# Vers une fouille sémantique des brevets : Application au domaine biomédical

Nizar Ghoula*, Khaled Khelif*, Rose Dieng-Kuntz*

*INRIA Sophia Antipolis 2004 route des Lucioles 06902, BP93
Sophia Antipolis - France
{Nizar.ghoula, Khaled.Khelif, Rose.Dieng}@sophia.inria.fr

**Résumé.** Les brevets sont une source d'information très riche puisque ce sont des documents qui servent à décrire les inventions. L'accès aux documents de brevets en ligne est possible grâce aux efforts des offices nationaux de la propriété intellectuelle. Par ailleurs, ayant des objectifs différents, la présentation de ces documents a pris des formes variées loin d'être unifiées. Ce papier présente une méthode et un système permettant l'analyse de brevets "Patent Mining" pour générer des annotations sémantiques. L'idée principale est de pouvoir prendre en considération la structure des brevets pour pouvoir trouver un lien entre le contenu du brevet et les concepts des différentes ontologies.

# 1 Introduction

## 1.1 Contexte

Le traitement des documents de propriété intellectuelle, tels que les brevets, est important pour l'industrie, les affaires et les communautés juridiques. Récemment, les communautés de recherche académiques et en particulier, les chercheurs de traitement automatique de la langue naturelle et de la recherche documentaire ont reconnu l'importance du traitement des brevets. En fouillant les brevets scientifiques, nous pouvons remarquer un volume important d'informations sur la biologie, les substances et les procédures médicales. En effet, l'extraction des informations de ces brevets permet de donner une idée précise sur : (i) par exemple les interactions biomédicales et l'effet pharmacologique résultant, et (ii) la propriété intellectuelle dans un certain contexte biologique.

Durant ces dernières années, de grands efforts ont été exercés pour mettre les données relatives aux brevets sous une forme électronique et les présenter au public via les services en ligne. De nos jours, nous remarquons que ces services présentent et fournissent des structures de données hétérogènes, ce qui rend difficile à mettre en œuvre une analyse automatique des brevets.

Dans ce papier, nous présentons l'approche **PatAnnot** fondée sur les principes du web sémantique et qui se réfère aux notions de métadonnée et ontologies pour faciliter l'extraction des connaissances et la recherche d'informations relatives aux brevets.

Ce travail rentre dans le cadre du projet européen **Sealife** (Schroeder et al, 2006) qui a pour objectif la réalisation d'un navigateur Web sémantique pour le domaine des sciences de la vie, qui exploitera les ressources du Web en les rendant partageables, accessibles et manipulables par plusieurs utilisateurs dans différents domaines biomédicaux et ce afin de favoriser le partage des connaissances.

## 1.2 Annotations sémantiques sur les brevets

Notre travail vise à faciliter la génération automatique des annotations sémantiques à base d'ontologies sur les brevets accessibles en ligne et repose sur une approche basée sur les principes et les technologies du web sémantique. Ces annotations peuvent être utilisées par les moteurs de recherche sémantiques afin d'extraire les connaissances incluses dans les brevets et les présenter selon le profil de l'utilisateur.

Une annotation est une description permettant d'avoir une information du type métadonnée facilitant l'exploitation, l'accès, la recherche et la reconnaissance d'une ressource. L'annotation peut se baser sur un modèle conceptuel comme par exemple une ontologie afin d'avoir un aspect sémantique lui permettant d'être utilisable, accessible et reconnue par un ensemble d'acteurs ou d'agents. Ainsi une annotation sémantique permet d'établir un lien entre une entité d'une ressource donnée et sa représentation sémantique décrite dans le modèle qui est en général une ontologie relative au domaine où la ressource évolue.

La formalisation du modèle d'annotation se base sur des ontologies ; l'utilisation de la hiérarchie de l'ontologie peut (i) permettre aux annotateurs de choisir le niveau approprié de détail de l'annotation, (ii) diminuer l'ambiguïté de la connaissance et (iii) aider à réduire les erreurs au cours du processus d'annotation. Dans le contexte du web sémantique, l'utilisation des formalismes standards tels que RDF (Lassila et Swick, 2001) ou OWL (McGuinness et al, 2004) pour représenter de telles annotations permet de faciliter leur réutilisation par différents outils d'annotations et moteurs de recherche.

