Formalisation et Construction d'une Ontologie dans le Domaine des Infections Orthopédiques

Damien De Nizza, James Ortiz, Hubert Meurisse, Pierre-Yves Schobbens

Formalisation et Construction d'une Ontologie dans le Domaine des Infections Orthopédiques

Damien De Nizza[1], James Ortiz[2], and Hubert Meurisse[1], Pierre-Yves Schobbens[2]

[1] Radiology Unit RDGN/CUMG
Catholic University of Louvain
{damien.denizza, hubert.meurisse, jean-paul.trigaux}@uclouvain.be

[2] Computer Science Faculty
University of Namur
{jor, pys}@info.fundp.ac.be

Résumé : Le diagnostic des infections orthopédiques est un processus complexe, long, fastidieux et coûteux. Sa complexité vient du fait que beaucoup d'infections peuvent prendre plusieurs mois pour se développer. Dans le cas d'infections aigües, une réaction rapide est nécessaire pour endiguer le mal. Le diagnostic d'infections orthopédiques implique une approche pluridisciplinaire car elle nécessite les compétences et l'expertise conjointe d'orthopédistes, radiologistes, nucléaristes et biologistes moléculaires. Le travail conjoint de spécialistes issus de domaines médicaux différents a deux conséquences importantes. D'une part, la spécificité et la richesse du vocabulaire propre aux différentes disciplines a tendance à complexifier la sémantique du diagnostique. D'autre part, les données étant de nature hétérogène (par exemple, des régions d'intérêt, des uptakes scintigraphiques, des séquences ADN, des données cliniques, etc), il n'est pas toujours évident d'effectuer informatiquement des recoupements et de déduire automatiquement des informations diagnostiques. Dans cet article, nous proposons la formalisation et la construction d'une ontologie définissant la sémantique de ce domaine à partir de Ressources Terminologiques et Ontologiques (RTO). Cette ontologie est intégrée Ã une plateforme d'aide au diagnostic à laquelle est couplée un moteur d'inférence probabiliste basé sur la logique Markovienne, ce qui permet d'intégrer la notion d'incertitude des raisonnements sur les données.
Mots-clés : Ontologies médicales, Corpus textuel, UMLS, Inférences probabilistes.

1 Introduction

Les infections ostéo-articulaires [Trampuz & Widmer, 2006] constituent un problème de santé publique majeur. Parmi celles-ci, les infections

orthopédiques constituent un groupe important. Il s'agit, généralement, d'infections des prothèses articulaires, survenant après une arthroplastie totale des articulations [Osmon *et al.*, 2013]. En 2004, 1.5 million de prothèses ont été posées dans le monde. En 2011, 20.950 arthroplasties du genou et 28.400 arthroplasties de hanche ont été réalisées en Belgique (chiffre INAMI, en provenance des mutualités socialistes), soit une augmentation de 800% par rapport à 2006 pour l'articulation du genou, et de 200% pour l'articulation de la hanche. Elles sont responsables de complications sévères et représentent un problème économique significatif. La fréquence des infections de prothèses ostéo-articulaires est de 1 à 2% des prothèses (chiffres probablement sous-évalués) et de l'ordre de 5% dans la chirurgie de révision des arthroplasties. Ces patients sans diagnostic formel coûtent cher à la société de par les hospitalisations, les bilans biologiques et d'imagerie répétés visant à préciser la nature de leur problème orthopédique. Quand une infection est diagnostiquée, il reste encore à déterminer à quel germe elle appartient (le Staphylocoque et le Streptocoque sont les plus connus [Trampuz & Widmer, 2006]) pour choisir un traitement approprié. La guérison d'une infection osseuse peut prendre beaucoup de temps. Si, de surcroît, l'infection est couplée à la présence d'une prothèse, alors les choses se compliquent car il faut le plus souvent enlever celle-ci, traiter l'infection et, lorsque la guérison est assurée, remettre une nouvelle prothèse.

