

ACTES DE LA CONFÉRENCE

IC 2005

16^{es} journées francophones d'Ingénierie des Connaissances

1 – 3 JUIN 2005

NICE

Présidente du comité de programme

Marie-Christine Jaulent, INSERM U729, PARIS

Président de la plateforme AFIA :

Fabien Gandon (INRIA Sophia Antipolis, Equipe ACACIA)

Présidente du comité d'organisation : Monique Simonetti (INRIA Sophia Antipolis,

Organisation de Colloques nationaux et internationaux)

Site Web de la plate-forme : <http://www-sop.inria.fr/acacia/afia2005/welcome.html>

Comité de programme IC2005

Présidente : M.-C. Jaulent, SPIM, INSERM U729, Paris

Membres :

- P. Albert, ILOG, Paris
- N. Aussenac-Gilles, IRIT, Toulouse
- B. Bachimont, INA, Paris & UTC, Compiègne
- C. Barry-Gréboval, LARIA, Amiens
- B. Biébow, LIPN, Villetteuse
- J.-F. Boujut, GILCO, Grenoble
- S. Calabretto, laboratoire LIRIS, Lyon 1
- J. Charlet, AP-HP, Paris
- S. Darmoni, CHU Rouen
- S. Despres, université Paris 5
- F. Darses, CNAM, Paris
- R. Dieng, INRIA Sophia-Antipolis
- J.-G. Ganascia, LIP6, Paris
- N. Girard, INRA, Toulouse
- G. Kassel, LaRIA, Amiens
- J.-M. Labat, CRIP5, Paris 5
- Ph. Laublet, LaLICC, Univ Paris 4, Paris
- C. Le Bozec, SPIM, INSERM U729, Paris
- M. Lewkowicz, Tech-CICO, Troyes
- N. Matta, Tech-CICO, Troyes
- A. Mille, LIRIS, Lyon 1
- S. Moisan, INRIA-Sophia, Sophia Antipolis
- J.-C. Moisdon, CGS, Paris
- A. Napoli, INRIA-Lorraine, Nancy
- Y. Prié, LIRIS, Lyon 1
- M. Revenu, GREYC, Caen
- C. Reynaud, LRI, Orsay
- N. Souf, CERIM, Lilles
- P. Tchounikine, LIUM, Le Mans
- R. Teulier, GRID, ENS Cachan
- F. Trichet, LINA, Nantes
- R. Troncy, ISTI, CNR, Pise
- M. Zacklad, Tech-CICO, Troyes

Avant propos

L'ingénierie des connaissances propose des concepts, méthodes et techniques permettant de modéliser, de formaliser, d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large.

L'essor et l'utilisation croissante des Sciences et Technologies de l'Information et de la Communication (STIC) dans des environnements professionnels divers modifient parfois profondément, les conditions de la représentation et de l'échange des informations et des connaissances entre acteurs au sein d'organisations. L'ingénierie des connaissances appréhende les changements induits par l'utilisation des STIC dans la mesure où elle «réfléchit» sur l'instrumentation technique des contenus pour leur exploitation dans un cadre où ils sont mobilisés pour leur signification. La discipline contribue aux nouvelles «technologies de la connaissance» en développant une ingénierie permettant de diversifier et d'exploiter les modes d'inscription de la connaissance, les modalités d'organisation et de diffusion des savoirs et de démultiplier les interactions entre les utilisateurs.

Du fait de son intérêt pour la connaissance en tant qu'objet à construire, à exprimer, à transmettre, à acquérir ou à exploiter, l'ingénierie des connaissances s'associe à de nombreuses disciplines :

- d'une part, dans sa démarche d'ingénierie, l'IC mobilise les concepts et techniques de la représentation des connaissances, les méthodes d'analyse et de conception à objets, le raisonnement à base de cas, l'ingénierie documentaire ou l'ingénierie éducative, la conception de systèmes d'information, etc,
- d'autre part, dans sa démarche de modélisation des connaissances, l'IC doit se rapprocher de disciplines permettant de caractériser et décrire les connaissances d'un domaine et d'évaluer leur mise en œuvre dans les SBC. La sociologie, la gestion ou l'ergonomie peuvent ainsi concourir à une démarche d'explicitation de ce que sont les connaissances dans un contexte humain et organisationnel.

Placée sous l'égide du GRACQ (Groupe de Recherche en Acquisition des Connaissances - <http://www.irit.fr/ACTIVITES/EQ_SMI/GRACQ/indexact.html>), les journées francophones sont le lieu d'échange et de réflexion de chercheurs français pluridisciplinaires sur les problématiques spécifiques de l'ingénierie des connaissances. Les articles sélectionnés cette année témoignent à la fois des objectifs communs porteurs de la discipline et de la diversité des champs d'application qui reflètent les usages potentiels de l'IC dans les organisations.

Les thèmes abordés dans cette édition 2005 se regroupent de façon assez équilibrée autour des problématiques de construction et d'exploitation d'ontologies d'une part et des problématiques d'ingénierie des connaissances au sein d'organisations et d'entreprises d'autre part. Certains articles amènent des propositions méthodologiques pour la construction d'ontologies à partir de corpus textuels ou à partir de la réutilisation de bases de connaissances déjà existantes. En particulier, on observe un nombre croissant de contributions sur le développement d'outils théoriques et pratiques pour l'alignement

automatique d'ontologies existantes, thématique rendue cruciale par le développement du web sémantique. Dans ce domaine mais aussi plus largement dans le domaine du monde documentaire, l'indexation ou l'annotation à partir d'ontologies pour la recherche intelligente d'informations sont également très bien représentés dans cette édition. Par ailleurs, plusieurs articles témoignent de l'ouverture vers d'autres disciplines comme la gestion des connaissances et mémoire d'entreprise, les systèmes d'information, la théorie des organisations, les systèmes de travail coopératif ou l'ingénierie éducative.

En 2005, les 16^{es} journées francophones d'ingénierie des connaissances se sont déroulés dans le cadre de la plateforme AFIA (Association Française pour l'Intelligence Artificielle) ayant pour vocation de rassembler l'ensemble des communautés intéressées par le traitement de l'information nécessitant de l'intelligence, sous toutes ses formes. Cette plate-forme 2005 a été organisée par l'Institut National de Recherche en Informatique et en Automatique (INRIA). Je tiens à remercier ici vivement Fabien Gandon (INRIA Sophia Antipolis, Equipe de Recherche ACACIA) et Monique Simonetti (INRIA Sophia Antipolis, Organisation de Colloques nationaux et internationaux) pour le remarquable travail qu'ils ont réalisé.

Enfin, je remercie Jean Charlet pour son soutien et sa disponibilité ainsi que l'ensemble des membres du comité de programme d'IC2005 qui ont rendu ma tâche de coordination agréable par le sérieux de leur travail et la convivialité de nos échanges.

Marie-Christine Jaulent
INSERM, U729, Paris

TABLE DES MATIÈRES

Conception et exploitation d'une base de métadonnées de traitements géographiques, description des connaissances d'utilisation <i>Yann Abd-el-Kader</i>	1
Apprentissage de relations entre termes MedDRA dans UMLS pour la détection du signal en pharmacovigilance <i>Iulian Alecu, Cédric Bousquet, Marie-Christine Jaulent</i>	13
Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques <i>Florence Amardeilh, Philippe Laublet, Jean-Luc Minel</i>	25
Construction d'ontologies médicales à partir de textes : propositions méthodologiques <i>Audrey Baneyx, Jean Charlet</i>	37
De l'ingénierie des connaissances à la gestion des compétences <i>Giuseppe Berio, Mounira Harzallah</i>	49
Dérivation d'un arbre de décision pour la mise en oeuvre de stratégies thérapeutiques dans le cas des maladies chroniques <i>Jacques Bouaud, Brigitte Séroussi, Jean-Jacques Vieillot</i>	61
Les annotations pour gérer les connaissances du dossier patient <i>Sandra Bringay, Catherine Barry, Jean Charlet</i>	73
Construction d'une ontologie du droit communautaire <i>Sylvie Despres, Sylvie Szulman</i>	85
Trois méthodes d'analyse pour conceptualiser le contenu de différentes sections des monographies des médicaments <i>Catherine Duclos, Jérôme Nobécourt, Alain Venot</i>	97
Construction guidée de graphes de transducteurs pour l'extraction d'évènements spatio-temporellement localisés <i>Manal EL Zant, Liliane Pellegrin, Michel Roux, Hervé Chaudet</i>	109
Aligner des ontologies lourdes : une méthode basée sur les axiomes <i>Frédéric Furst, Francky Trichet</i>	121

Comment ne pas perdre de vue les usage(r)s dans la construction d'une application à base d'ontologies ? Retour d'expérience sur le projet KmP <i>Alain Giboin, Fabien Gandon, Nicolas Gronnier, Cécile Guigard, Olivier Corby</i>	133
Modèles Sémantiques de Composants pour l'Ingénierie des Systèmes d'Information <i>Gwladys Guzélian, Corine Cauvet</i>	145
Indexation de documents AV : Ontologies, patrons de conception et d'utilisation <i>Antoine Isaac, Bruno Bachimont, Philippe Laublet</i>	157
Modélisation de l'activité d'annotation discursive pour la conception d'un collecticiel support à l'herméneutique <i>Gaelle Lortal, Myriam Lewkowicz, Amalia Todirascu-Courtier</i>	169
Un traitement sémantique par ontologie pour l'indexation de documents dans un référentiel métier <i>Wilfried Njomgue Sado, Dominique Fontaine</i>	181
Intégration de multiples ontologies en anatomie pathologique <i>David Ouagne, Christel Daniel-Le Bozec, Eric Zapletal, Maxime Thieu, Marie-Christine Jaulent</i>	193
Une étude approfondie pour le choix des connaissances à capitaliser en amont de la construction d'une mémoire d'entreprise <i>Inès Saad, Camille Rosenthal-Sabroux, Michel Grundstein</i>	205
COWOS : un modèle des situations de travail collectif pour l'apprentissage de l'organisation <i>Neil Taurisson, Pierre Tchounikine</i>	217
Alignement d'ontologies pour OWL-Lite : l'apport d'un classifieur sémantique <i>Raphaël Troncy, Umberto Straccia, Henrik Nottelmann</i>	229
Introduction aux ontologies sémiotiques dans le Web socio sémantique <i>Manuel Zacklad</i>	241
<i>Index des auteurs</i>	253

Conception et exploitation d'une base de métadonnées de traitements géographiques, description des connaissances d'utilisation

Yann Abd-el-Kader

IGN - laboratoire COGIT
2/4 av. Pasteur 94165 Saint-Mandé cedex
yann.abd-el-kader@ign.fr

Résumé :

Cet article présente un modèle de métadonnées de traitements géographiques, conçu pour décrire les programmes informatiques, logiciels et algorithmes utilisés à l'Institut Géographique National. En particulier, nous montrons comment décrire les connaissances d'utilisation. Le besoin de fournir des modes d'emploi adaptés aux différents contextes d'utilisation nous mène à produire des descriptions opérationnalisables, exploitées par l'application qui permet la consultation de la base de métadonnées construite.

Mots-clés : description des traitements géographiques, modèle de métadonnées, adaptation au contexte d'utilisation, gestion des connaissances, ontologie.

1 Introduction

L'information géographique – base de données vecteurs, base de données images, cartes papiers – est construite, analysée, transformée. Pour cela elle fait l'objet de traitements, communément réalisés par des Systèmes d'Information Géographiques (SIG), bientôt par des services Web. Des organismes développent aussi leurs propres programmes informatiques de traitements géographiques répondant à leurs besoins spécifiques. Par exemple à l'Institut Géographique National (IGN) les équipes de production et les laboratoires de recherche conçoivent et implémentent, entre autres, des traitements d'images, de généralisation¹, d'appariement de base de données. Il s'ensuit que les développeurs et utilisateurs de l'IGN ont besoin d'aide pour partager, rechercher et connaître ces traitements.