Dans le contexte de l'exploitation des brevets, ces annotations peuvent faciliter :

- La recherche d'informations : puisqu'elles offrent une vue structurelle sur le contenu de brevets et permettent aux moteurs de recherche de tirer profit des modèles de connaissances du domaine (i.e. Ontologies).
- La classification : les ontologies peuvent améliorer la performance des algorithmes de classification des brevets, puisqu'ils offrent un modèle formel ainsi qu'une représentation explicite des connaissances contenues dans les documents. Ces informations peuvent être utiles pour trouver des relations entre les brevets et les classes représentatives du schéma de classification.
- Le raisonnement : liant automatiquement la connaissance incluse dans les brevets en exploitant des inférences possibles sur des annotations sémantiques.

# 2 L'approche PatAnnot

L'idée capitale est de pouvoir prendre en considération la structure des brevets afin de retrouver un lien entre, d'une part, les connaissances contenues dans les documents et, d'autre part, les concepts de l'ontologie utilisée. Nous avons commencé par parcourir la littérature des brevets existante en ligne à travers les sites officiels des offices nationaux et internationaux des brevets et quelques moteurs de recherche spécifiques aux brevets afin de déterminer les caractéristiques de ces documents en termes de structure, contenu et typologie. Ainsi l'annotation sémantique générée portera sur trois aspects principaux : la structure, les métadonnées et le contenu textuel des documents de brevets.

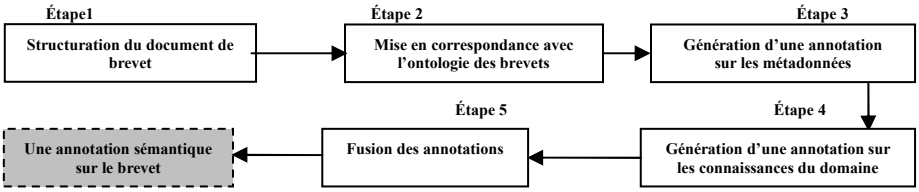


FIG. 1 – Description de l'approche PatAnnot

La figure 1 présente l'approche globale en résumant ses différentes étapes qui se basent essentiellement sur ces quatre points :

1. La construction d'un modèle de représentation sémantique (i.e. une ontologie) des documents de brevets décrivant les aspects structurels ainsi que sémantiques de ces ressources. Nous avons baptisé cette ontologie PatOnto.
2. La structuration des ressources à utiliser dans un format unique en élaborant un moyen de transformation des documents de brevets d'un format d'origine vers un format standard qui permet de faciliter d'autres traitements.
3. La fourniture d'un moyen d'extraction de connaissances à partir des documents de brevets qui se base sur une représentation sémantique de ces ressources et sur des modèles du domaine (i.e. une ontologie de brevets et une de domaine).
4. Le regroupement des différents types de connaissances extraites dans une annotation sémantique.

La figure 2 résume l'architecture du système qui implémente l'approche PatAnnot.