Pour traiter ce problème, nous proposons d'intégrer sous la forme d'une plateforme d'aide au diagnostic la combinaison, d'une part, d'une ontologie formalisant le domaine des infections orthopédiques selon l'angle de la démarche diagnostique et, d'autre part, d'un moteur d'inférence probabiliste qui permet l'introduction de la notion d'incertitude (les déductions diagnostiques n'étant jamais entièrement manichéennes).

Dans cet article, nous présentons des objectifs du projet ORTHOGEN[1]. Dans la section 3, nous présentons les travaux d'élaboration et de structuration de l'ontologie ORTHOGEN. Dans la section 4, nous détaillons les choix réalisés concernant la modélisation de l'incertitude ainsi que l'architecture du moteur d'inférence. Enfin, nous concluons et présentons les travaux à venir.

2 Le Projet ORTHOGEN

Ce projet vise à développer une plateforme d'aide au diagnostic des infections de prothèses ostéo-articulaires combinant des modules d'imagerie et de biologie moléculaire, un système d'organisation intelligente

1. Système d'information Intégré pour la Traçabilité et la Gestion Multi−Paramètres des Infections Orthopédiques - projet financé par la Région wallonne DGTRE : WALEO3

des données et des interfaces utilisateur multidisciplinaires et adaptatives, tout en assurant une traçabilité complète du diagnostic. L'objectif de cette structuration de l'information est de rendre possible la réalisation d'inférences automatiques sur les données dans le but d'aider le médecin dans l'élaboration de son diagnostic. Dans cette optique, il convient de modéliser le domaine d'application dans une représentation formelle et non-ambigüe. C'est dans ce cadre que le recours aux ontologies, et plus particulièrement aux ressources terminologiques et ontologiques (RTO), trouve son intérêt.

La construction d'ontologies medicales constitue un défi important, tant pour la communauté de l'ingénierie des connaissances que pour les spécialistes médicaux et les utilisateurs. Etant donnée la complexité des domaines à modéliser, les systèmes de traitement de l'information qui doivent fonctionner dans le monde médical ne peuvent être réellement efficaces que s'ils s'appuient sur des ontologies basées sur des RTO et construites pour un domaine concerné en vue d'une application particulière. La communauté scientifique gravitant autour de l'ingénierie des connaissances travaille depuis plusieurs années sur le problème de la construction de RTO à partir de corpus de textes. Le développement de ces RTO a pour but de faciliter l'usage des terminologies internationales et revêtissent une importance particulière pour le recueil d'information (aide au codage des diagnostics et des actes, réalisation d'études épidémiologiques, etc) et pour l'accès aux connaissances médicales.

Un autre point lié avec le domaine des ontologies médicales est l'incertitude. Le processus d'élaboration du diagnostic médical est une démarche prospective dans laquelle le médecin cherche à maximiser la fiabilité de son interprétation de la pathologie à laquelle est sujet le patient qui le consulte. Ce constat vient notamment du fait que les connaissances du domaine reposent sur une information statistique traduisant cette notion d'incertitude, à l'instar de la phrase "Mycobacterium Smegmatis est résistant au Rifampicine avec une probabilité supérieure à 90%". Ces dernières années, des travaux ont été réalisés dans la représentation et le raisonnement sur l'incertitude dans le Web sémantique [Declerck & Charlet, 2011] et visent la manière de combiner ces langages avec les formalismes probabilistes [Yang & Calmet, 2005][Haase & Völker, 2008]. Pour ce projet, nous avons privilégié une approche basée sur la logique Markovienne [Domingos *et al.*, 2008] et qui combine données et probabilités sans étendre le langage de représentation utilisé (OWL2-DL). L'objectif étant de conserver la transparence du langage et compatibilité avec les moteurs déterministes existant.

Le projet ORTHOGEN comprend quatre modules principaux : (1) **Le Système d'information** qui intègre notamment les problématiques d'accès à l'information (dossier médical informatisé) et d'accès aux médias (issus de laboratoires ou d'un plateau technique d'imagerie).

