournis par le biologiste. Ces annotations sont ensuite utilisées, soit pour faire de la recherche documentaire (retrouver un article parlant d'un gène particulier ou d'un phénomène biologique), soit dans un scénario plus complexe de recherche d'informations, comme la recherche de relations entre gènes ou autres entités biomédicales.

3.2.2 Validation des annotations

Afin de valider nos annotations, nous avons adopté une approche centrée utilisateur : nous avons choisi au hasard un corpus de test (2540 phrases) parmi les documents fournis par les biologistes et nous avons présenté les suggestions proposées par *MeatAnnot* aux biologistes à travers une interface de validation pour qu'ils évaluent leur qualité. Cette interface a été conçue de manière à présenter les annotations dans un format compréhensible (textuel) pour les biologistes, qui ne sont pas spécialistes de RDF.

Etant dans un contexte de recherche d'informations (RI), nous nous sommes basés sur des mesures classiques de RI et nous les avons adaptés à notre cas.

Au cours de cette phase, nous avons remarqué aussi que quelques suggestions de MeatAnnot sont considérées correctes mais inutiles pour les biologistes car elles décrivent soit des connaissances de base soit des connaissances vagues. Nous avons donc introduit une nouvelle mesure de qualité nommée *utilité* pour mesurer le taux des suggestions utiles.

	Mesures
Précision	$\frac{\text{Nb suggestions correctes}}{\text{Nb toutes les suggestions extraites}}$
Rappel	$\frac{\text{Nb suggestions correctes}}{\text{Nb suggestions qui auraient dû être extraites}}$
Utilité	$\frac{\text{Nb suggestions utiles}}{\text{Nb suggestions correctes}}$

TAB. 2 - *Mesures de la qualité des annotations*

[5] http://www.w3.org/RDF/

	Suggestions	Correctes	Manquantes	Utiles	Précision	rappel	Utilité
Résultat	454	372	224	357	0,82	0,62	0,96

TAB. 3 **-** *Qualité des suggestions proposées par MeatAnnot*

La seconde colonne du tableau 3, décrit le nombre de relations extraites correctement à partir des textes. La différence avec le nombre de suggestions proposées par *MeatAnnot* est due principalement aux erreurs générées par les outils de TALN (catégorie grammaticale ou rôle linguistique incorrect) et aux termes manquants du thésaurus de UMLS. Néanmoins, nous avons obtenu une bonne précision puisque 82% des suggestions sont correctes.

La troisième colonne décris le nombre de relations existant dans le texte mais que *MeatAnnot* n'a pas pu extraire. Ce silence est dû dans certains cas aux erreurs générées par les outils de TALN mais principalement aux relations déduites par les biologistes en lisant la phrase mais qui ne peuvent pas être générées automatiquement.

Example de phrase: "*Upon interferon-gamma induction, after viral infection for example, a regulator of the proteasome, PA28 plays a role in antigen processing*."

Dans cet exemple, *MeatAnnot* extrait automatiquement la relation "PA28 plays_role antigen processing", mais le biologiste en lisant la phrase peut déduire, en utilisant ses connaissances implicites, une autre relation qui est "interferon-gamma have_effect PA28".

Enfin, *MeatAnnot* a une bonne *utilité* puisque 96% des suggestions correctes sont considérées utiles par les biologistes. Ces résultats montrent bien que *MeatAnnot* génère des annotations de bonne qualité, condition essentielle dans un contexte de recherche d'informations.

3.3 L'utilisation des annotations : MeatSearch

Le but étant d'utiliser les annotations générées par *MeatAnnot* ainsi que celles éditées par *MeatEditor* afin de faciliter la validation/l'interprétation des résultats des expériences, nous avons développé le système *MeatSearch* basé sur le moteur CORESE (Corby et al, 2004) et composé d'interfaces permettant d'interroger et de raisonner sur la base d'annotations.