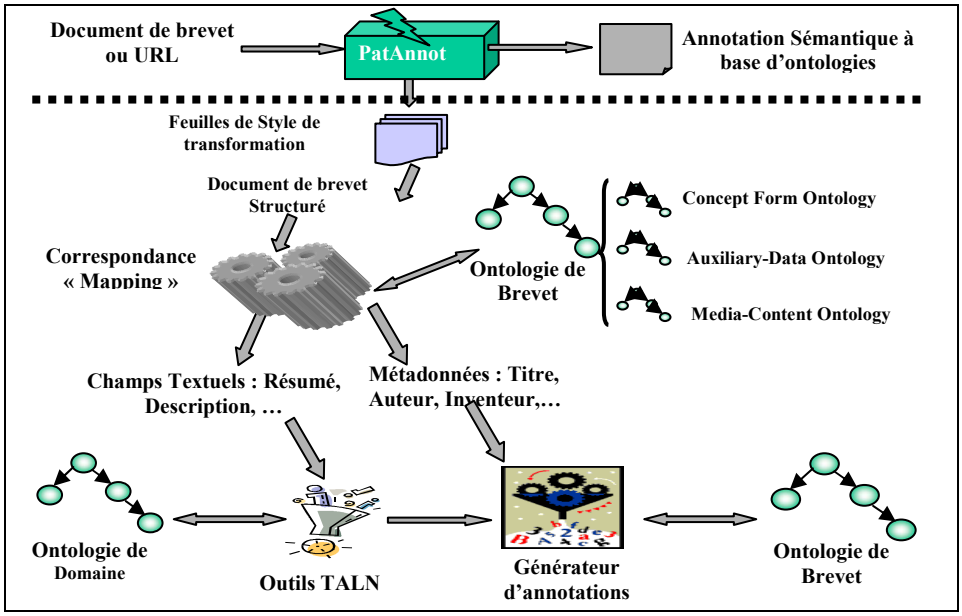


FIG. 2 – Description globale détaillée de l'architecture du système.

Nous pouvons remarquer que l'ontologie de brevet intervient dans différents niveaux de l'architecture du système grâce à sa modularité, ce qui permet de raffiner l'annotation sémantique et de l'enrichir à chaque étape.

# 3 Ontologie de brevet : PatOnto

La connaissance de la structure du document de brevet fournit non seulement une meilleure image sur la morphologie du document mais peut aussi réellement mener le procédé d'analyse. Ainsi la représentation de ces documents est une tâche importante. Jusqu'ici la recherche et le développement dans le domaine de l'analyse des brevets ont été limités à des approches qui ne tiennent pas compte du contenu sémantique des brevets. Ainsi les tâches orientées contenu sont assurées manuellement. De ce fait, une représentation sémantique des documents de brevets ne doit pas porter sur un seul aspect de ces documents. Nous avons donc conçu une ontologie modulaire PatOnto qui modélise une représentation sémantique des brevets. La figure 3 décrit les différentes composantes de cette ontologie.

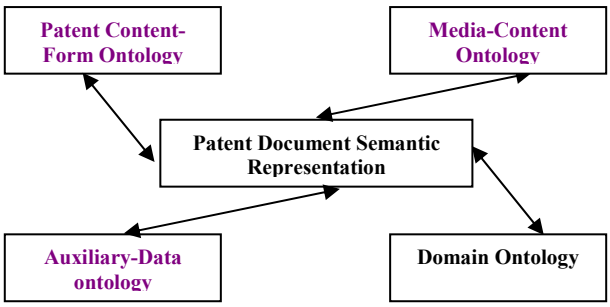


FIG. 3 – Ontologies impliquées dans la représentation sémantique d'un brevet.

1. L'ontologie ''Patent Content Form'' : basée sur notre analyse de la structure de brevet dans les différentes règles et normes de l'organisation mondiale de la propriété intellectuelle (OMPI[1]), cette ontologie décrit la structuration des documents de brevets en termes de sections, sous-sections, etc. fournissant un modèle hiérarchique des éléments de brevet. Une demande internationale de brevet doit être éditée sous forme de brochure. La brochure doit contenir : (i) une page de garde normalisée, (ii) la description de l'invention, (iii) les revendications, (iv) les schémas, et (v) le rapport international de recherche. La figure 4 est une description d'un extrait de cette ontologie que nous avons formalisé en OWL, et qui contient 119 concepts et 95 relations:

[1] http://www.wipo.int/portal/index.html.fr?ex

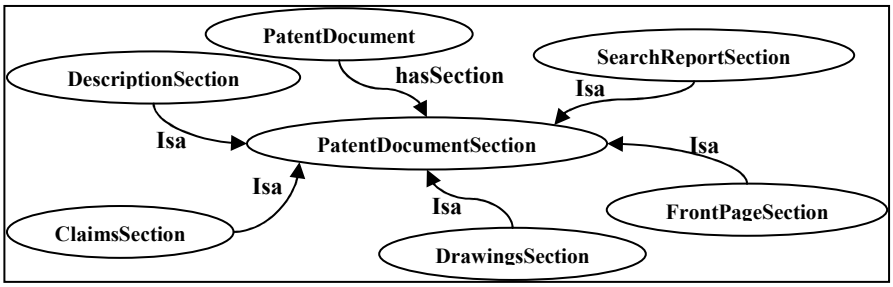


FIG. 4 – Extrait de l'ontologie « Patent Content Form ».