(2) **Le Kit génétique** vise la création d'un processus d'identification moléculaire directe et rapide des agents infectieux à partir des sites suspectés d'infection. Ce processus d'identification repose sur la Polymerase Chain Reaction (PCR). (3) **La Gestion des connaissances** qui comprend une description formelle des spécificités liées au domaine d'application, tant au niveau médical (pathologies, traitements, etc.) qu'au niveau technique (modalités d'imagerie, dossier et suivi du patient, etc). Ce module inclut également un moteur capable d'inférences déterministes et probabilistes. (4) **Le Diagnostic par l'image** : Ce module concerne la mise en place des mécanismes de détection des différentes lésions observables à travers les clichés provenant des différentes modalités d'imagerie (tant morphologiques que fonctionnelles).

3 L'Ontologie ORTHOGEN

Le développement de l'ontologie été réalisé en collaboration avec les spécialistes médicaux afin d'assurer un niveau de qualité lors de l'inférence. Une collection de termes spécifiques au domaine a été constituée à partir de corpus de textes et de rapports de diagnostic concernant des patients ayant subi des reprises totales ou partielles de prothèse de hanche. L'ontologie ORTHOGEN se base sur une segmentation standard UMLS à partir de ces termes et est écrite à l'aide du langage OWL2. Dans le cadre du processus de construction nous avons utilisé le "Metathesaurus" ainsi que le "Réseau Sémantique" d'UMLS. Pour chaque terme extrait des documents, nous récupérons dans un premier temps le concept du "Metathesaurus" auquel il correspond. A partir de ce dernier, nous récupérons le type sémantique auquel il appartient ainsi que l'ensemble des ancêtres dont il hérite jusqu'à la racine du "Réseau Sémantique". De manière incrémentale, nous construisons l'ontologie par l'ajout successif de sous-arbres de telle façon que, lorsque deux sous-arbres partagent un même ancêtre, ceux-ci sont fusionnés pour ne faire qu'un. Pour finir, nous enrichissons l'ontologie résultante avec les relations existantes parmi l'ensemble des types sémantiques issus de UMLS. La Figure 1 illustre le type de résultat obtenu. Les concepts constituant les feuilles de l'arborescence correspondent aux termes extraits des corpus de textes et de rapports de diagnostics du domaine. Intégrés au sein de cette taxonomie, ces termes sont désormais exprimés selon la nomenclature standardisée du métathesaurus. En remontant récursivement les noeuds parents, on abouti dans les classes du "Réseau Sémantique" de UMLS qui défini une taxonomie de haut niveau permettant une classification générale des concepts médicaux. Ces types sémantiques sont reliés entre eux à travers un ensemble de propriétés permettant d'exprimer de manière riche la sémantique propres aux concepts du domaine.