Le but de notre travail est de fournir cette aide, qui n'existe actuellement que sous forme de documentations éparses aux formats hétérogènes. Dans cet article

¹ il s'agit de simplifier la représentation des cartes lorsque l'échelle croît, sauvegardant ainsi leur lisibilité pour l'œil humain à la perception limitée (laboratoire COGIT, Conception Objet et Généralisation de l'Information Topographique).

nous montrons comment nous avons défini un modèle de métadonnées adapté aux besoins de consultation identifiés, et quels ont été nos sources d'inspiration. En particulier, nous nous intéressons à la façon de décrire les connaissances d'utilisation des traitements. Afin d'adapter les modes d'emploi au contexte propre à chaque utilisateur (préconditions sur le format des données, disponibilité des traitements annexes et de l'environnement informatique, capacité de l'utilisateur à programmer), nous mettons en œuvre des mécanismes d'inférence. En effet, tous les cas de figure ne peuvent être stockés à l'avance dans la base de métadonnées ; il faut donc dériver l'information recherchée des métadonnées présentes. Les règles qui permettent cette dérivation représentent la connaissance des experts, que notre modèle tente de capturer sous une forme opérationnalisable. Ces règles peuvent ainsi être exploitées par l'application que nous avons développée, application qui permet la consultation et la saisie des descriptions de traitements géographiques.

2 Décrire les traitements géographiques

2.1 Les ressources à décrire, les besoins des utilisateurs

Les traitements auxquels nous nous intéressons existent avant tout sous forme informatique (programmes et bibliothèques de fonctions utilisables dans un contexte de développement, logiciels, SIG, plug-in, services Web), mais également sous forme non-implémentée (algorithmes). L'analyse des requêtes typiques exprimées par les utilisateurs (*"Où sont disponibles les programmes d'appariement?"*, *"Est-ce qu'il existe un programme de généralisation adapté aux bâtiments de formes irrégulières?"*, *"Comment fonctionne le programme Accordéon?"*, *"Quels sont les avantages de Lamps2 par rapport à Geoconcept?"*, etc.) montre que pour un traitement donné les besoins d'information portent sur cinq thèmes principaux : les métadonnées qui l'identifient (nom, date, auteur, etc.), "ce qu'il fait", "comment il fonctionne", "comment l'utiliser" et "quelle est son évaluation". L'enquête auprès des utilisateurs a permis de révéler des besoins spécifiques au domaine géographique : par exemple pour comprendre ce que fait un traitement, il est utile de fournir des illustrations graphiques sous forme d'échantillons des données, ainsi qu'une description de l'évolution des propriétés des objets géographiques avant et après traitement. Il est également des besoins qui nécessitent la mise en œuvre d'un raisonnement "expert" ; nous prenons section 3 l'exemple des modes d'emplois à adapter au contexte d'utilisation.

2.2 Le modèle de métadonnées

Nous avons dû créer notre propre modèle de métadonnées car dans le domaine de l'information géographique il existe des normes pour décrire les données, mais peu encore pour décrire les traitements. Le comité technique de l'International Organization for Standardization sur l'information géographique et la géo-

tique² propose bien un modèle, l'ISO 19119, mais l'ensemble des descripteurs que fournit ce dernier est trop sommaire pour nos besoins. Finalement c'est dans le domaine très actuel du Web sémantique que notre état de l'art des descriptions de traitements nous a mené. C'est en effet en nous inspirant d'OWL-S³, langage de description des services Web, que nous avons défini les principales facettes de description des traitements. On retrouve, à travers le choix des cinq facettes de description adoptées pour notre modèle (fig.1), les trois facettes d'un service selon OWL-S : ServiceProfile (ce que le service fait), ServiceModel (comment fonctionne le service) et ServiceGrounding (comment y accéder)(Coalition, 2004). Source d'inspiration, OWL-S ne répondait néanmoins pas pleinement à nos besoins car, premièrement, seule une partie de nos traitements se présentent sous forme de services Web, les autres possédant des spécificités requérant des descripteurs particuliers; deuxièmement, OWL-S est conçu dans un but de planification et d'exécution automatique : les destinataires des descriptions sont des agents logiciels. Or, à l'opposé, si nos descriptions de traitements sont formulées pour être l'objet de raisonnement (cf. §3), elles n'en demeurent pas moins au final destinées à la consultation humaine.

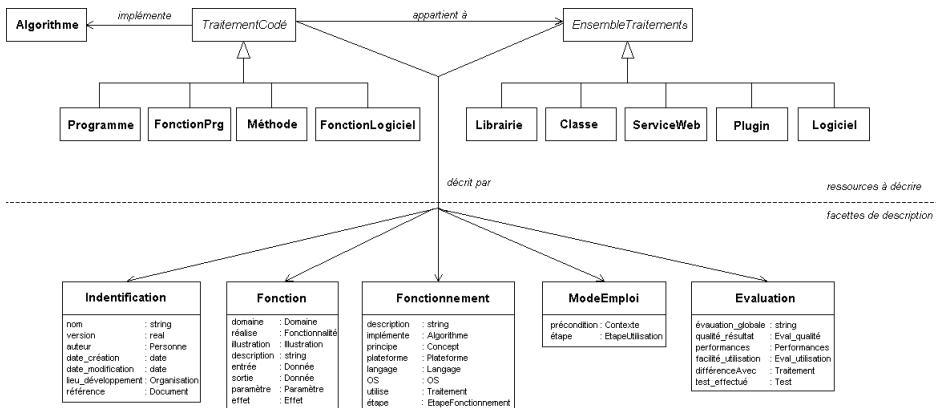


FIG. 1 – Les principales classes du modèle de métadonnées

3 Décrire les connaissances d'utilisation

3.1 Les connaissances à représenter

Délimitation de l'objectif

L'utilisation des traitements géographiques nécessite la mobilisation d'ensembles étendus de connaissances. Des connaissances informatiques ou géographiques,

²<http://www.isotc211.org/>

³Ontology Web Language for Services (Coalition, 2004)

contextuelles ou générales, théoriques ou empiriques, explicites ou tacites⁴... toutes ces connaissances que possède l'expert et qui manquent au novice.

Notre ambition n'est pas de représenter toutes ces connaissances, mais seulement une partie. Celle qui, dans un contexte donné, permettra l'utilisation effective du traitement voulu. Pour circonscrire notre cadre de travail, nous commençons par poser des hypothèses sur l'utilisateur. On postule ainsi que les savoir-faire et concepts informatiques de base sont maîtrisés.

Par ailleurs, il n'est pas envisageable de représenter l'intégralité des connaissances nécessaires au paramétrage de certains traitements complexes, ni forcément de fournir toutes les explications nécessaires à une utilisation experte. Par exemple, une partie de la thèse de S.Bard est consacrée au paramétrage des traitements évaluant la qualité des résultats de traitements de généralisation cartographique (Bard, 2004). Nos descriptions ne peuvent se substituer totalement aux manuels, articles et thèses : en raison de ses propriétés propres (taille non limitée, expressivité), le format traditionnel "texte en langue naturelle" reste irremplaçable pour les modes d'emploi des traitements complexes. Les descriptions de notre base de métadonnées ont donc une vocation complémentaire plutôt que concurrente des documentations en langue naturelle. Elles doivent les référencer.

En résumé nous souhaitons apporter aux utilisateurs une aide certes limitée en complexité mais néanmoins suffisamment complète pour la majorité des besoins communément rencontrés. Une des principales conditions pour atteindre cet objectif sera la capacité du modèle proposé à permettre l'explicitation des connaissances tacites.

Des modes d'emploi adaptés au contexte d'utilisation

Nous allons être amenés à distinguer plusieurs contextes d'utilisation des traitements, selon que l'utilisateur est prêt à programmer ou non, suivant les contraintes liées aux types de données qu'il utilise, suivant encore l'environnement logiciel dont il dispose ou qu'il est contraint d'utiliser pour des raisons de compatibilité avec d'autres traitements. Ces quelques exemples montrent que l'aide apportée à l'utilisateur ne peut être figée. L'adapter automatiquement en fonction du contexte suppose deux choses. Premièrement, les connaissances générales sur l'environnement des traitements doivent être représentées. Il faut fournir un cadre qui permette à l'expert de les exprimer. Deuxièmement, ces connaissances doivent être formalisées de façon à être exploitables par l'application de gestion des métadonnées.

3.2 MASK, une source d'inspiration

Il existe divers modèles de gestion de connaissances. C.Bandza en a fait une synthèse (Bandza, 2000). Nous nous sommes notamment inspirés de ceux propo-

⁴La distinction entre connaissances tacites et connaissances explicites a fait l'objet de plusieurs définitions dans le domaine de la gestion de connaissances (knowledge management) (Ermine, 2003)(Bandza, 2000) . Retenons simplement deux critères : une connaissance est dite explicite si son existence est identifiée, et s'il existe un support quelconque qui en permet la transmission, sinon elle est tacite.

sés par la méthode MASK (Method for Analysing and Structuring Knowledge), dont le but est le recueil et la capitalisation des savoirs tacites d'experts (Bézard & Ariès, 2003). L'ancêtre de MASK est MKSM (Methodology for Knowledge System Management), conçu par J-L.Ermine afin de représenter les connaissances au sein du C.E.A. (Commissariat à l'énergie Atomique). L'élaboration de la méthode a pour origine un constat, celui de la difficulté d'acquérir les connaissances tacites des experts parfois difficilement exprimables. D'où la nécessité de fournir un cadre afin de faciliter ce qui est avant tout un problème d'acquisition des connaissances.

On trouvera dans (Ermine, 2003) la description des six modèles MASK traduisant des points de vue *Connaissances fondamentales, Activités, Contexte historique, Savoir-faire, Concepts et Historique des solutions et leurs justifications*.

Concernant les connaissances heuristiques à représenter pour l'adaptation des traitements (par exemple un expert peut connaître plusieurs façons de contourner un problème d'incompatibilité entre deux programmes, mais ignorer laquelle est préférable), nous nous sommes inspirés des travaux existants dans le domaine de la planification. Par exemple, (Clouard *et al.*, 1998) proposent des descripteurs *déclencheur, précondition, évitement, etc.* pour piloter les applications de traitement d'images.

3.3 Nos choix de modélisation

L'utilisateur exprime son besoin, en réponse il doit se voir indiquer la séquence d'instructions à suivre pour réaliser son besoin. C'est ce but qui a guidé nos choix de modélisation des connaissances d'utilisation des traitements.

Chaque mode d'emploi est composé d'étapes. Une même étape peut se réaliser de différentes façons suivant le contexte. Par exemple, "importer des données dans le SIG ou le programme considéré", se traduira par "appliquer tel traitement de conversion de format", puis "appliquer tel traitement de changement de projection", etc.

Comme dans MASK, plusieurs types de connaissances sont représentés dans notre modèle. Nous en avons introduit quatre. Les concepts (par exemple "distance euclidienne") et les modes d'emploi (spécifiques à un traitement – p.ex. "faire une requête topologique avec Geoconcept" –, ou génériques – p.ex. "interfacer du code Java et du C" –, "améliorer un Modèle Numérique de Terrain"⁵) sont des connaissances destinées à la lecture humaine. C'est-à-dire qu'elles sont manipulables informatiquement (elles font l'objet de requêtes et sont affichées), mais leur signification n'est pas exploitée par l'application ; elles ne sont pas dotées de sémantique opérationnelle. Cela aurait été le cas si, pour reprendre nos exemples, le "concept distance" euclidienne avait été défini dans un langage formel qui aurait effectivement permis le calcul, ou si le mode d'emploi "convertir des données au format shape en MIF" avait été décrit dans un langage comme

⁵Les MNT décrivent la forme et la position de la surface du sol. Ils comportent souvent des défauts, des artefacts qu'il est possible de corriger, par exemple en évitant qu'une rivière remonte ou qu'une crête soit plate.

OWL-S, permettant l’invocation effective du service Web réalisant le changement de format⁶.

Au contraire des concepts et modes d’emploi destinés donc à la lecture humaine, les règles d’adaptation et les heuristiques de choix de ces règles sont dans notre modèle des connaissances opérationnalisables (cf. Tab.1). En effet elles sont utilisées par l’application pour déterminer les réalisations des étapes. Elles devraient également être intelligibles pour l’utilisateur lambda, mais les aspects procédural et déclaratif sont difficiles à concilier. En l’état actuel de nos travaux les règles d’adaptations sont consultables et saisissables sous forme de règles Si contexte Alors Adaptation (lien entre une étape et sa réalisation), mais les heuristiques ne sont accessibles qu’au développeur de l’application de gestion des métadonnées.