2. Les métadonnées du document de brevet sont modélisées par l'ontologie ''Auxiliary-Data''. C'est une modélisation des informations explicites qui décrivent directement des documents de brevets, et des informations implicites qui exigent un traitement linguistique avancé. Les métadonnées explicites dans ces documents sont utiles pour la description des ressources documentaires puisqu'elles concernent le titre de l'invention, le nom de l'inventeur, la classification du document et toute autre information bibliographique. Ainsi les métadonnées explicites sont bien définies, elles suivent des règles spécifiques. Les métadonnées implicites se composent des concepts qui doivent être extraits en appliquant des niveaux plus élevés d'association entre les documents de brevets avec leur contenu textuel. Ce genre d'informations peut être extrait de la section des références située au niveau de la section de page de garde. La figure 5 décrit un extrait de cette ontologie que nous avons développé en OWL et qui est composée de 109 concepts et 54 relations.

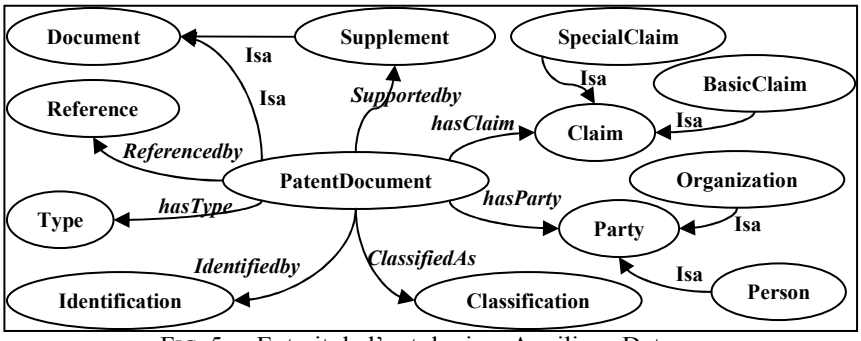


FIG. 5 – Extrait de l'ontologie « Auxiliary-Data ».

3. L'ontologie ''Patent Media-Content'' décrit le contenu non textuel du document de brevet, en effet les illustrations d'une invention font croître la complexité du document car ce ne sont pas des images ordinaires mais divers genres de schémas qui n'ont pas la même interprétation. Le but d'une telle ontologie est de modéliser les objets multimédias qui constituent une partie du document de brevet. Cette ontologie est constituée de 59 concepts et 18 relations, modélisés en OWL. Les concepts que nous avons conçus sont disjoints, par exemple, un objet multimédia ne peut pas être classé en même temps comme étant une formule et une figure. La figure 6 présente un extrait de cette ontologie.

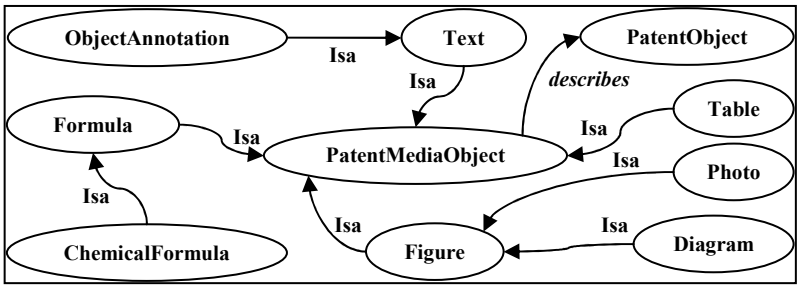


FIG. 6 – Extrait de l'ontologie « Patent Media-Content ».

En conclusion, l'ontologie de brevets PatOnto que nous avons conçue et développée est constituée de trois sous-ontologies décrites précédemment, cette ontologie servira comme référence dans le processus de génération des annotations sémantiques sur les brevets.

L'annotation sémantique est basée aussi sur une ontologie de domaine, de ce fait comme nous traitons les brevets du domaine biomédical nous avons choisi le vocabulaire UMLS (Humphreys et Lindberg, 1993) élaboré par NLM (National Library of Medecine of Bathesda) et constitué (i) d'un métathesaurus qui énumère tout le vocabulaire médical et (ii) d'un réseau sémantique constitué de 134 types sémantiques et de 54 relations pouvant exister entre eux. Ce réseau sémantique peut être exploité comme une ontologie (Khelif et al., 2007).

La section suivante reprend le processus de génération des annotations sémantiques décris par le schéma d'architecture de la figure 2.