MeatSearch traduit les résultats de CORESE en présentation graphique et/ou textuelle qui est plus compréhensible par les biologistes. Il fournit aussi des informations complémentaires, telles que le document ou la phrase à partir desquels elle est extraite, les auteurs et les personnes qui ont validé l'annotation ou fourni les articles. Cette richesse en information et cette traçabilité des annotations offrent de très intéressants scénarios d'utilisation.

3.3.1 L'utilisation de CORESE

Pour la formalisation de nos ontologies ainsi que nos annotations, nous avons choisi les langages RDFS et RDF, qui sont deux recommandations du W3C, respectivement pour la représentation des ontologies légères et pour la description des ressources du web en utilisant les annotations basées sur les ontologies.

Ce choix nous a permis d'utiliser CORESE afin de permettre de:

– Naviguer dans la base d'annotations en tenant compte de la structure des ontologies.

- Ajouter des règles qui complètent la base d'annotations.
- Raisonner sur des d'annotations construites à partir de sources différentes et hétérogènes afin de déduire des connaissances à la fois implicites et explicites sur un gène.
- Utiliser différents niveaux d'accès (admin, public, groupe…) à la base d'annotations.

3.3.2 Exemples d'utilisation

CORESE fournit un langage de requêtes pour les données RDF qui est proche du langage SPARQL[6] en cours d'élaboration au W3C; Ce langage de requêtes permet d'écrire des requêtes composées de combinaisons booléennes de triplets RDF.

Comme exemple, la requête suivante permet de retrouver toutes les relations entre le gène « cav3.2 » et les parties d'un corps humain :

```
select ?g ?r ?b where
?g rdf:type m:Gene_or_Genome. ?g = 'cav3.2'. ?g ?r ?b.
?b rdf:type Body_Part__Organ_or_Organ_Component
```

Cette requête interne est générée automatiquement par *MeatSearch* à partir de l'interface graphique de requête manipulée par l'utilisateur et le résultat de cette requete est formaté en une représentation graphique (Figure 5) qui facilite sa visualisation.

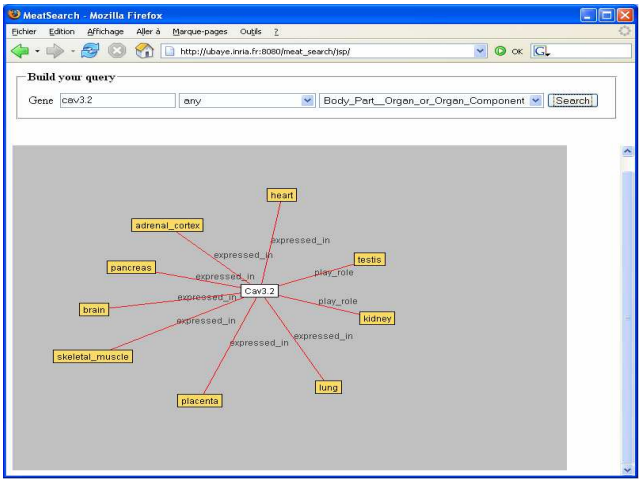

FIG. 5 - *Résultat de la requête précédente renvoyé par MeatSearch*

CORESE fournit aussi un langage de règles qui nous a permis de déduire de nouvelles connaissances à partir de celles qui existaient déjà. Les règles sont appliquées sur la base d'annotations pour compléter et ajouter de nouvelles informations dans le but de réduire le silence dans la phase de RI. Ces règles ont été produites progressivement à travers les discussions avec les biologistes. Nous présentons ici un exemple d'utilisation de règles :

[6] http://www.w3.org/TR/rdf-sparql-query/

"Pour chaque récepteur qui active une fonction moléculaire, si cette fonction joue un rôle dans le fonctionnement de l'organisme alors le récepteur joue le même rôle"

Cette règle est exprimée comme suit:

```
IF  ?r   rdf:type   m:Receptor
    ?r   m:activates   ?mf     ?mf   rdf:type   m:Molecular_Function     ?mf   m:play_role ?of
    ?of   rdf:type   m:Organism_Function
THEN      ?r   m:play_role   ?of
```

Le dernier exemple concerne l'ajout de métadonnées sur les annotations, ce qui permet de rajouter de l'information sur :

- La source de la ressource : le biologiste qui a fourni la ressource à annoter.
- La source de l'annotation : générée par *MeatAnnot* ou ajoutée/validée par un biologiste.
- Le thème général de l'annotation.