Les étapes des modes d’emploi n’étant pas toutes à mettre au même niveau, et ne se suivant pas toujours en séquence, il n’est pas satisfaisant de les présenter sous forme de liste plate. C’est pourquoi l’auteur d’un mode d’emploi est libre d’employer une numérotation hiérarchique arborescente (comme dans un livre on trouve chapitre 1, section 2.1, sous-section 2.2.3, etc.). Le *ou logique* permet d’indiquer les alternatives entre étape de même profondeur. Les structures de contrôles *for*, *while*, *split* (exécution simultanée), etc⁷. utilisées par OWL-S pour l’agencement des services Web ne nous ont pas paru utiles dans le cadre de nos objectifs.

La notation du point précédent permet de contourner la difficulté de saisir en mode texte des graphes de type et/ou habituellement visualisés sous forme graphique. On touche là une différence notable de nature entre les diagrammes MASK ou UML et nos métadonnées avant tout textuelles. Ceci dit, il est toujours possible lors de la saisie d’inclure des images qui seront affichées à titre d’illustration (les références peuvent en effet pointer sur tous type de documents : images, URL, ou plus classiquement manuels dont on a souligné en 2.1 la complémentarité avec nos métadonnées).

Concernant les instructions, deux types ont été distingués, selon que l’utilisateur soit prêt à programmer ou simplement à utiliser un logiciel. Beaucoup de parties du code des programmes sont routinières ; en informatique géographique c’est le cas des instructions qui permettent de faire des requêtes sur les bases de données géographique, charger des données dans les structures de données permettant de les manipuler, d’invoquer les méthodes des objets courants, etc. D’où l’intérêt de décrire des instructions types et les modèles de code (*code template*) qui leur correspondent.

⁶Précisons cependant à propos de la sémantique opérationnelle que les concepts sont définis dans une ontologie et organisés avec les relations de subsumption (distance Euclidienne et distance de Hausdorff sont deux spécialisation du concept de distance qui lui-même est un spécialisation de concept mathématique, etc.). Comme notre application exploite ces relations on peut considérer que les concepts ne sont pas totalement dénués de sémantique opérationnelle – même sommaire.

⁷le détail des descripteurs de *processus* selon OWL-S peut être trouvé dans (Coalition, 2004)

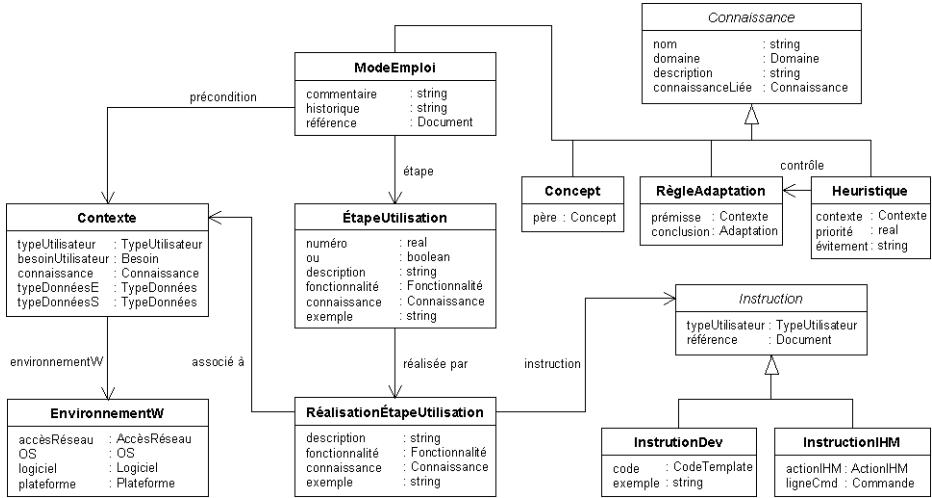


FIG. 2 – Détail du modèle de métadonnées : utilisation des traitements

4 Adapter les modes d'emploi au contexte

4.1 Les règles d'adaptation des modes d'emploi

Les modes d'emploi doivent être adaptés aux contextes d'utilisation. Mais comme il n'est pas envisageable de stocker à l'avance les adaptations de tous les cas de figure possibles, il faut nécessairement mettre en place un mécanisme pour dériver l'information recherchée de celle présente dans la base de métadonnées, le contexte ayant été spécifié par l'utilisateur. Par exemple, si un traitement requiert des données *shape* et que l'utilisateur possède des données *MIF*, le serveur de métadonnées doit lui indiquer comment effectuer une conversion de format à l'aide des traitements adéquats. Le tableau 1 montre en pseudo-langage quelques règles à exprimer. Ces règles constituent une partie des connaissances à faire exprimer par les experts. Dans le cas présent, à la difficulté liée à l'aspect parfois tacite desdites connaissances s'ajoute la contrainte de les acquérir sous une forme opérationnalisable. Notre réponse à ce double problème est la suivante.

Par la présence dans le formulaire de saisie d'un champ *Règle d'adaptation*, l'auteur de description de traitement est incité à songer aux éventuelles connaissances d'adaptations qu'il n'aurait pas spontanément formulées. Les règles d'adaptations déjà existantes, consultables, servent d'exemples suggestifs.

L'expressivité des règles repose grandement, dans la partie *prémissse*, sur la capacité du **Contexte** à représenter la variété des cas de figure possibles. Il faudrait donc également rendre possible la saisie de nouveaux **Besoin** particuliers susceptibles de caractériser le contexte. Ce travail est en cours, comme celui qui vise à opérationnaliser les connaissances heuristiques jusqu'à présent cantonnées à la simple forme de conseils à l'utilisateur.

```
Conversion de format :
```

```
SI données_utilisateur.format = F1
ET traitement.entrées.format = F2
ET F1 != F2
ALORS nouvelle_étape(conversion(F1, F2))
```

```
Interfaçage de langages de programmation :
```

```
SI besoin_utilisateur.langage = L1
ET traitement.langage = L2
ET L1 != L2
ALORS nouvelle_étape(interfacer(L1, L2))
```

```
Indisponibilité logiciel requis :
```

```
SI non_disponible(X)
ET accès.réseau = intranet
ALORS nouvelle_étape(client_Citrix) //connexion machine distante
```

TAB. 1 – Exemple de règles pour l'adaptation des mode d'emploi au contexte

4.2 Quel formalisme pour stocker ces règles ?

Les règles d'aptation au contexte sont des métadonnées de traitements comme les autres. Elles sont donc stockées dans le même format, en l'occurrence XML. Notre modèle de métadonnées présenté plus haut (fig. 1 et 2) est en fait le *modèle conceptuel* (la forme du diagramme de classes étant un impératif du projet de recherche dans lequel s'inscrit notre travail) qui trouve sa traduction dans le *modèle d'implémentation* au format XML Schema. Pour mettre en œuvre un moteur d'inférence plusieurs voies s'offraient à nous. Comme Gandon et Sadeh, générer du CLIPS avec une feuille XSLT pour utiliser le moteur d'inférence JESS était une solution (Gandon & Sadeh, 2004).

Nous en avons choisi une autre : celle d'écrire une feuille XSLT générant elle-même les feuilles XSLT correspondant à la forme opérationnalisable des règles. Une telle feuille produit, par application sur un fichier XML contenant le contexte d'utilisation spécifié par l'utilisateur et sur la base de métadonnées⁸, le mode d'emploi adapté attendu au format HTML, qu'il n'y a plus qu'à afficher.

Une solution à laquelle nous n'avons également pas manqué de songer est d'utiliser les mécanismes d'inférences associés au langage OWL. En effet si parfois les ontologies jouent le rôle de terminologies ou de modèles de connaissances hors de toute exploitation informatique, il arrive aussi – c'est le cas dans le domaine du Web Semantique – qu'elles soient destinées à faire l'objet d'inférences. C'est d'ailleurs parce qu'à cet égard les usages sont variables qu'OWL se décline en

⁸ XPath `document('file')/path` permet d'accéder à plusieurs sources XML au cours d'une même passe XSLT.

3 sous-langages (OWL Lite, DL et Full (W3C, 2004b)) selon que l'on cherche expressivité ou capacités d'inférence.

Mais il appert que les types d'inférences permis de façon générique par OWL ne correspondent pas à nos besoins. Il s'agit principalement de la vérification de consistance (*est-ce qu'une classe peut avoir une instance ?*), de la classification (*A est-elle une sous classe de B ?*) et de la classification d'instance (*à quelle classe appartient un individu ?*) (Knublauch *et al.*, 2004). Ces inférences peuvent être effectuées de façon générique sur les ontologies OWL car elles reposent sur la sémantique formelle des relations unissant les concepts et les individus, sur l'existence de fonctions d'interprétation (Euzenat, 2004). Par exemple l'inférence qui permet de déduire que "*Corse et Ajaccio sont deux lieux apparentés puisque Napoléon est né dans les deux à la fois*" repose sur la fonction d'interprétation du mot clé **cardinality**, affecté en l'occurrence dans l'ontologie considérée de la valeur "1" pour la relation **lieu_de_naissance**.

Contrairement à ce type d'exemple, les règles d'adaptation au contexte que nous voulons mettre en œuvre s'appuient sur des relations spécifiques à notre modèle. Ces règles doivent donc être écrites de façon *ad hoc*.

5 Consulter les métadonnées, en saisir de nouvelles

5.1 Le serveur de métadonnées, l'interface utilisateur

Depuis un an, la base de métadonnées créée conformément au modèle défini est consultable et modifiable via un site intranet de l'IGN. Entre l'utilisateur du site équipé d'un navigateur Web et la base de métadonnées se trouve le serveur d'application qui effectue la reformulation des requêtes fournies, l'adaptation au contexte, la gestion de l'IHM, etc. ; nous sommes dans le cadre d'une architecture 3 tiers classique.

L'interface utilisateur propose les fonctionnalités traditionnelles de systèmes d'aide : recherche plein-texte de mots clés ou navigation dans les index. A cela s'ajoute la possibilité de soumettre via des formulaires HTML des requêtes telles que "*quels sont les traitements dont l'auteur est Léon et la fonctionnalité réalisée caricature ?*". Nous gérons les relations de subsumption grâce à la génération dynamique d'index inversés. Le principe est simple : étant donné un ensemble d'index décrivant les relations unissant des entités, on construit de nouveaux index décrivant les relations inverses. Typiquement, partant de pages Web contenant des listes de mots, on construit les index inverses qui décrivent pour chaque mot les pages Web qui les contiennent. Ainsi, la requête précédente prendra en compte tous les traitements qui réalisent une sous-fonctionnalité de *caricaturer* (*amplification, dilatation, etc.*). Puisque nous avons automatisé l'export des classifications de notre base de métadonnées en OWL (cf. §4.2), un autre moyen que la construction d'index inversés pour retrouver les liens de subsumption aurait été l'utilisation d'une API permettant le requêtage OWL. Nous avons pour l'ins-

FIG. 3 – Illustration d'un cas typique de consultation

tant écarté cette possibilité pour des questions de performance. En effet coupler un moteur OWL au moteur XSLT sur lequel repose le cœur de l'application était possible mais lourd (génération de fichiers intermédiaires, temps de parsing⁹) et nous avons donc pour l'instant choisi de tout faire en XSLT.

Les résultats des requêtes peuvent être triées selon le critère désiré. La présence de liens hypertextes vers les diverses métadonnées de la base est systématique, ce qui donne à la navigation une souplesse semblable à celle des Topic Maps. Topic Maps dont les avantages et l'affichage via le générateur de pages Web Omnigator d'Ontopia¹⁰ nous sont connus, mais dont la philosophie s'opposait aux principes

⁹Lors d'une transformation XSLT le fichier XML source est chargé en mémoire vive sous la forme d'un DOM (Document Object Model) XML. Un DOM XML prend en mémoire vive environ 10 fois la taille du fichier XML source, ce qui pose des problèmes de performance à partir d'un certain seuil. Dans l'éventualité d'une utilisation de requêtes OWL nous devrons voir comment les outils proposés se comportent sur cette question.

¹⁰<http://www.ontopia.net/omnigator/models/index.jsp>

de modélisation que nous souhaitions mettre en œuvre.