# 4 Génération des annotations sémantiques

La génération des annotations sémantiques sur les brevets est un processus qui suit les étapes suivantes que nous allons détailler par la suite :

1. Structurer le document de brevet ;
2. Générer les annotations sémantiques correspondant aux métadonnées utilisant l'ontologie de brevet.
3. Générer les annotations sémantiques sur le contenu en se basant sur l'ontologie de brevet et l'ontologie de domaine et fusionner les annotations.

Nous comptons sur la structure quasi uniforme des brevets fournis par les organismes principaux de la propriété intellectuelle dans le monde qui éditent plus de 80%[2] des brevets accordés dans le monde (OMPI, USPTO[3],…). Nous supposons que prendre en considération la structure et le contenu textuel des documents de brevets va nous permettre de générer des annotations sémantiques plus riches.

## 4.1 Structuration des documents de brevets

Les documents de brevets accessibles en ligne sont en majorité en format HTML. En analysant progressivement cette représentation, nous avons proposé une méthode de transforma-

[2] http://istanbulpatent.com/en/patent.htm, §.2
[3] http://www.uspto.gov/

tion de ce format initial vers le format standard XML. L'utilisation de XML résout le problème en assurant une description structurée du document de brevet, mais il n'y a aucune manière générique simple qui permet de convertir un document HTML vers XML en tenant compte de l'arborescence de ses métadonnées. Ainsi nous avons proposé un processus basé sur les transformations XSLT qui permet de (i) charger le document de brevet en le représentant sous forme d'une arborescence et (ii) le transformer en XML.

*Phase 1 : Construction de l'arbre HTML du document de brevet*

Les balises HTML sont utilisées par les navigateurs pour gérer l'affichage d'une page HTML. En effet ces balises sont conçues pour permettre le repérage des structures logiques ayant une influence sur la présentation physique.

Le format HTML tolère la non fermeture de certaines balises comme *<br>, <hr>, <p>, <li>* etc. ainsi que la fermeture d'une balise avant la fermeture de ses fils, par exemple, *<font>toto</font>*. Par conséquent, la construction d'un arbre de représentation du document est impossible puisque nous avons une ambiguïté dans la classification des nœuds. Comme solution, nous avons appliqué une API existante appelée ''HTMLCleaner'', cette API permet de corriger le code HTML d'un document et d'extraire son arborescence. Un tel arbre se compose d'une racine, des nœuds internes et des feuilles ; un nœud correspond à une étiquette HTML, une feuille peut être un texte ou une étiquette.

*Phase 2 : Génération de la représentation XML*

L'extraction des éléments significatifs à partir d'un document HTML n'est pas exécutée sur le document lui-même mais sur une représentation abstraite mieux adaptée aux manipulations avancées. Nous avons donc choisi la représentation arborescente du document HTML et à l'aide des expressions XPATH, comme par exemple *' //html/body/table[1]/tr',* qui est un chemin unique dans l'arbre HTML., le processeur XSLT peut parcourir cet arbre et extraire les métadonnées et le contenu textuel. Nous avons développé un ensemble de feuilles de style permettant d'assurer la transformation vers XML à travers des procédures récursives permettant de suivre à chaque fois un chemin unique dans l'arbre à la recherche des métadonnées spécifiques pour l'ajouter dans l'arbre XML à générer. La figure 7 montre un exemple d'arbre XML généré d'un document de brevet.

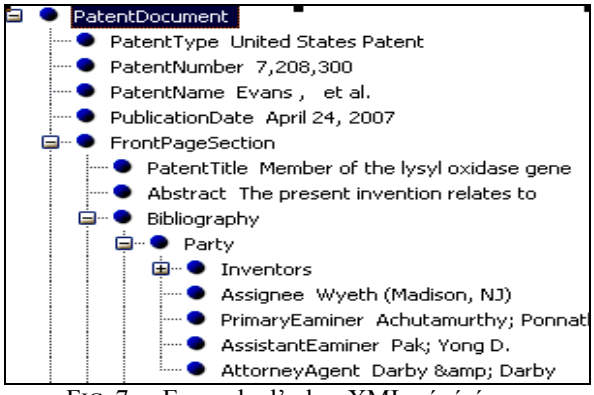


FIG. 7 – Exemple d'arbre XML généré.