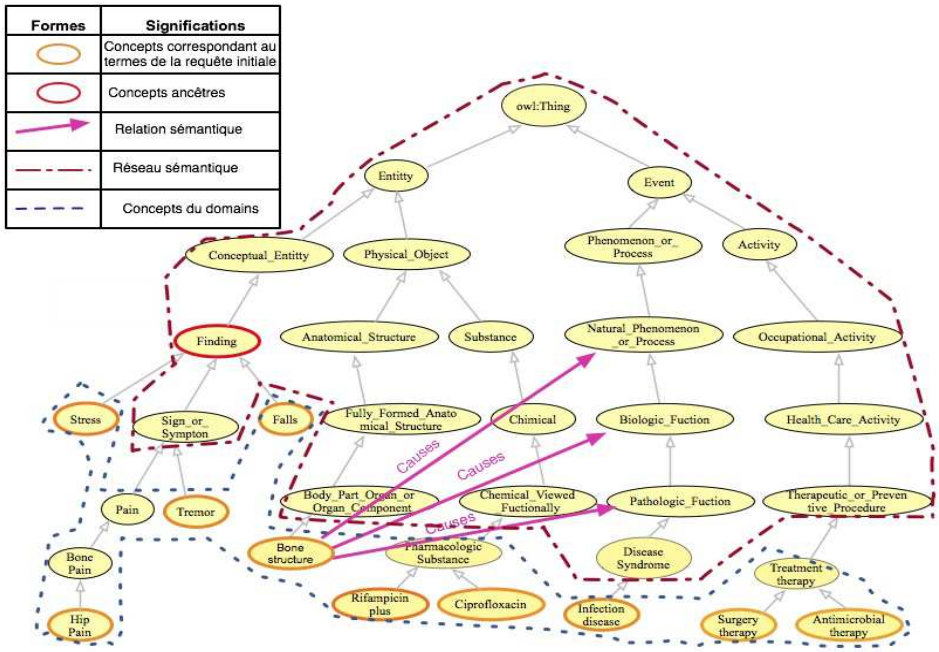

FIGURE 1 – Récupération des Ancêtres à partir des termes et des concepts.

3.1 Limitations

Bien que nécessaire pour la description formelle et précise des concepts médicaux, la richesse d'expressivité de la taxonomie obtenue montre quelques limites lorsque l'on aborde des questions plus opérationnelles. En effet, toutes les relations et classes obtenues ne sont pas indispensables pour les inférences diagnostiques. Un certain nombre d'entre-elles ont une utilité dans l'explication de la sémantique d'un concept mais ont tendance à éloigner dans la taxonomie des concepts qui, pris conjointement, apportent une valeur ajoutée importante aux inférences diagnostiques. Par exemple, les concepts de "liseré radiologique" et de "largeur", pourtant dans des branches éloignées dans la taxonomie, permettent d'interpréter un liseré en tant que descellement de prothèse lorsque sa largeur dépasse un certain seuil. Naturellement, plus il y a de classes et relations, plus le travail du moteur Markovien s'alourdit et l'on souhaiterait optimiser cela. Une autre conséquence est que l'écriture des règles logiques qui vont piloter les inférences devient plus complexe. En effet, celles-ci utilisent ces classes et relations dans leur prémisses et conclusions. Dès lors, en fonction de la définition des concepts dans l'ontologie, la taille et la complexité des règles varie.

Une autre limitation de l'utilisation directe de cette taxonomie est que la classification des concepts n'est pas toujours adaptée au cadre de la démarche diagnostique telle que vue dans le projet ORTHOGEN. Par exemple, la technique de la PCR est considérée comme un examen laboratoire pratiqué en tant que procédure diagnostique pour renseigner sur la présence potentielle de germes infectieux. On s'attendrait donc à retrouver ce concept dans la taxonomie en tant que "Event : : Activity : : Occupationnal_Activity : : Health_Care_Activity : : Diagnostic_Procedure", or elle est classée en tant que "Event : : Activity : : Occupationnal_Activity : : Research_Activity : : Molecular_Biology_Research_Technique". De même, le concept de "Pus" est rangé en tant que "Body_Substance" alors que nous souhaiterions le classer comme information diagnostique car sa présence chez le patient renseigne sur la possibilité d'une infection. Cependant, nous ne souhaitons pas altérer la structure de la taxonomie pour que ces définitions de concepts conservent une compatibilité entière avec le standard UMLS.

Enfin, UMLS étant centré sur la description de concepts médicaux, certaines composantes relatives aux aspects opérationnels de la démarche diagnostique sont moins facilement modélisables, à l'instar des processus d'extraction de données utiles à partir des images, des guidelines diagnostiques ou encore de données plus techniques relatives à la configuration du système informatique au sein de l'hôpital (noeuds DICOM [2] du PACS [3], dossier médical informatisé, etc).