5.2 Le contenu de la base de métadonnées

En son état actuel, la base de métadonnées est riche d'une centaine de descriptions de traitements (SIG et packages Java pour la manipulation de primitives géographiques principalement), mais aussi surtout de descriptions de fonctionnalités géographiques (200), de type de données (35 au niveau abstrait, c'est-à-dire non implémenté), de concepts et de modes d'emploi, descriptions dont les ensembles respectifs forment les classifications qui servent de base à l'indexation des traitements – indexation sémantique pourrait-on dire puisque le sens des ressources est fixé est normalisé dans le cadre de travail – mais de façon générale sont utiles pour qui veut avoir un aperçu synthétique du domaine (Fig.4)

Sans préjuger de la valeur de nos classifications, de toutes façons appelées à être enrichies et éventuellement révisées au fil du temps¹¹, les rendre disponibles sous forme d'ontologie OWL pourrait constituer une participation au projet d'interopérabilité des services Web géographiques et plus généralement au partage de la connaissance. Nous-même aurions été heureux d'avoir accès sous forme d'ontologie à la classification des fonctionnalités selon une partie de la communauté géographique (ce que nous n'avons pu trouver que sous forme de parutions écrites).

C'est pourquoi nous avons créé des feuilles XSLT exportant les classifications de notre base de données vers des ontologies en OWL. La figure 4 montre ainsi la hiérarchie des fonctionnalités avec l'éditeur Protégé 3.0 bêta¹².

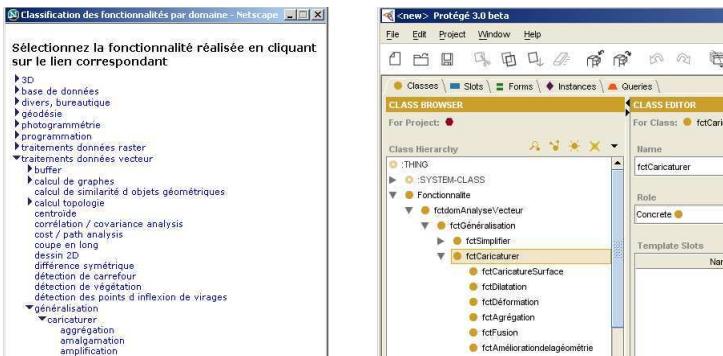


FIG. 4 – Classification des fonctionnalités vu via notre application Web (export HTML) et via Protégé 3.0 bêta (export OWL); l'ordre varie à l'affichage mais les classifications sont identiques.

¹¹Le versionnage des ontologies OWL est prévu ((W3C, 2004a), §6) mais cela ne résout pas tous les problèmes causés par les révisions dans les indexations devenues obsolètes. La description des évolutions et les mises à jours seraient des pistes à creuser.

¹²<http://protege.stanford.edu/index.html>

6 Conclusion

Si, en leur état actuel, le modèle de métadonnées défini et l'application développée permettent bien de répondre à la majorité des requêtes simples que l'analyse des besoins a fait apparaître, il reste néanmoins des cas où l'utilisateur ne peut être renseigné de façon satisfaisante, comparé à l'information que pourrait fournir un expert humain. En rendant opérationnalisable une partie des connaissances d'utilisation des traitements, nous avons fait un pas vers un système d'aide "intelligent". Que ce soit pour l'aide au paramétrage, la modélisation des connaissances générales du domaine géographique, ou l'aide à la programmation, les pistes ne manquent pas pour continuer dans cette voie. Pour de tels besoins, nous retenons de l'expérience de notre travail qu'avant même l'apport de mécanismes d'inférences simulant l'expert, c'est dans le modèle de métadonnées que réside "l'intelligence" dans la mesure où il permet la structuration et la transmission des connaissances. En cela nous nous situons dans la lignée des modèles de gestion de connaissances que nous avons étudiés.

Références

- BANDZA C. (2000). *Des méthodes de formalisation des connaissances et de MKSM en particulier.* thèse professionnelle du mastère MSIT, Management des Systèmes d'Informations et des Technologies, HEC-Mines, <http://www.hec.ensmp.fr/Theses/Theses2000/Bandza.doc>.
- BARD S. (2004). *Méthode dévaluation de la qualité de données géographiques généralisées - Application aux données urbaines.* thèse de doctorat, université Paris 6.
- BÉZARD J.-M. & ARIÈS S. (2003). *La méthode MASK - Présentation pour la capitalisation des connaissances.* <http://perso.wanadoo.fr/serge.aries/presentation/MASKmet/frame.htm>.
- CLOUARD R., ELMOATAZ A. & REVENU M. (1998). *Une modélisation explicite et opérationnelle de la connaissance en Traitement d'Images.* RFIA'98, Clermont-Ferrand.
- COALITION T. O. S. (2004). *OWL-S : Semantic Markup for Web Services.* <http://www.daml.org/services/owl-s/1.0/owl-s.html>.
- ERMINÉ J.-L. (2003). *La Gestion des connaissances.* Hermès Science Publication (diff. Lavoisier), 2003. *Voir aussi :* J.-L. Ermine, Introduction à la méthode MASK, <http://perso.wanadoo.fr/serge.aries/presentation/MASKint/frame.htm>.
- EUZENAT J. (2004). *Chouette alors ! Un langage d'ontologies pour le web.* conférence invitée d'IC'2004, 15èmes journées francophones d'ingénierie des connaissances, Lyon.
- GANDON F. & SADEH N. (2004). *Gestion de connaissances personnelles et contextuelles, et respect de la vie privée.* In actes d'IC'2004, 15èmes journées francophones d'ingénierie des connaissances, Lyon.
- KNUBLAUCH H., MUSEN M. & NOY N. (2004). *Creating Semantic Web (OWL) Ontologies with Protégé.* Stanford Medical Informatics, <http://protege.stanford.edu/plugins/owl/publications/2004-07-06-OWL-Tutorial.ppt>.
- W3C (2004a). *OWL Web Ontology Language Guide.* <http://www.w3.org/TR/owl-guide/>.
- W3C (2004b). *OWL Web Ontology Language Reference.* <http://www.w3.org/TR/owl-ref/>, traduction française de J.J.Solari datant du 4 mai 2004 : <http://www.yoyodesign.org/doc/w3c/owl-ref-20040210/>.

Apprentissage de relations entre termes MedDRA dans UMLS pour la détection du signal en pharmacovigilance

Iulian ALECU¹, Cédric BOUSQUET¹, Marie-Christine JAULENT¹

¹INSERM U729, Paris, F-75006, France
{Iulian.Alecu,Cedric.Bousquet,
Marie-Christine.Jaulent}@spim.jussieu.fr

Résumé : Ce travail est motivé par la problématique de recherche d'informations spécifiques au domaine de la pharmacovigilance. Un signal de pharmacovigilance est la relation détectée statistiquement entre un médicament et un groupe d'effets indésirables exprimant des conditions cliniques similaires. Le regroupement pertinent des termes cliniques normalisés rend un signal plus spécifique et sa détection plus sensible. Notre objectif est de trouver les regroupements pertinents des effets indésirables en employant une méthode d'apprentissage automatique sur l'UMLS - un grand métathésaurus de la médecine. A partir des relations pertinentes regroupant des termes cliniques normalisés et non définies explicitement dans MedDRA, le thesaurus d'origine, nous avons extrait des règles permettant de prédire ces relations à l'intérieur d'UMLS. En appliquant ces règles nous avons réussi à prédire les relations pour deux couples de concepts. Ces résultats démontrent l'intérêt de l'utilisation des méthodes d'apprentissage automatique sur l'UMLS pour l'extraction de relations non définies explicitement dans un système terminologique.

Mots-clés: Terminologies médicales; Appariement terminologique; Réseau sémantique; UMLS; MedDRA.

1 Introduction

Les différents points de vue selon lesquels on peut manipuler de l'information dans un domaine applicatif nécessitent des regroupements différents au sein d'une même organisation de l'information. Dans le cadre du travail présenté, nous nous sommes posés la question du regroupement d'information dans un contexte très particulier, rattaché au domaine médical, celui de la pharmacovigilance. Cette discipline cherche à mettre en évidence des relations statistiques entre des groupes d'effets indésirables (EI) observés et des médicaments, ce que l'on appelle un signal de pharmacovigilance. Une fois découverte, une telle relation peut mener à des études plus étendues afin d'établir la relation de causalité et donc l'imputabilité du médicament pour l'apparition des EI. L'objectif à long terme du projet est d'apporter une réponse à la détection du signal en s'appuyant sur le regroupement des

informations concernant les EI. Dans cet article nous nous intéressons essentiellement à la construction de l'organisation sous-jacente au regroupement des informations.

L'organisation de l'information pour un domaine d'application donné nécessite au moins deux étapes.

Une première étape de *normalisation terminologique* assure qu'une entité du domaine sera décrite au moyen des mêmes termes préférentiels, toutes les fois qu'elle sera indexée. Cette opération donne lieu à un vocabulaire contrôlé.

Ensuite, par un travail consensuel de *normalisation conceptuelle* du domaine (Charlet, 2003), des experts déterminent la position des termes du vocabulaire contrôlé dans une structure relationnelle (souvent hiérarchique). Le thesaurus résultant a l'avantage de regrouper l'information à des niveaux de granularité différents. L'inconvénient majeur d'un thesaurus est son manque de flexibilité. Ainsi, l'intégration d'un terme nouveau (ou l'exclusion d'un terme) dans le thesaurus n'est ni facile ni immédiate. Cette inflexibilité se retrouve aussi dans la structure relationnelle. Généralement, l'utilisation d'un thesaurus pour faire des regroupements d'information basés sur des critères autres que ceux proposés par les auteurs du thesaurus risque fortement de mener à des résultats insatisfaisants.

L'utilisation d'une ontologie formelle pour faire des regroupements d'information résout le problème de l'inflexibilité et s'intègre dans un effort commun pour la mise au point des nouvelles technologies dans l'organisation de l'information.

Dans le contexte spécifique des informations sur les effets indésirables, une première étude développée dans notre laboratoire a montré qu'une ontologie formelle des concepts désignant les EI permet d'effectuer des regroupements de cas pertinents et d'améliorer ainsi la détection de signaux (Henegar 2004). En même temps nous nous sommes retrouvés face à des difficultés importantes pour développer cette ontologie qui reste aujourd'hui très partielle. La méthodologie de construction adoptée, qui s'appuyait essentiellement sur les connaissances d'un seul expert, n'était pas adaptée à l'ampleur de la tâche. Nous souhaitons donc proposer une méthodologie originale qui réutilise des connaissances déjà acquises dans le domaine, comme par exemple les connaissances déjà présentes dans les thesaurus médicaux, et en particulier le métathésaurus UMLS (Unified Medical Language System).

L'hypothèse qui sous-tend notre méthodologie consiste à dire qu'il existe à l'intérieur d'UMLS des connaissances implicites qui permettraient de suggérer automatiquement des relations de subsomption entre concepts lors de la construction d'une ontologie dans le domaine médical.

La méthodologie adoptée et présentée dans cet article s'apparente à une méthode d'apprentissage. Grâce à l'ontologie amorcée précédemment, nous avons un certain nombre de couples de concepts qui sont directement liés par une relation de subsomption. C'est notre base d'exemples. La relation de subsomption a été créée à partir des définitions formelles fournies par l'expert et n'existe pas nécessairement dans UMLS. Nous cherchons donc à apprendre si parmi les différentes façons de parcourir UMLS d'un concept à un autre, il y en a certaines qui peuvent suggérer une relation de subsomption en se basant sur les exemples connus. La méthode présentée explicite dans un premier temps le formalisme des exemples qui vont être utilisés lors de l'apprentissage. Ensuite, la méthode d'apprentissage elle-même est fournie. Elle

s'appuie sur une méthode statistique classique (l'odds ratio). Nous présentons ensuite, une étape de vérification sur deux couples exemples qui permet de conforter l'intérêt de l'approche. Notre contribution est aujourd'hui dépendante de la validation dans notre contexte d'application. Néanmoins, des briques sont posées pour une stratégie d'extraction des connaissances à partir d'UMLS pouvant avoir des retombées dans bien d'autres secteurs médicaux que la pharmacovigilance.