## 4.2 Génération de l'annotation des métadonnées

Une annotation de métadonnées décrit les données du document de brevet et pas son contenu textuel, elle décrit des métadonnées en se basant sur les concepts et les relations de l'ontologie ''PatOnto''. La sémantique d'un concept et son emplacement dans le document doivent être pris en compte en se basant sur les différentes parties de l'ontologie. En effet, l'ontologie ''Patent Content-Form'' est employée pour annoter la structure de brevet en classant son contenu et en développant sa hiérarchie, l'ontologie ''Auxiliary-Data'' est employée pour décrire la signification d'un concept et pour annoter son contenu, tel que les références, la classification, les intervenants etc.

Cette étape du processus d'annotation se base sur la correspondance entre le document XML de brevet et l'ontologie. La détection des concepts dans le document n'est pas une tâche triviale ; de ce fait, nous avons proposé un document de correspondance qui va permettre au générateur d'être indépendant de l'ontologie et du document en question. La figure 8 est un extrait de ce document. Nous allons par la suite décrire les deux étapes d'annotation.

*Phase 1 : Annotation basée sur l'ontologie « Patent Content-Form »*

Cette annotation est la principale qui sera raffinée pendant les étapes suivantes du processus de génération des annotations sémantiques. En utilisant un fichier de configuration que nous avons proposé, la correspondance consiste à parcourir le document XML du brevet et l'ontologie et à construire un fichier de correspondance (figure 9).

Si le nœud courant correspond à un concept ***C*** de l'ontologie, alors ce concept ***C*** est ajouté au document de correspondance et une recherche dans l'ontologie est effectuée afin de trouver les propriétés appropriées qui peuvent relier ***C*** à d'autres concepts de l'ontologie :

- Pour chaque propriété ***P*** trouvée, le processus explore le document XML de brevet afin de trouver un autre concept ***C'*** pouvant être lié à ***C*** par cette propriété. Si un tel concept ***C'*** existe, alors la propriété ***P*** et le concept ***C'*** sont ajoutés au document de correspondance;
- Si aucune propriété n'est trouvée, alors le concept ***C*** est considéré comme un nœud isolé;

Le document de correspondance résultant est de la forme suivante :

```
<ResultMapping>
  <Concept>
    <Cname>PatentDocument</Cname>
    <OnProperty>
        <Property>hasPatentType</Property>
        <AppliedOn>PatentType</AppliedOn>
    </OnProperty>
    ……
```

FIG. 8 – Extrait du fichier de correspondance.

Parcourant ce fichier de correspondance et le document de brevet, un processus récursif permet la génération de l'annotation selon les étapes suivantes :

- Construire en mémoire un objet « annotation » de type DOM qu'une méthode spécifique se charge de l'initialiser avec tous les « espaces de noms » requis ;
- Pour un concept ***C*** du document de correspondance :

- Construire une instance de ***C*** dans l'objet « annotation » avec la description formelle de la syntaxe RDF ;
- Pour chaque propriété ***P*** figurant dans la description « OnProperty » du document de correspondance comme nœud fils du ***C***, construire une instance de ***P*** suivant la syntaxe RDF et trouver le concept ***C'*** qui est associé à l'étiquette XML figurant dans le nœud fils « AppliedOn » :
  - si ***C'*** se trouve dans le document de configuration comme concept qui a des propriétés comme nœuds fils, alors le processus refait le même traitement en gardant la référence du nœud XML ;
  - Sinon, instancier le concept et mettre la valeur du nœud XML correspondant comme attribut qui donne la valeur de cette ressource en RDF ;

L'annotation générée est une première version de l'annotation sémantique de brevet qui va être enrichie tout au long du processus d'annotation.

*Phase2 : Annotation basée sur l'ontologie « Auxiliary-Data »*

Des parties textuelles telles que les références, les revendications et la description sont employées pour raffiner l'annotation de métadonnées. Par exemple, au niveau des références, nous classons ces références en nous basant sur les concepts de cette ontologie (références de brevets américains, références étrangères et citations, …) puisque les brevets sont liés et une analyse de plusieurs brevets fournit des résultats plus consistants que ceux obtenus en étudiant chaque brevet à part. Ainsi, l'annotation sur la bibliographie peut être enrichie. Pour cela nous avons conçu une méthode qui permet lors de traitement d'une partie appropriée du document, de générer une annotation sémantique en se fondant sur l'ontologie ''Auxiliary-Data'', ensuite cette annotation est ajoutée à son emplacement dans l'annotation principale.