3.2 Contextualisation

Une des attentes formulées dans la description du projet ORTHOGEN concerne l'adaptabilité du modèle d'information avec d'autres domaines médicaux. En considérant cet aspect avec les limitations évoquées au point précédent, l'utilisation directe de la taxonomie obtenue après segmentation de UMLS n'apparait pas indiquée. Plus précisément, sur ce point de l'adaptabilité, il convient d'avoir une structure stable à partir de laquelle on puisse étendre les concepts spécifiques aux domaines et dans laquelle on sache où les chercher. Cependant, les variations de taxonomie obtenues en segmentant UMLS pour différents domaines peuvent être importantes. De plus, de par la classification qui est faite au sein du métathesaurus, certains concepts ne se trouvent pas là où seraient attendus dans le cadre d'une démarche diagnostique (par exemple, les concepts de "PCR" et "Pus" discutés précédemment).

A la lueur de ces éléments, l'idée que nous avons retenu n'est pas de travailler directement sur la segmentation du réseau sémantique d'UMLS

2. Digital Imaging and Communication in Medicine
3. Picture Archiving and Communication System

pour gérer l'élaboration du diagnostic et les réaliser inférences probabilistes. Il s'agit plutôt d'exploiter ce qui constitue la base commune aux différents domaines, à savoir le processus d'élaboration d'un diagnostic médical en tant que tel. Nous avons donc modélisé, sous la forme d'une ontologie, une abstraction de la démarche diagnostique. Dans celle-ci, à partir des classes de haut niveau, on peut étendre les grands composants de cette démarche tels que les entités de l'environnement (acteurs, dossier médical, etc), les entités processus (examens, pipelines d'imagerie, etc), les entités processées (images, post-processing, rapports de laboratoire, etc) ou les informations diagnostiques (symptômes, facteurs de risques, observations, conclusions). Ainsi, au même titre que la taxonomie obtenue par segmentation d'UMLS constitue le réseau sémantique qui défini les concepts médicaux du domaine, cette seconde ontologie leur donne le contexte dans lequel ils vont être exploités.

Cette approche permet de conserver la richesse du réseau sémantique d'UMLS (sans l'altérer) et de bénéficier d'une structure ontologique mieux adaptée à nos besoins. Par adaptée, nous signifions une structure : (1) Dans laquelle la proximité sémantique entre les concepts est plus forte. (2) Dans laquelle nous retrouvons les éléments d'informations qui sont pertinents pour le projet ORTHOGEN et que l'on ne retrouve pas dans UMLS (comme par exemple la définition de services d'extraction des données qui associe, par exemple, le résultat d'un pipeline de segmentation d'image à une instance du concept de liseré ainsi qu'à ses propriétés). (3) Qui nous permet de rédiger des règles diagnostiques de complexité raisonnable en vue d'optimiser le travail du moteur d'inférence. (4) Qui facilite l'adaptabilité à d'autres domaines médicaux.

3.3 Ontologies Modulaires : Trois Niveaux

Toutes ces nouvelles descriptions servant à remplir les contraintes d'adaptabilité et de robustesse du modèle ont amené à pousser plus loin la réflexion concernant la répartition des données au sein du modèle. En effet, jusqu'à présent, deux niveaux ont été envisagés : un réseau sémantique pour la définition des concepts et une description du contexte d'aide au diagnostic. Cependant, la satisfaction de ces contraintes entraine avec elle l'incorporation de nouvelles données servant à décrire l'environnement diagnostique. Parmi ces informations, certaines même ne sont pas réellement pertinentes pour les inférences diagnostiques (mais permettent de trouver, dans l'environnement, les outils conceptuels utiles ou une description précise des concepts). Sachant la lourdeur du travail du moteur d'inférence, on ne peut envisager de le surcharger davantage. Ce constat nous a amené à pousser la réflexion plus loin sur l'intégration, la répartition et la formalisation des données en les balançant sur trois ontologies. Parmi celles-ci, on retrouve le réseau sémantique pour la définition