2 Contexte et état de l'art

2.1 Les méthodologies de construction des ontologies

Une ontologie est la spécification explicite et formelle d'une conceptualisation partagée, en vue de la réalisation d'une tâche (Bachimont, 2000). Le processus général de modélisation est complexe, mais il s'agit d'abord d'identifier les concepts utilisés dans le domaine pour la tâche, leurs propriétés et les relations qu'ils entretiennent. Plusieurs approches pour la construction d'ontologies ont été publiées (Aussenac 2000). On distingue en général les méthodes descendantes, orientées sur la tâche à réaliser en fonction de laquelle les connaissances du domaine sont choisies et les méthodes ascendantes qui utilise le sens des mots pour organiser les connaissances. Dans le domaine médical, des expériences sont menées qui privilégient les approches ascendantes en donnant beaucoup de place aux rôles respectifs de l'expert et des textes dans la conception de l'ontologie par le choix des primitives (leMoigno 2002). Ces travaux s'appuient sur le fait qu'il existe des sources textuelles importantes comme les comptes rendus médicaux. Par contre, en pharmacovigilance, il n'existe pas de textes reflétant une pratique de la discipline. Les sources textuelles qui sont à notre disposition sont les thésaurus médicaux.

2.2 Les thésaurus médicaux

Les thesaurus en médecine sont utilisés pour enregistrer et d'échanger des informations et des connaissances médicales sous une forme normalisée (Cimino 1996, Zweigenbaum 1999). Ces thésaurus sont construits toujours dans un but précis. Ainsi, le thesaurus MeSH est employé pour l'indexation des connaissances médicales, en particulier les articles scientifiques (NLM 2001). D'autres ont pour vocation l'établissement de statistiques, par exemple de mortalité et de morbidité avec la classification internationale des maladies CIM-10 (OMS 1993). La nomenclature SNOMED Internationale (Côté 1998) est utilisée pour enregistrer des informations cliniques détaillées. S'ils ne sont pas dédiés à cela, ces thésaurus contiennent des informations relatives aux EI.

MedDRA est le thésaurus actuellement utilisé pour la normalisation des EI en pharmacovigilance. Les études sur la pertinence de regroupements des termes MedDRA désignant des EI ont mis en évidence des problèmes portant spécialement sur la pauvreté relationnelle et le positionnement des termes qui présentent parfois des granularités différentes sur un même niveau hiérarchique. La structure relationnelle

dans laquelle ces termes sont organisés est d'autant plus importante qu'elle a un impact direct sur la spécificité et la sensibilité des signaux de pharmacovigilance détectés (Yokotsuka 2000 ; Brown, 2002). Nous avons montré dans un article récent que MedDRA ne permet pas le regroupement pertinent d'information pour la détection du signal et que ceci est en partie lié au manque de représentation formelle des termes (Bousquet 2005).

Devant ces constats, nous nous sommes plus particulièrement intéressés à l'Unified Medical Language System® (UMLS) de la National Library of Medicine (NLM) qui présente un statut particulier par rapport à ces thésaurus (Lindberg 1993).

UMLS¹

L'objectif déclaré de l'UMLS est de mettre ensemble et de rendre interrelationnels la majorité des standards terminologiques utilisés actuellement par les divers domaines de la santé afin d'obtenir une terminologie complète de la médecine. La construction d'UMLS, ne vise pas un domaine spécifique et reste la plus générale possible, laissant aux utilisateurs la possibilité de l'adapter conformément à leurs besoins. L'effort de la NLM se concentre seulement dans l'intégration des terminologies au sein d'une base de connaissances sans apporter de modifications à leurs composants linguistiques ou structurels.

La couverture terminologique du domaine médical offerte par UMLS est impressionnante (60 familles de terminologies médicales, 2 millions de termes et 900 mille concepts, 12 millions de relations entre les concepts), MedDRA y étant incluse. Les thesaurus inclus dans l'UMLS ont été créées dans des contextes sensiblement différents, ayant comme point commun le domaine de la médecine.

Les spécificités à l'intérieur des différents thesaurus présentent deux aspects :

- un aspect *structuré*. On trouve des thesaurus plutôt *profonds* (AOD le thesaurus « Alcool and Other Drugs » - 10 niveaux et environ 2 000 termes par niveau), ainsi que des thesaurus *larges* (MedDRA - 5 niveaux hiérarchiques et environ 15 000 termes par niveau).
- un aspect qualitatif ou *sémantique*, lié aux critères qui ont mené au positionnement taxinomique des termes les uns par rapport aux autres. Par exemple pour un même terme, dans certains thesaurus l'aspect sémiologique du terme est le plus important, dans les autres c'est l'aspect lésionnel. Autrement dit, les relations taxinomiques n'ont pas le même sens d'un thesaurus à l'autre.

Les termes synonymes provenant de plusieurs thesaurus sont associés dans UMLS à un seul concept. De plus, les relations taxinomiques originaires des thesaurus sont conservées entre les concepts. UMLS est donc un réseau de concepts comprenant des relations de subsomption (Petiot 1996).

En conclusion, UMLS par l'entrelacement des thesaurus, facilite beaucoup la comparaison des concepts entre eux ainsi que leurs façons d'être regroupés.

¹ Nous avons utilisé sa version 2003AC (<http://umlsks.nlm.nih.gov>).

Etant donnée sa forme de réseau, on peut trouver dans UMLS plusieurs chemins entre deux concepts qui peuvent être composés d'un nombre d'arcs différent (Bodenreider, 2003). Tous les arcs ne traduisent pas le même degré de généralisation/specialisation. Ainsi, un chemin court d'un arc entre deux concepts peut correspondre à un passage « général – spécifique » rapide. En contrignant les trajets à avoir un grand nombre d'arc, on traverse de façon préférentielle des concepts issus de thésaurus présentant une plus grande profondeur avec une organisation conceptuelle du domaine plus précise.

Par ailleurs, UMLS propose un typage sémantique des concepts au sein du réseau sémantique. Ce typage réalise un regroupement très large et général des concepts (e.g. type « syndrome »). Il y a actuellement 135 types sémantiques pour une totalité d'1 million de concepts dans le réseau UMLS.

3 Matériel

La *base de connaissance* UMLS est utilisée dans sa version 2003AC mise à disposition en ligne et en tant que web service par la NLM.

Une *base d'exemples* est constituée. Elle est composée de 190 couples de concepts. Pour un couple donné, les concepts sont liés par la relation « is-a » dans l'ontologie des EI que nous avons à notre disposition mais ne sont pas en relation directe dans MedDRA. L'exemple que nous allons présenter lors des différentes étapes composant la méthode est celui du couple « Purpura » – is_a – « Bleeding Diathesis ».

L'apprentissage se fait sur 188 couples sélectionnés aléatoirement. Les deux couples restants sont utilisés pour vérifier les résultats obtenus à partir de la base d'apprentissage. Une méthode de validation de type « bagging » est envisagée dans le futur qui s'appuiera sur une série de tirages aléatoires de 188 couples. Dans le cadre du travail présenté, un seul tirage a été réalisé.

4 Méthode

Nous appelons « *trajet* » l'information contenue dans un chemin existant entre deux concepts (nœuds) connexes dans l'UMLS (incluant : concepts intermédiaires, informations liées aux relations). L'objectif de la méthode mise au point consiste à trouver des règles d'association pour chaque couple de la base d'apprentissage entre les trajets UMLS existants pour ce couple et la relation « is_a » dans l'ontologie des EI. L'idée sous-jacente à cette méthode est de prédire la relation de subsumption à partir des trajets existants entre deux concepts MedDRA dans l'UMLS.

La méthodologie élaborée pour mettre en œuvre et évaluer cette méthode a suivi trois étapes distinctes :

- 1) Acquisition de tous les trajets UMLS pour chaque couple de la base d'apprentissage.

2) Extraction de modèles de trajets définissant les règles d'association.

3) Validation des résultats ainsi obtenus. Dans l'état d'avancement actuel du travail, cette étape consiste à vérifier que les règles d'association appliquées sur un concept exemple permet de suggérer des concepts candidats à être liés au concept exemple par une relation de subsumption.

4.1 Formalisme pour la représentation des relations et trajets

Le modèle formel de représentation des relations de l'ontologie est un couple du type :

- $C_F - R - C_P$ où R est la relation « is_a_ontologique », C_F le *concept fils* et C_P le *concept père*.

Dans l'UMLS ce couple va prendre la forme $C_F - T_{UMLS} - C_P$, où :

- $T_{UMLS} = \{t_{UMLS} \mid C_F - t_{UMLS} - C_P\}$ est l'ensemble des trajets qui mettent en relation C_F et C_P .
- et $t_{UMLS} = R_{UMLS_1} - C_1 - R_{UMLS_2} - C_2 - \dots - C_{n-1} - R_{UMLS_n}$, un trajet en UMLS composé de relations et concepts UMLS (R_{UMLS} , C);

Les trajets sont décrits par les propriétés suivantes :

- pour les concepts
 - le nom (« arrhythmia », « stomach ulcer »...),
 - le type sémantique attribué à l'intérieur d'UMLS («disease», « finding »).
- pour la relation R_{UMLS} , que nous allons appeler « *arc* »
 - le type. Ce type a été attribué lors de la construction du metathesaurus. Il existe 11 types définis selon plusieurs critères. Ils indiquent la direction pour les relations purement hiérarchiques : CHD (child), PAR (parent), RN (narrower), RB (broader). Les autres types indiquent des relations associatives (part-of, caused-by, ...)
 - le nom du thesaurus d'origine.

4.2 Acquisition des données sur les trajets

Dans cette section nous présentons la méthodologie employée pour l'acquisition des trajets, les choix que nous avons fait concernant les informations pertinentes à retenir ainsi que les raisons qui nous ont amené à faire ces choix. Pour l'acquisition des trajets, nous avons implémenté un algorithme classique d'exploration en profondeur de graphes en éliminant les cycles (Bodenreider, 2001).

Les tests que nous avons faits pour la connectivité des concepts au sein de l'UMLS ont mis en avant que tous les types de relations ne sont pas susceptibles de nous aider. Ainsi, nous avons trouvé plusieurs centaines de trajets en deux arcs entre les concepts « Fracture » et « Hépatite ». Cette proximité purement topologique dans le réseau est trompeuse car ces deux concepts ont des sens éloignés. Sur la base de ce constat, nous

avons choisi de considérer exclusivement les relations de type hiérarchique incluses en UMLS. Ces relations sont typées comme : CHD, PAR, RN et RB.

Afin de mieux déterminer le domaine de couverture des concepts pris en compte nous avons restreint la liste des thesaurus source en ne gardant que ceux qui sont en anglais et dont le domaine d'application est la médecine clinique. Une quinzaine de thésaurus a été ainsi écartée.

Le choix du nombre des relations composant un trajet a été extrêmement difficile. Le constat qu'un trajet plus long passerait par des concepts intermédiaires plus proches sémantiquement plaide en faveur d'un choix des trajets longs. En effet, en empruntant un trajet long, on choisit systématiquement les relations appartenant à des thesaurus où la structure est plus profonde et en conséquence plus riche du point de vue relationnel. Dans une telle structure il est évident que les différences sémantiques entre les concepts subsumés et subsumants sont plus fines.

En ce qui concerne la variabilité en longueur des trajets, le traitement statistique n'est pas pertinent pour des trajets de longueurs différentes. L'acquisition de trajets de taille variable nous aurait obligé à créer des ensembles de trajets de même longueur pour pouvoir soumettre chaque ensemble au même traitement statistique.

Suite à ces constats, nous avons décidé qu'un nombre de 3 relations pour chaque trajet est un compromis avantageux entre la longueur et les ressources disponibles.

A la fin de cette étape nous obtenons pour chaque couple $C_F - R - C_P$ un nombre de trajets du type $C_F - t_{UMLS} - C_P$ qui vérifient les conditions :

- t_{UMLS} est composé uniquement de 3 relations R_{UMLS} (1)
- les R_{UMLS} sont d'un type inclus dans l'ensemble {CHD, PAR, RN, RB} et ont comme origine un des thesaurus que nous avons conservés. (2)

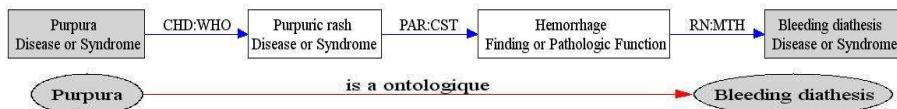


Fig. 1 : Exemple d'un trajet (en haut) pour le couple ontologique « Purpura » – is_a – « Bleeding Diathesis ». WHO, CST, MTH sont des acronymes représentant le thesaurus d'origine de la relation..