Grâce à cette ontologie, nous pouvons décomposer les parties textuelles du brevet en plusieurs petites parties significatives ayant chacune un lien avec tout le document pour préparer la phase de traitement automatique de la langue et l'annotation de contenu basée sur l'ontologie de domaine.

```
<AuxDt:hasUri>
- <AuxDt:URI rdf:about="http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=%
   2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN%
   2F4732856">
   <rdf:type rdf:resource="http://www.inria.fr/Edelweiss/2007/AuxiliaryDataOntology#PatentReference" />
  </AuxDt:URI>
</AuxDt:hasUri>
```

FIG. 9 – Extrait de l'annotation basée sur l'ontologie « Auxiliary-Data ».

## 4.3 Génération de l'annotation du contenu basée sur l'ontologie de domaine

Les documents de brevet ont une terminologie spécifique et concrète qui affecte n'importe quel genre de traitement linguistique. La terminologie biologique semble fréquente et importante de ce fait, nous avons utilisé UMLS (Humphreys et Lindberg, 1993) comme ontologie de domaine pour traiter les brevets biomédicaux. Si au cours de l'annotation le processus rencontre une partie textuelle, il la sauvegarde en mémoire et garde une trace du nœud approprié de l'objet « annotation », ce contenu textuel va être annoté par un autre processus. Ainsi nous avons conçu un ensemble de méthodes qui interrogent le module Mea-

tAnnot (Khelif et al, 2007). MeatAnnot fait partie du projet MEAT (Memoire d'Expériences pour l'Analyse du Trascriptome) élaboré au sein de l'équipe edelweiss/Acacia, il est générique et indépendant de toute ontologie ou plateforme et assure la détection des concepts existant dans le texte en se basant sur l'ontologie du domaine en question. MeatAnnot repose sur des outils TALN (Traitement Automatique de la Langue Naturelle) tel que TreeTagger (Helmut, 1994), GATE (Cunningham et al, 2002) ainsi que d'autres extensions propres à edelweiss dédiées à la détection des relations sémantiques et l'extraction des concepts UMLS.

Après avoir collecté toutes ces informations linguistiques, MeatAnnot permet de générer une annotation RDF décrivant le texte proposé en entrée.

Grâce à l'utilisation des ontologies, nous avons pu générer des annotations sémantiques sur les documents de brevets décrivant leur contenu sémantique, ces annotations sont regroupées dans une base d'annotations qui est chargée dans le moteur de recherche sémantique CORESE (Corby et al, 2004). Deux services Web ont été également développés pour encapsuler les différents processus, le premier prend en entrée un document brevet et fournit son annotation sémantique, et le deuxième permet de répondre aux requêtes en interrogeant Corese via SPARQL[4].

# 5 Conclusion

## 5.1 Discussion

L'approche ''PatAnnot'' que nous venons de présenter dans cet article vise à fournir un support méthodologique et technique pour faciliter la fouille des documents de brevets, considérés comme une source inestimable d'informations scientifiques. En effet, le système implémentant l'approche est un système modulaire développé en java et utilise les technologies standards du web sémantique. Notre système est générique et se compose de modules réutilisables ; (i) l'ontologie modulaire ''PatOnto'' que nous avons conçue et construite est indépendante du domaine et couvre toute information structurelle dans un document de brevet, (ii) l'intégration des feuilles de style XSLT qui permettent de générer les documents XML est facile et flexible, (iii) l'outil MeatAnnot est générique et indépendant des ontologies utilisées, (iv) le générateur permet de rassembler toutes les parties des annotations relatives à plusieurs ontologies en une seule annotation sémantique d'un document de brevet qui porte sur plusieurs aspects de cette ressource. Enfin, PatAnnot a été testé et validé sur les deux plus grandes bases des brevets : USPTO et EPO, et ce en annotant automatiquement plus de 1000 brevets.