des concepts propres au domaine. Les deux autres ontologies vont servir à définir, d'une part, l'environnement d'aide au diagnostic (services, définition des types de descripteurs pour les caractéristiques, typages des données, stratégies d'extractions, types d'examens, etc.) et, d'autre part, les diagnostics individuels (sur lesquels se feront les inférences probabilistes). Toutes deux s'appuient sur l'abstraction de la démarche diagnostique évoquée au point précédent. Cette complémentarité entre ontologies permet de récupérer, pour un concept donné (par exemple, un liseré radiologique), d'une part, les informations relative à sa représentation et son exploitation (par exemple, tel algorithme de traitement d'image est capable d'extraire un liseré radiologique et de calculer sa largeur codée en "float") et, d'autre part, les instanciations du concept à travers les différents diagnostics qui sont pratiqués par les médecins (par exemple, Mr Dupont a "ce" liseré de 1.5mm en bordure de prothèse). Cette approche s'inscrit dans la continuité, d'une part, du travail d'optimisation de la structure de l'ontologie sur laquelle le moteur probabiliste va réaliser les inférences et, d'autre part, du renfort de l'adaptabilité à d'autres domaines médicaux.

4 Raisonnement et Incertitude

L'approche probabiliste est un domaine qui essaie de trouver des mécanismes efficaces pour modéliser le raisonnement, tenant compte de l'incertitude de certaines connaissances. Dans ce domaine, les modèles graphiques probabilistes permettent de fournir un outil compact et expressif pour modéliser l'incertitude et la complexité. Ceux-ci joignent dans la même représentation la théorie des probabilités et la théorie des graphes. Il existe deux principaux types de modèles probabilistes graphiques : ceux dirigés (Réseaux Bayésiens [Tighe & Tawfik, 2008]) et ceux non-orientés (Réseaux Markoviens Domingos *et al.* [2008]). Nous avons opté pour la seconde approche car : Les réseaux Markoviens, à la différence des réseaux Bayesiens, permettent la gestion des cycles dans les graphes. Une relation de réciprocité entre deux nœuds est dès lors possible.

4.1 Architecture des moteurs d'inférence

L'architecture du moteur d'inférence probabiliste pour l'ontologie de domaine est construite à partir des modules suivants : "Module Interface ORTHOGEN", "Module Processeur Ontologies", "Module Système Principal" ("Module Moteur Markovien" (Alchemy/Tuffy [4]), "Modules de Traduction". La Figure 2 illustre l'architecture du moteur d'inférence. L'architecture est constitué d'un système principal avec lequel interagissent

4. Markov Logic Network inference engine http ://hazy.cs.wisc.edu/hazy/tuffy/doc/

deux composants externes. Premièrement, le Module Processeur d'Ontologie responsable de la traduction des ontologies OWL2 en logique du premier ordre dans une représentation efficace pour les calculs probabilistes (en forme normale conjonctive). Deuxièmement, le Module Interface avec la plateforme ORTHOGEN. Troisièmement, le Module Système Principal est constitué de deux composants : Module Moteur Markovien, responsable du raisonnement et du processus d'apprentissage. Le processus d'apprentissage de poids est responsable de l'application des techniques d'apprentissage probabiliste (apprentissage automatique ou itératif). Dans le processus de raisonnement, le système utilisera deux informations indispensables : un ensemble de formules pondérées et une requête. Les résultats du raisonnement sont retournés par le moteur Markovien à la couche d'interface et renseigne deux informations : les possibilités d'instanciation des concepts présents dans la requête qui a été adressée au moteur et, pour chacune, une information de probabilité. Concrètement, le moteur est capable de renseigner sur la probabilité d'existence d'une classe de pathologie sur base des informations présentes dans l'ontologie.