4.3 Extraction des modèles de trajets

L'objectif principal de cette étape est de trouver des règles d'association entre la relation « $is_a_ontologique$ », caractérisée par le couple de concepts d'ancre C_F, C_P et les trajets acquis. Nous définissons dans un premier temps, une classification des trajets. Les classes de trajets sont formées en considérant les types sémantiques pour les noeuds et les types de relation pour les arcs. Pour chaque couple de la base, les trajets sont organisés par groupes sémantiques selon cette classification (Figure 2).

Nous posons comme modèle de couple $C = (T_F, T_P)$, où T_F, T_P sont les types sémantiques des concepts C_F, C_P , et le modèle de trajet $M = R_{UMLS1} - T_1 - R_{UMLS2} - T_2 - R_{UMLS3}$, où R_{UMLS} est un type de relation et T le type sémantique du concept intermédiaire. (Figure 2)

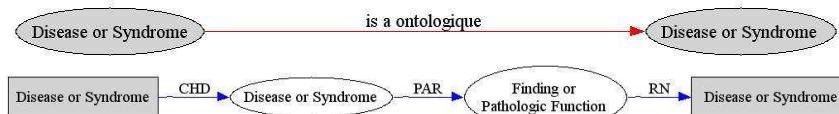


Fig. 2 : Le modèle de trajet de l'exemple du trajet présenté dans la figure 1.

Après la création de modèles de trajet nous avons obtenu comme résultat un tableau qui comprenait pour chaque modèle de couple C le nombre de trajets du modèle M qui ont été trouvés. Pour l'exemple montré en figure 2, 4 trajets du modèle présenté (figure 2 en bas) ont été trouvés pour l'ensemble des couples appartenant au modèle de couple présenté (figure 2 en haut).

Ensuite nous avons calculé le rapport des cotes (*odds ratio-OR*) pour le nombre d'occurrences de chaque modèle de trajet M , afin d'exprimer de façon synthétique la spécificité statistique de chaque association modèle de trajet – modèle de couple ($C - M$). Pour chacune des associations on construit le tableau de contingence dans lequel :

Tableau 1 : Tableau de contingence 2x2.

	M_j	$\neg M_j$
C_i	a	b
$\neg C_i$	c	d

- a est le nombre de trajets des couples du type C_i avec le modèle M_j
- b est le nombre de trajets des couples du type C_i avec un modèle autre que M_j
- c est le nombre de trajets basés sur le modèle M_j mais qui ne décrivent pas les couples du type C_i
- d est le nombre de trajets qui ne sont pas basés sur le modèle M_j et qui ne décrivent pas les couples du type C_i

L'odds ratio est le rapport entre (a/c) et (b/d) .

Nous avons choisi pour chaque C toutes les M pour lesquels $OR > 1$ (Li 2003). Nous avons obtenu ainsi un ensemble de règles d'association $C \rightarrow \{(M_1, OR_1); (M_2, OR_2); \dots; (M_n, OR_n)\}$ pour chaque modèle de couple.

4.4 Vérification

Nous considérons les deux couples de concepts choisis aléatoirement pour la vérification. Pour un couple donné, nous avons identifié le modèle auquel ce couple (C) appartient, c'est-à-dire les types sémantiques de C_F et C_P . Ensuite, nous avons retenu les modèles de trajets (M) qui ont été associés à ce modèle au cours de l'apprentissage, c'est-à-dire les règles d'apprentissage

Nous avons exploré l'UMLS à partir du C_F en respectant les modèles de trajets définis dans la règle d'association. Nous avons retenu tous les concepts qui se trouvent au bout de chaque trajet qui respecte un des modèles. Nous avons appelé ces concepts, *concepts présumés subsumant* (C_{PS}).

Soit MT l'ensemble des modèles de trajets pour lesquels on trouve au moins un trajet dans le cas d'un couple utilisé pour la vérification. Pour mettre en évidence la pertinence des C_{PS} trouvés nous avons mis au point un système de scores. Ainsi nous avons considéré comme score maximum la somme des OR des éléments de MT . Un score individuel est défini pour chaque C_{PS} comme la somme des OR des modèles de trajets (M) pour lesquelles nous avons trouvé au moins un trajet qui menait à C_{PS} . Ces scores individuels ont été exprimés comme un pourcentage du score maximum que nous avons appelé *niveau de pertinence* du C_{PS} . Une forte valeur du niveau de pertinence signifie que le C_{PS} a été trouvé en suivant des trajets correspondant à des T fortement associés au C modélisant (C_F, C_P).

5 Résultats

5.1 Base de trajets

19234 trajets ont été recueillis pour 128 (68%) couples sur 188. Dans 60 couples (31%) aucun trajet en 3 arcs n'a été trouvé. Le nombre moyen de trajets par couple était de 205.

Dans le tableau suivant nous présentons quelques exemples de trajets entre les concepts « Purpura » (C_F) et « Diathèse hémorragique » (C_P). Pour ce couple on trouve 57 trajets dans l'UMLS.

Tableau 2 : Exemples de trajets entre « Purpura » et « Diathèse hémorragique ».

R_{UMLS1}	C_1	R_{UMLS2}	C_2	$-R_{UMLS3}$
CHD	Purpura, Thrombocytopenic	PAR	Hemorrhage	RN
CSP	<i>Disease or Syndrome</i>	CST	<i>Finding Pathologic Function</i>	MTH
CHD	Purpura Fulminans	PAR	Hemorrhage	RN
SNM	<i>Finding Disease or Syndrome</i>	CST	<i>Finding Pathologic Function</i>	MTH
CHD	Purpuric rash	PAR	Hemorrhage	RN
WHO	<i>Disease or Syndrome</i>	CST	<i>Finding Pathologic Function</i>	MTH
PAR	Disease of capillaries	CHD	Ecchymosis	PAR
CST	<i>Disease or Syndrome</i>	CST	<i>Finding Pathologic Function</i>	MDR

5.2 Traitement statistique

Les 19234 trajets ont produit 488 modèles de trajets. Dans 443 (91%) des modèles de trajets, le nombre de trajets correspondants était inférieur à 100. Les 45 modèles de trajets restants généralisent 12670 trajets, soit 65,8% du total des trajets. Les modèles de couples sont au nombre de 22. Pour 4 modèles de couples on a trouvé un nombre de trajets de 15674, soit 81% du nombre total des trajets. 10 modèles de couples n'ont

aucun modèle de trajet correspondant qui ait un OR supérieur à 1. Le nombre moyen des modèles de trajet avec un OR supérieur à 1 par modèle de couple est de 9,8.

5.3 Vérification

Nous avons réalisé une vérification sur deux couples aléatoirement choisis parmi les 190 de la base d'exemple. Le premier couple a les caractéristiques suivantes :

- $(C_F, C_P) = (\text{« Tachycardia, Paroxysmal »}, \text{« Tachycardia »})$
- le modèle de couple $C = (\text{« Disease or syndrome »}, \text{« Finding »})$

En explorant l'ensemble de ces modèles de trajets on part du C_F pour aboutir à 43 C_{PS} . La figure 3 montre le niveau de pertinence de chaque C_{PS} .

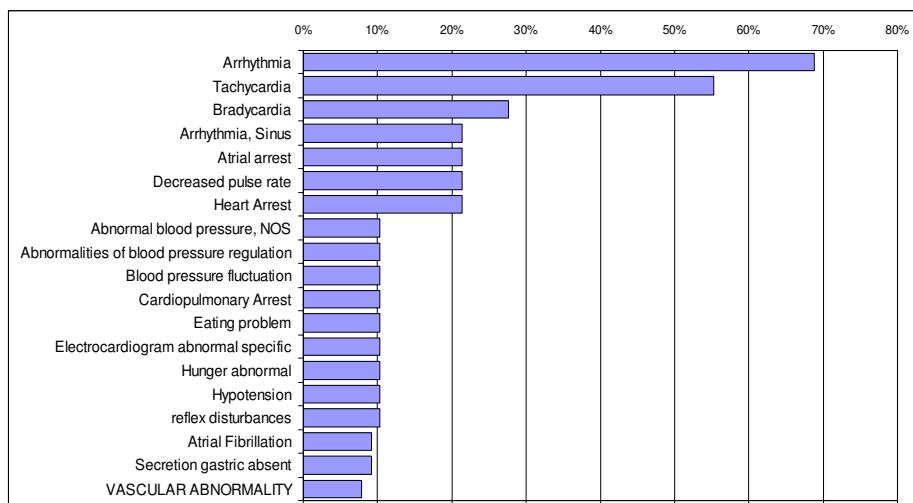


Fig. 3 : Les scores des C_{PS} trouvés pour le concept de validation « Tachycardia, Paroxysmal ». Un nombre de 24 C_{PS} ayant un score de pertinence inférieur à 8% n'ont pas été représentés dans cette figure.

Notre méthode met en évidence une différence nette entre les deux premiers concepts classés et les autres. Le concept que nous recherchions est classé deuxième dans cette liste (« tachycardia »). Le concept « arrhythmia », classé en premier, est le concept qui subsume « tachycardia » dans l'ontologie des effets indésirables. Pour l'autre couple utilisé pour la vérification (« Tachycardia, Supraventricular », « tachycardia ») nous avons obtenu des résultats similaires. Entre les plus pertinents (16) C_{PS} trouvés 5 sont des arythmies et 8 des affections cardiovasculaires entraînant une arythmie.

6 Discussion et conclusion

Plusieurs études se sont attachées à identifier dans l'UMLS des relations en un arc entre des concepts ciblés. Le taux partiel de couverture obtenu grâce à ces relations nous a encouragé à développer une méthode capable d'identifier des relations comportant un nombre d'arcs supérieur à un. Nous avons proposé dans ce travail une méthode originale pour l'apprentissage de nouvelles relations de subsumption dans une classification basée sur une hiérarchie stricte. Notre cible, la terminologie MedDRA, devrait bénéficier d'une telle approche pour retrouver les liens hiérarchiques permettant de regrouper ses concepts de façon pertinente pour la détection du signal. Nous avons identifié des modèles de trajets entre concepts UMLS qui sont spécifiques pour une relation de type *is_a* ontologique. Cette étude montre l'intérêt de l'UMLS pour l'extraction de connaissance et l'aide à la construction d'une ontologie.

Bien que cette méthode soit en grande partie automatisable, l'intervention d'un expert est nécessaire afin de valider les concepts proposés subsumants proposés par le système et éventuellement choisir un ou plusieurs couples parmi les couples qui présentent les niveaux de pertinence les plus élevés.

Une validation plus ample (encore embryonnaire dans le travail accompli) ainsi que l'affinage des paramètres de la méthode d'apprentissage rendront certainement la validation experte moins importante pour le résultat final.

En ce qui concerne la description des trajets nous avons limité cette première étude aux types attachés aux relations taxinomiques (PAR, CHD, etc.). Des informations supplémentaires disponibles dans l'UMLS doivent être prise en compte, comme par exemple les relations associatives issues de la SNOMED CT ou autres. De la même façon la méthodologie d'apprentissage des règles d'association peut encore être optimisée. La prise en compte du facteur de Jaccard, du cosinus, du Laplace et d'autres mesures statistiques, comme des indicateurs de pertinence des règles trouvées seraient des optimisations envisageables (Li 2003).

Nous avons choisi d'une façon expérimentale des trajets en 3 arcs. Des comparaisons de résultats obtenus avec des trajets d'autres longueurs viendraient compléter cette étude et préciser son intérêt.

Dans la même direction l'affinage de l'apprentissage pourra porter sur l'utilisation de regroupements proposés par l'UMLS en s'appuyant sur des relations existantes entre les types sémantiques. De la même façon, une relation qui relie les mêmes concepts dans plusieurs terminologies peut être considérée plus pertinente.

A notre connaissance, il n'existe aucune méthode disponible, qui pour un nombre de concepts et un domaine d'interprétation établis, soit capable de filtrer les relations tout en conservant la sémantique que l'on souhaite leur donner. La méthode que nous proposons a comme objectif à long terme de mettre à disposition une telle méthode, qui devrait être suffisamment sensible pour détecter les relations dont nous avons besoin et avoir une bonne spécificité afin d'éliminer les nombreuses relations contenues dans l'UMLS qui n'apportent rien à la modélisation du domaine.