## 5.2 Travaux connexes

Dans de nombreux domaines l'analyse des brevets est une tâche très importante dont le but est de décrire l'état de l'art des inventions et préserver la propriété intellectuelle. Parmi les outils aidant à effectuer cette tâche nous citons :

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

Le système Vigitext qui repose sur la méthode d'exploration contextuelle pour la fouille des documents techniques (en particulier les brevets) à des fins de veille technologique (Gougon, 2000). Patent Cafe[5], qui aide à mener des recherches professionnelles dans différentes bases de brevets et exploite l'information sur la propriété intellectuelle provenant des différents offices. BioPatentMiner (Mukherjea et al, 2005) facilite la recherche d'information dans les brevets biomédicaux, en identifiant les termes biologiques et les relations entre les brevets dans le but de fournir une approche basée sur le web sémantique. PATExpert (Mark et al, 2006) est un nouveau projet visant à fournir une approche web sémantique et des techniques avancées de traitement de brevets. Notre approche fondée sur les ontologies et les annotations sémantiques est originale par rapport à ces outils. En outre, notre correspondance entre les documents XML et les ontologies qui diffère de (Amann et al, 2001) et (Xiao et al, 2006) par sa généricité et sa manipulation des technologies du Web Sémantique.

## 5.3 Perspectives

Ce travail sur l'exploitation de brevets illustre une application intéressante de Web sémantique, très utile pour la gestion de connaissances des compagnies et des communautés collaborant par le Web. Comme perspective nous pouvons ajouter un module basé sur des outils de TALN qui permet d'extraire à partir de la partie revendications le type du brevet (modèle d'utilité, brevet de conception ou brevet d'usine). L'approche peut être appliquée à d'autres domaines que le domaine biomédical : par exemple les domaines techniques. D'ailleurs, les principes de notre approche (c.-à-d. l'exploitation de la structure des documents et des techniques de TALN sur leur contenu textuel pour produire des annotations sémantiques) pourraient être généralisés à d'autres genres de documents structurés (par exemple fiches patients, fiches d'incident, etc.).

## 5.4 Remerciements

Nous remercions la commission européenne pour le financement de ce travail dans le cadre du projet européen Sealife (IST-2006-027269).

# Références

Amann B., Fundulaki I., Scoll M., Beeri C. et Vercoustre A.M. (2001). Mapping XML Fragments to Community Web Ontologies , WebDB 2001.

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

Corby O., Dieng-Kuntz R. and Faron-Zucker C. (2004): Querying the Semantic Web with the CORESE engine ECAI'2004.valencia Spain.

Gougon B. (2000): Utilisation de l'exploration contextuelle pour l'aide à la veille technologique : Réalisation du système informatique VIGITEXT. Thèse de doctorat de l'université Paris-Sorbonne.

[5] http://www.patentcafe.com/

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

Humphreys B. and Lindberg D. (1993): The UMLS project: making the conceptual connection between users and the information they need. Medical Library Association 81(2).

Khelif, K., Dieng-Kuntz, R., and Barbry, P. (2007): An ontology-based approach to support text mining and information retrieval in the biological domain. In Special Issue on Ontologies and their Applications of the Journal of Universal Computer Science (JUCS), (Accepted for publication).

Lassila O. and Swick R. (2001). W3C Resource Description framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/REC-rdf-syntax/.

Mark G., Achim S., Stören B, Martin R, Thomas E.: Application of semantic Technologies for representing Patent Metadata. AST Workshop Informatik'2006, Dresden.

McGuinness D. L., Van Harmelen F. (2004), OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/.

Mukherjea S., Bamba B. and Kantar P (2005): Information Retrieval and Knowledge Discovery Utilising a Biomedical Patent Semantic Web. IEEE TKDE 2005, p.1099-1110.

Schroeder M., Burger A., Kostkova P., Stevens R., Habermann B. and Dieng-Kuntz R. Sealife (2006): A Semantic Grid Browser for the Life Sciences Applied to the Study of Infectious Diseases. (HealthGrid 2006) 120:167--78.

Xiao L., Zhang L., Huang G., Shi B (2004) Automatic Mapping from XML Documents to Ontologies. CIT'2004, p. 321-325

## Summary

Patents are a rich source of information since they are used to describe the inventions. Due to the efforts of national offices of intellectual property, most of patent literature is accessible by the web. In addition, having different objectives, the presentation of these documents takes varied and not unified forms. In this paper, we describe our approach for generating semantic annotations on patents, relying on the structure and on a semantic representation of patent documents. We use both the structure of the patent documents and their textual content processed using Natural Language Processing (NLP) tools.