Ces informations une fois combinées peuvent nous fournir des annotations contextuelles et multi-points de vue. L'annotation ci-dessous (figure 6) décrit un article fourni par un biologiste nommé Pascal et parlant du développement des poumons.

```
<do:paper  rdf:about='http://www-sop.inria.fr/acacia/meat/lungrepair.pdf'>
    <do:providedBy>Pascal</do:providedBy >
    <do:relatedTo >
       <m: Organ_or_Tissue_Function rdf:about='lung_development #'/>
    </do:relatedTo >
    ....Annotation…
    <do:generatedBy>MeatAnnot</do:generatedBy>
    <do:validatedBy>Pascal</do:validatedBy>
    …Annotation…
</do:paper>
```

FIG. 6 - *Exemple de métadonnée sur une annotation*

MeatSearch utilise ces métadonnées pour proposer différentes vues sur la base d'annotations qui sont liées aux utilisateurs (source d'annotation), au contexte (thème général de l'annotation) et à la méthode de génération de l'annotation (automatique ou manuelle). D'autre part, l'interrogation de ces métadonnées par CORESE nous permet d'avoir plus d'informations sur les annotations et de vérifier leurs cohérences afin de les valider.

4 Conclusion

4.1 Discussion

Dans cet article, nous avons présenté une méthode pour la construction d'une mémoire d'expérience pour le domaine des puces à ADN. Cette méthode peut être généralisée pour n'importe quel domaine scientifique (chimie, physique…) ayant des besoins de validation et d'interprétation de résultats d'expériences. En effet, les modules que nous avons présentés sont réutilisables et reposent sur des technologies standards et libres. Le système MeatAnnot est indépendant des outils de TALN utilisés et peut se baser sur n'importe quelle ontologie.

Notre approche propose des solutions à quelques problèmes posés dans la discussion finale du groupe de travail du W3C dans le domaine des sciences de la vie[7] :

[7] http://www.w3.org/2004/10/swls-workshop-report.html

- Annotations sémantiques de bonne qualité : les annotations générées par *MeatAnnot*.
- La prise en compte du contexte : nos métadonnées sur les annotations.
- La possibilité de raisonnement sur les annotations: l'utilisation de CORESE.

Une autre originalité de ce travail consiste dans (a) l'intégration de métadonnées qui permettent d'avoir d'autres moyens pour plus de raisonnement et d'informations sur la base d'annotations, (b) l'utilisation de plusieurs technologies (TALN, ontologies, annotations sémantiques, CORESE) afin de proposer une réelle application de web sémantique.

Ce travail illustre l'intérêt du web sémantique pour la communauté des biologistes.

Enfin, nous pensons qu'une phase d'évaluation des annotations en amont (cf. §3.2.2) est nécessaire vu que la phase de génération est coûteuse et généralement irréversible.

4.2 Travaux connexes

La méthode sur laquelle *MeatAnnot* repose peut être comparée avec (a) les travaux exploitant l'extraction d'informations dans le domaine biomédical (Alamarguy et al, 2005) (Staab, 2002) (b) ceux sur la génération d'annotations sémantiques pour le web sémantique (Handschuh et al, 2003). Reposant sur des techniques linguistiques, notre approche diffère des méthodes basées sur les techniques d'apprentissage proposées par (Nédellec, 2002). Contrairement à (Golebiowska et al, 2001) notre approche permet de créer des annotations consistantes, non seulement, en des instances de concepts mais en des *instances de relations*, et le tout en reposant sur une ontologie déjà existante. Ces instances de relations peuvent relier les différents concepts de l'ontologie et pas seulement les gènes ou les protéines comme décrit dans (Proux, 2000).