Par rapport aux critères d'évaluation de notre approche, cette méthode basée sur UMLS est une réutilisation effective d'une connaissance déjà acquise. Cette méthode

est actuellement utilisée pour enrichir l'ontologie existante, elle facilite le travail de l'expert et aboutit à un coût de développement plus réduit. Le coût de construction et la reproductibilité dans des contextes différents restent encore à estimer.

Références

- AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus . Actes des journées francophones d'Ingénierie des connaissances, Toulouse, 2000; p 93-104
- BACHIMONT B.(2000) engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances. *In : Ingénierie des connaissances*.
- BODENREIDER O. (2001), Circular hierarchical relationships in the UMLS : Etiology, Diagnosis, Treatment, Complications and Preventions, *Proc. of AMIA 2001*, p.57-61
- BODENREIDER O. (2003), Strength in numbers: Exploring redundancy in hierarchical relations across biomedical terminologies, *AMIA 2003 Symposium Proceedings*,101.
- BOUSQUET C, LAGIER G, LILLO-LE LOUËT A, LE BELLER C, VENOT A, JAULENT M.C., (2005) Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reaction. *Drug Saf 2005*; 28(1):19-34.
- BROWN EG. (2002) Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART. *Drug Saf 2002*; 25(6):445-52.
- CHARLET J. (2003), L'Ingénierie des connaissances, Développements, Résultats et Perspectives pour la gestion des connaissances médicales, *Mémoire d'habilitation à diriger des recherches*, 2003
- CIMINO J. J., « Coding Systems in Health Care », VAN BEMMEL J. H., MCCRAY A. T., Eds., *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*, p. 71– 85, Stuttgart, 1996.
- COTÉ R.A., ROTHWELL D.J., PALOTAY J.L., BECKET R.S., BROCHU L., SNOMED International. College of American Pathologists (1993)
- HENEGAR C, BOUSQUET C, LILLO-LE LOUËT A.,DEGOULET P., JAULENT M.C. (2004) A knowledge based approach for automated signal generation, *Medinfo 2004*; 2004:626-30.
- LE MOIGNO S., CHARLET J., BOURIGAULT D., DEGOULET P., JAULENT M-C., Terminology extraction from text to build an ontology in surgical intensive car,. Proc AMIA Symp., 430-4 (2002)
- LI J., ZHANG Y. (2003) Direct interesting rule generation, *Proceedings of The Third IEEE International Conference on Data Mining (ICDM)*, 2003, 155 – 162, IEEE computer society.
- LINDBERG D.A., HUMPHREYS B.L., MCCRAY A.T., « The Unified Medical Language System », Methods Inf Med, vol. 32, n° 4, 1993, p. 281-91.
- NATIONAL LIBRARY OF MEDICINE, Bethesda, Maryland, « Medical Subject Headings », 2001,disponible à www.nlm.nih.gov/mesh/meshhome.html.
- ORGANISATION MONDIALE DE LA SANTE, Genève, « Classification statistique internationale des maladies et des problèmes de santé connexes— Dixième révision », 1993.
- PETIOT D., BURGUN A., LE BEUX P. (1996) Modelisation of a criterion of proximity : application to medical thesauri. Brender J. *Medical Informatics Europe 1996*. IOS Press,149-53.
- YOKOTSUKA M, AOYAMA M, KUBOTA K. (2000) The use of a medical dictionary for regulatory activities terminology (MedDRA) in prescription-event monitoring in Japan (J-PEM). *Int J Med Inf 2000*; 57(2-3):139-53.
- ZWEIGENBAUM P., « Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances », *Innovation Stratégique en Information de Santé*, no 2-3, 1999, p. 27-47.

Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques

Florence Amardeilh^{1,2}, Philippe Laublet¹, Jean-Luc Minel¹

¹ Laboratoire LaLICC, Université Paris IV,
`{prenom.nom}@paris4.sorbonne.fr`

² Mondeca, Département R&D, Paris
`{prenom.nom}@mondeca.com`

Résumé : Dans cet article, nous présentons une plate-forme de peuplement semi-automatique d'ontologies à partir de documents textuels. Notre plate-forme fournit un environnement permettant la mise en correspondance des extractions linguistiques avec l'ontologie du domaine de l'application cliente à l'aide de règles d'acquisition de connaissance. Ces règles s'appliquent, pour chaque étiquette linguistique pertinente, aussi bien à un concept, qu'à un de ses attributs ou encore à une relation sémantique entre plusieurs concepts. Elles déclenchent l'instanciation de ces concepts, attributs et relations dans la base de connaissance relative à l'ontologie du domaine. Ce papier détaille le processus et présente les premières expérimentations réalisées à partir d'un cas client provenant de l'édition juridique.

Mots-clés : Méthodologie de population de ressources terminologiques et ontologiques, Modèles de connaissances, Ontologies, Fouille de textes, Web Sémantique.

1 Introduction

Dans les métiers liés au monde documentaire comme l'édition, la documentation, la veille stratégique, etc., les professionnels doivent chaque jour traiter de grands volumes de données provenant de diverses sources documentaires. A partir de l'ensemble des documents qu'ils sont chargés d'étudier, ceux-ci doivent d'abord sélectionner les ressources documentaires pertinentes pour leur travail puis, pour chacune d'entre elles, extraire manuellement l'information pertinente. Cette information sert ensuite à annoter le document par un ensemble de descripteurs (termes du thesaurus et entités nommées comme les noms de personnes) et à enrichir leur base de connaissance (entités nommées, attributs de ces entités nommées et relations sémantiques entre ces entités nommées).

Dans le contexte du Web Sémantique, le contenu d'un document peut être décrit et annoté à l'aide de langages de représentation des connaissances comme RDF, XTM et OWL. RDF, the Resource Description Framework¹, est un formalisme de

¹ ORA L. & SWICK R. (1999). Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation.

représentation des connaissances, issu des réseaux sémantiques, dont la syntaxe utilise XML. Il sert à décrire des ressources documentaires par un ensemble de métadonnées (auteur, date, source, descripteurs, etc.). Ces métadonnées sont constituées sous la forme de triplets : (sujet, verbe, objet) ou (objet 1, relation, objet 2) ou encore (ressource, propriété, valeur) selon le type de description nécessaire.

Les Topic Maps sont un autre formalisme de représentation des connaissances qui dispose aussi d'une syntaxe basée sur XML (Park, 2003). Les Topic Maps définissent un ensemble de sujets relatifs à un même domaine avec des interactions entre eux formant ainsi une carte sémantique de la connaissance. Un sujet représente tout ce qui peut être décrit ou pensé par un humain. Il peut participer à une ou plusieurs relations, appelées associations, dans lesquelles il joue un rôle spécifique. Les sujets ont également au moins un nom et des propriétés intrinsèques, appelées occurrences. Ce langage permet une grande flexibilité de représentation des connaissances, particulièrement pour la modélisation de relations sémantiques complexes (n-ary).

OWL, Ontology Web Language², permet de formaliser une ontologie (Gruber, 1993), ou plus globalement des ressources terminologiques et ontologiques (Bourigault, 2004), par la définition des concepts utilisés pour représenter un domaine de connaissance. Ce langage permet de décrire ces concepts par un ensemble de propriétés, de relations et de contraintes. Le formalisme utilisé correspond à ceux de certaines logiques de description.

Dans nos projets, nous utilisons RDF pour décrire le contenu d'une ressource documentaire, OWL pour modéliser l'ontologie qui représentera une vision métier ou applicative du domaine étudié et les Topic Maps pour construire la base de connaissance qui contiendra les instances des concepts, propriétés et relations décrits dans l'ontologie du domaine. Les informations pertinentes au domaine, contenues dans les ressources documentaires, servent à instancier la base de connaissance et à créer les annotations documentaires pour pouvoir être interprétables par les machines qui pourront les partager, les publier, les rechercher et les utiliser de manière générale (Laublet, 2002).

L'annotation documentaire et le peuplement d'ontologie dépendent fortement des informations extraites des articles par les professionnels. Ce traitement manuel des documents est extrêmement coûteux en temps et en ressources. L'ensemble du processus pose également des problèmes en terme de productivité et de qualité. Pour toutes ces raisons, les entreprises cherchent de plus en plus à mettre en place des solutions basées sur l'utilisation d'outils linguistiques pour extraire (semi-) automatiquement les informations pertinentes des documents textuels. Ces outils du traitement automatique du langage naturel devront s'intégrer étroitement aux futures applications du Web Sémantique et seront même essentiels au développement, à l'acceptation et à l'utilisation du Web Sémantique (Bontcheva, 2003). Grâce aux fonctionnalités offertes par les technologies du Traitement du Langage Naturel, et notamment celles de l'Extraction d'Information, des solutions adaptées aux besoins du

² HENDLER J., HORROCKS I. et al. (2004). OWL web ontology language reference, W3C Recommendation.

Web Sémantique peuvent être implémentées telles que : la construction semi-automatique de vocabulaires/terminologies d'un domaine à partir d'un corpus documentaire représentatif ainsi que leur maintenance (Bourigault, 2004) ; l'enrichissement semi-automatique de bases de connaissance par les entités nommées et les relations sémantiques extraites des documents textuels après validation (Kyriakov, 2003) ; l'annotation sémantique de ressources documentaires (Kahan, 2001) (Handschoh, 2002) (Vargas-Vera, 2002).

Néanmoins, nous avons constaté dans nos propres projets applicatifs que les outils linguistiques et la modélisation de l'ontologie du domaine client sont implémentés indépendamment l'un de l'autre et non pas l'un pour l'autre comme c'est le cas dans les travaux de recherche cités ci-dessus. C'est pourquoi nous nous sommes intéressés à la mise en place d'une passerelle entre d'un côté les résultats de ces outils linguistiques et de l'autre l'ontologie.

Dans cet article, nous présentons donc une nouvelle plateforme d'annotation documentaire et d'acquisition de connaissances. Dans la prochaine section de ce papier, nous montrerons l'émergence de la problématique actuelle et nous décrirons la mise en place de notre solution. Ensuite, nous présenterons les résultats de premières expérimentations dans le domaine de l'édition juridique. Ce projet permettra aussi d'illustrer notre travail tout au long de cet article. Enfin, nous synthétiserons ces résultats afin d'apporter de nouvelles réflexions pour conclure dans la dernière partie sur les perspectives futures de nos travaux de recherche.

2 Intégration d'outils linguistiques dans un portail Web Sémantique

2.1 Les outils utilisés

Notre solution est basée sur l'outil « Intelligent Topic Manager™ » (ITM) de la société Mondeca. ITM est une plateforme logicielle pour la gestion de connaissance et l'exploitation d'ontologies. ITM intègre un portail sémantique, décrit dans (Amardeilh, 2004), fournissant quatre fonctions clefs : l'Edition, la Recherche, la Navigation et la Publication. L'ontologie cliente, formalisée en OWL, impose ses contraintes de modélisation à la base de connaissance (implémentée en Topic Maps), aux interfaces utilisateurs ainsi qu'à toutes les fonctionnalités du portail. Les éléments de la base de connaissance pointent vers les documents, accessibles par URL sur Internet ou dans un système de gestion de contenus.