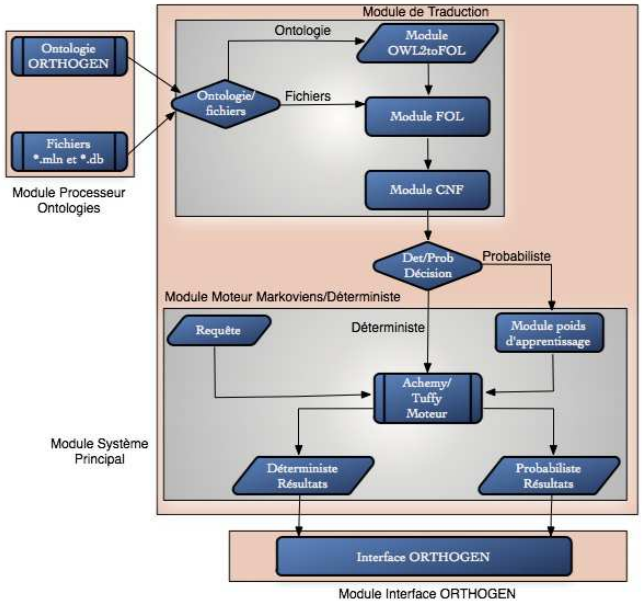

FIGURE 2 – Architecture du moteur d'inférences

5 Conclusion

Dans cet article, nous proposons la formalisation et construction d'une ontologie du domaine ostéo-articulaire et l'implémentation d'une moteur

d'inférence probabiliste. Cette ontologie permettant d'aborder l'integration multidisciplinaire du diagnostic des infections orthopédiques qui fait appel à des concepts hétérogènes puisés dans l'anamnèse du patient, dans l'imagerie et dans l'analyse moléculaire. Pour extraire les données de l'environnement (imagerie, dossier médical, résultats de laboratoire), des algorithmes de traitements d'images ont été mis en place et, en ce qui concerne la biologie moléculaire, nous utilisons la PCR pour extraire des informations génétiques. Ces données sont ensuite caractérisées et annotées sémantiquement au sein d'un modèle ontologique en trois niveaux. Ce dernier s'appuie sur le méta-thesaurus UMLS afin de rester conforme à une nomenclature standard et pouvoir s'adapter plus facilement à d'autres domaines médicaux. Pour assurer les raisonnements logiques aidant à la détermination du diagnostic, nous utilisons un moteur d'inférence. Afin d'intégrer la notion d'incertitude dans ces inférences, le fonctionnement est basé sur la logique Markovienne, ce qui permet de réaliser à la fois des raisonnements déterministes ainsi que probabilistes.

Nous sommes actuellement en phase de développement d'un prototype. L'objectif est de déployer la plateforme au sein de l'hôpital Universitaire de Mont-Godinne. Un groupe de radiologues, nucléaristes et orthopédistes du site se chargera d'utiliser l'outil et de formuler une critique objective sur les résultats obtenus dans le but d'établir une validation de l'approche et de la solution globale.

Références

DECLERCK G. & CHARLET J. (2011). Intelligence artificielle, ontologies et connaissances en médecine les limites de la mécanisation de la pensée. *Revue d'Intelligence Artificielle*, p. 445–472.

DOMINGOS P., KOK S., LOWD D., POON H., RICHARDSON M. & SINGLA P. (2008). Markov logic. *Probabilistic Inductive Logic Programming*, p. 92–117.

HAASE P. & VÖLKER J. (2008). Uncertainty reasoning for the semantic web i. p. 366–384. Berlin, Heidelberg : Springer-Verlag.

OSMON D. R., BERBARI E. F., BERENDT A. R., LEW D., ZIMMERLI W., STECKELBERG J. M., RAO N., HANSSEN A. & WILSON W. R. (2013). Diagnosis and management of prosthetic joint infection : clinical practice guidelines by the infectious diseases society of america. *Clin Infect Dis*, p. e1–e25.

TIGHE C. A. & TAWFIK A. Y. (2008). Using causal knowledge to guide retrieval and adaptation in case-based reasoning about dynamic processes. *Int. J. Know.-Based Intell. Eng. Syst.*, p. 271–281.

TRAMPUZ A. & WIDMER A. F. (2006). Infections associated with orthopedic implants. *Curr Opin Infect Dis*, p. 349–56.

YANG Y. & CALMET J. (2005). Ontobayes : An ontology-driven uncertainty model. In *International Conference on Computational Intelligence for Modelling*, CIMCA '05, p. 457–463.