Par rapport aux approches linguistiques pour extraire des relations sémantiques (Séguéla, 1999), nous ne visons pas l'aide à la création ou enrichissement d'une ontologie mais plutôt l'extraction d'informations à partir de textes pour générer des annotations sémantiques sur lesquelles raisonner pour la recherche d'information.

Le couple *MeatAnnot*/*MeatSearch* qui permet de générer des annotations sémantiques basées sur une ontologie et extraites à partir des textes et qui offre un système de recherche sémantique sur ces annotations, a plusieurs points communs avec le système proposé par (Muller et al, 2004) qui repose sur une ontologie mais ne l'utilise pas pour faire la recherche.

4.3 Perspectives

Comme perspective pour ces travaux, nous allons étudier l'extraction d'information à partir des graphiques et les tableaux, compte tenu de leur importance pour les biologistes, de manière à intégrer un nouveau module à *MeatAnnot* pour les prendre en compte. Nous approfondirons aussi les problèmes de la gestion de l'évolution de UMLS, ce qui impliquera d'approfondir l'évolution des annotations. Enfin, nous sommes en train d'étudier avec les biologistes plusieurs scénarios d'utilisation avec des requêtes typiques afin de leur faciliter l'utilisation de notre système et la navigation contextuelle dans la base d'annotations.

Références

Alamarguy L. et al (2005). Annotation de textes par extraction d'informations lexico-syntaxiques et acquisition de schémas conceptuels de causalité. EGC'2005, Paris, France

Briscoe, E. et J. Carroll (2002) Robust accurate statistical annotation of general text. In Proceedings of LREC'02, Las Palmas, Gran Canaria. 1499-1504.

Corby, O., R. Dieng-Kuntz, C. Faron-Zucker (2004), Querying the Semantic Web with the CORESE engine. ECAI'2004, Valencia, August 2004, IOS Press, p.705-709

Cunningham H., D. Maynard, K. Bontcheva et V. Tablan (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02.

Dieng-Kuntz R., Corporate Semantic Webs (2005), in Encyclopaedia of Knowledge Management, D. Schwartz ed., Idea Publishing Group,

Golebiowska J., R. Dieng-Kuntz, O. Corby, et D. Mousseau (2001), Building and Exploiting Ontologies for an Automobile Project Memory. .In Proc. K-CAP'01, Canada.

Handschuh S. , M. Koivunen, R. Dieng et S. Staab, eds (2003), Proc. of KCAP'2003 Workshop on Knowledge Markup and Semantic Annotation, Sanibel, Florida, October 26.

Helmut S. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing.

Humphreys B.L. et D. Lindberg (1993). The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc.;81(2):170–7.

Khelif K., R. Dieng-Kuntz, P. Barbry (2005) - Semantic web technologies for interpreting DNA microarray analyses: the MEAT system., Proc. of WISE'05, 20-22/11 New York

Muller, H.M., E. Kenny et P.W. Sternberg, (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature *PLoS* Biol., 2, E309

Nédellec C. (2002), Bibliographical Information Extraction in Genomics. IEEE Intelligent Systems & their Applications, p.76-80, March/April.

Proux D., F. Rechenmann et L. Julliard (2000). A pragmatic information extraction strategy for gathering data on genetic interactions Proc ISMB

Séguéla P. et N. Aussenac-Gilles (1999), Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. IC'99. pp 79-88. Palaiseau

Staab S., eds. (2002), Mining Information for Functional Genomics. IEEE Intelligent Systems & their Applications, p. 66-80, March-April.

Stoeckert, C.J. et H. Parkinson (2003) The MGED ontology: a framework for describing functional genomics experiments. Comp. Funct. Genomics, 4, 127 132.

Summary

This paper describes MEAT (Memory of Experiments for the Analysis of Transcriptomes), a project aiming at supporting biologists working on DNA microarrays. We provide methodological and software support to build an experiment memory for this domain. Our approach, based on Semantic Web Technologies, relies on formal ontologies and semantic annotations on scientific articles and on other knowledge sources (databases, experiment sheets). It can be extended to other domains requiring experiments and massive data analyses (as proteomics, chemistry…).