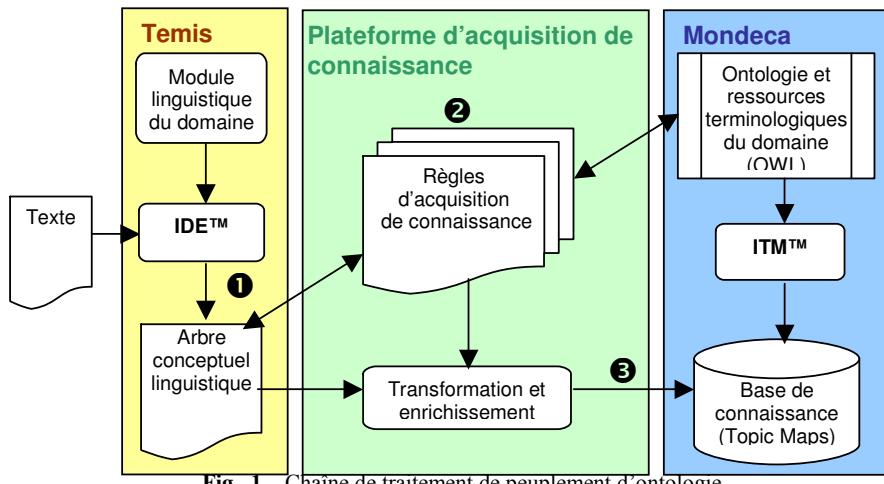
L'analyse linguistique est effectuée par l'Insight Discoverer™ Extractor (IDE) développé par la société Temis. Cet outil implémente une méthode d'automates à états finis (Grivel, 2001) s'appuyant sur un prétraitement regroupant la segmentation des documents en unités textuelles, la lemmatisation et l'analyse morpho-syntactique de ces unités textuelles. En sortie, l'IDE™ produit un arbre conceptuel étiqueté. Chaque nœud de l'arbre porte le nom d'une étiquette sémantique attribuée à l'unité textuelle extraite en fonction du domaine traité (cf. exemple section 2.2.1).

D'une part, le portail d'ITM™ ne permet pas un enrichissement (semi-)automatisé de sa base de connaissance. D'autre part, l'outil d'extraction d'information, l'IDE™, une fois les extractions réalisées sur un corpus documentaire, présente simplement les informations à l'utilisateur dans une interface html, sans les enregistrer dans une base de données ou plus encore dans une base de connaissance, pour être ultérieurement exploitées. Les deux sociétés ont décidé de collaborer sur plusieurs projets client (services de documentation, veille économique ou édition). Néanmoins, le paramétrage de leurs outils pour une application cliente est toujours réalisé indépendamment l'un de l'autre, chacun comportant ses propres contraintes.

En effet, Mondeca construit l'ontologie du domaine, si elle n'existe pas déjà, en fonction du client, de ses besoins et des données déjà existantes. Temis construit des modules d'extractions linguistiques propres à chaque projet tout en réutilisant, lorsque cela est possible, tout ou partie de modules d'extraction existants (comme celui des entités nommées). Par conséquent, les dénominations des étiquettes linguistiques de l'arbre conceptuel produit par l'IDE sont généralement indépendantes de celles des concepts de l'ontologie même si elles décrivent le même sujet. Il nous faut donc définir un moyen de les faire correspondre et ainsi pouvoir instancier les bons concepts de l'ontologie à partir des extractions linguistiques.

2.2 Intégration ITM/IDE dans le portail Web Sémantique

L'intégration entre les extractions linguistiques de l'IDE et les concepts ontologiques du domaine définis dans ITM doit se faire en plusieurs étapes : 1) parcours de l'arbre conceptuel résultant de l'analyse linguistique ; 2) définition manuelle des règles d'acquisition entre étiquettes linguistiques et concepts ontologiques; 3) déclenchement automatisé des règles d'acquisition sur les textes.



La chaîne de traitement décrite ci-dessus est appliquée à chacun de nos projets client. Nous allons illustrer ce processus à travers un projet concernant le domaine de l'édition juridique : les auteurs d'articles juridiques doivent se tenir informés de l'ensemble des textes de lois et des décisions des cours de justice. Ainsi pour tout texte paru, une référence est enregistrée dans la base de connaissance avec toutes ses propriétés ainsi que les références aux autres textes de lois cités. Le corpus utilisé dans notre exemple est constitué uniquement de comptes rendus de décisions issues de cours de cassation à propos de divorces ou de contrats de travail. Les comptes rendus (cf. Fig. 2) se divisent en deux parties : tout d'abord un en-tête semi-structuré représente les informations liées à cette décision (date, cour de cassation, n° de décision, n° de pourvoi, etc.) et ensuite le corps du document (texte non structuré) décrit dans l'ordre les parties impliquées, les motifs, l'argumentation avec les références aux textes de lois codés (dits « TC », par exemple « Code civil ») et non codés (dits « TNC » comme « Décret du 30 septembre 1953 »).

CIV. 1	D.S
COUR DE CASSATION	
Audience publique du 23 mars 2004	Cassation partielle
M. BOUSCHARAIN, président	Arrêt n° 510 F-D
Arrêt n° 510 F-D	
Pourvoi n° F 02-19.839	
(...)	
REPUBLIQUE FRANCAISE	
AU NOM DU PEUPLE FRANCAIS	
<hr/> LA COUR DE CASSATION, PREMIÈRE CHAMBRE CIVILE, a rendu l'arrêt suivant :	
Sur le pourvoi formé par Mme X, épouse Y, demeurant xxxxx, 75019 Paris,	
(...) Sur le rapport de Mme G-D, conseiller référendaire, les observations de Me B-H, avocat de Mme X, de la SCP VO, avocat de la société XYZ, les conclusions de Mme P, avocat général, et après en avoir délibéré conformément à la loi ;	
Sur le moyen unique, pris en sa seconde branche :	
Vu l'article L. 311-37 du Code de la consommation, dans sa rédaction antérieure à la loi n°2001-1168 du 11 décembre 2001 ;	
(...) 	

Fig. 2 – Extrait d'un compte rendu de décision d'une cour de cassation

2.2.1 Arbre conceptuel résultant de l'analyse linguistique

Comme nous l'avons mentionné plus haut, l'IDE produit un arbre conceptuel à partir de chaque analyse linguistique d'un compte rendu (cf. Fig. 3). A chaque nœud de cet arbre correspond une étiquette linguistique et sa forme textuelle issue du compte rendu, rappelée entre parenthèse. Notre solution doit parcourir cet arbre étiqueté évalué afin d'en extraire l'information pertinente et la rapprocher d'un concept de l'ontologie, que ce dernier soit un sujet, un attribut, une association ou un rôle dans la base de connaissance.

```

/REFERENCE DECISION(cassation 10400510)
/FORMATION(CIV . 1)
    /Chambre civile(CIV . 1)
/JURIDICTION(COUR DE CASSATION)
/DATE SEANCE(Audience publique du 23 mars 2004)
    /DATE(23 mars 2004)
        /MonthDayNumber(23)
        /month(mars)
        /YearNumber(2004)
/Noms-de-personnes(M. BOUSCHARAIN , président)
    Nom(M. BOUSCHARAIN)
    role(président)
        /Role/Juridique(président)
/DECISION/ARRET/ARRET DIFUSE(Arrêt n° 510 F-D)
    num(510 F-D)
/POURVOI(Pourvoi n° F 02-19.839)
    num(F 02-19.839)
...
/REFERENCE(article L. 311-37 du Code de la consommation)
    ref(article L. 311-37 du Code de la consommation)
        /ARTICLE unique(article L. 311-37)
            art num(L. 311-37)
        TEXTE(Code de la consommation)
        /CODE/Code consommation(Code de la consommation)

```

Fig. 3 – Extrait d'un arbre conceptuel d'un compte rendu de décision juridique

Le parcours de l'arbre est régenté par quelques principes de base : 1) Un arbre possède nécessairement une racine père, représentant ici le plus souvent le document ou le sujet principal (ici la décision elle-même). 2) Le parcours de l'arbre s'effectue en profondeur par ordre *préfixe* : partant de la racine, l'algorithme parcourt d'abord le fils gauche avant de parcourir le fils droit et ainsi de suite récursivement. 3) Deux parcours de l'arbre sont nécessaires : le premier pour acquérir les sujets avec leurs attributs et le second pour acquérir les associations avec les différents rôles joués par les sujets dans celles-ci.

Ces deux parcours sont primordiaux car tous les sujets ne jouent pas nécessairement un rôle dans une association. Ils ne seraient donc pas instanciés si l'arbre était parcouru en n'y repérant que les associations, puis leurs rôles et enfin leurs sujets. C'est notamment le cas des sujets « Personnalités », ayant des attributs « Nom » et « Rôle », qui ne participent à aucune association.

Afin de traiter l'arbre conceptuel, nous avons choisi, dans une première étape, d'implémenter les règles d'acquisition en langage XPath³. En effet, ce langage permet de parcourir un arbre (document XML, arbre conceptuel, etc.), d'atteindre directement n'importe lequel de ses noeuds et à partir d'un noeud quelconque de sélectionner n'importe lequel de ses descendants, descendants ou frères.

³ Site web du W3C : <http://www.w3.org/TR/xpath>

2.2.2 Définition des règles d'acquisition

Nom de l'étiquette linguistique	Nom du concept ontologique	Type dans la base	Contexte
/nom lex	Personne	Sujet	
/noms lex	Personne	Sujet	
/MEMBRES COUR	Personnalité Juridique	Sujet	\exists Descendant = /Juridique
	Personnalité Politique	Sujet	\exists Descendant = /Politique
/REFERENCE	Réf Editoriale Législative TNC	Sujet	$\exists!$ Fils = /article
	Réf Editoriale Législative TNC Article	Sujet	\exists Fils = /article
	Renvoi simple	Association	\exists Père = /REFERENCE DECISION
	Cible lien	Rôle	\exists Père = /REFERENCE DECISION
/art num	Num Article	Attribut	
/MOTIF			
	Origine Lien	Rôle	

Tableau 1 – Exemples de mise en correspondance d'étiquettes et de concepts

Chaque nœud de l'arbre conceptuel doit manuellement être rapproché de son concept ontologique correspondant, quelque soit son type (sujet, attribut, association et rôle)⁴. Pour cela, nous construisons des règles d'acquisition de la connaissance qui permettront de déclencher la création d'une instance du concept ontologique à chaque nœud correspondant de l'arbre conceptuel. Le tableau ci-dessus résume les différents cas possibles :

- Une étiquette correspond à un seul concept : « /art num » pour l'attribut « Num Article ».
- Plusieurs étiquettes correspondent au même concept : « /Nom lex » et « /Noms lex » pour le sujet « Personne ».
- Une étiquette correspond à plusieurs concepts du même type : « /MEMBRES COUR » pour les sujets « Personnalité Juridique » et « Personnalité Politique ».
- Une étiquette correspond à plusieurs concepts de types différents : « /REFERENCE » pour les sujets « Réf Editoriale Législative TNC » et « Réf Editoriale Législative TNC Article », l'association « Renvoi simple » et le rôle « Cible lien ».
- Une étiquette ne correspond à aucun concept de l'ontologie : « /MOTIF ».
- Un concept n'a pas d'équivalence dans l'ensemble des étiquettes existantes : le rôle « Origine Lien ».

⁴ Rappelons que le vocabulaire utilisé est celui des Topic Maps (Park, 2003).

Dans les cas où une étiquette peut donner linstanciation de plusieurs concepts, il faut alors utiliser le contexte des nœuds ascendants, descendants ou frères pour résoudre les ambiguïtés. Par exemple, si le nœud « /REFERENCE » a un nœud fils « /article », le sujet « Réf Editoriale Législative TNC Article » sera instancié, sinon il s’agira de « Réf Editoriale Législative TNC ».

La première partie d’un compte rendu, et par conséquent des extractions linguistiques, concerne la décision de la cour de cassation. Elle contient donc les attributs du concept représentant cette décision, i.e. « Réf Editoriale Jurisprudence » marqué par l’étiquette « /REFERENCE DECISION ». Il est donc possible de mettre en relation chacun des noeuds de cette première partie avec les attributs correspondants, telle l’étiquette « /FORMATION » avec « formation » (cf. Fig. 4).

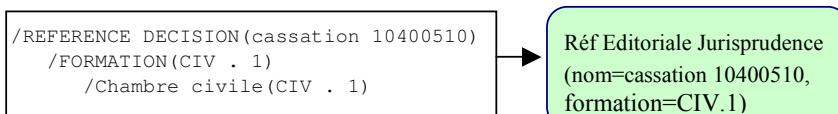


Fig. 4 – Extraction linguistique du Sujet « Réf Editoriale Jurisprudence »

La deuxième partie du document permettra de recueillir d’autres types d’instances de concepts, notamment les personnes, que ce soit des parties ou des personnalités juridiques (avocats, présidents, greffiers, etc.), et les références aux textes juridiques sur lesquels se base l’argumentation des différentes parties. Ces références seront instanciées selon leur concept, texte codé ou non codé, avec leurs attributs (date, type de texte, etc.) puis mises en relation avec la décision à travers l’association « Renvoi simple » et la spécification de leur rôle « Cible lien » (cf. Fig. 5).

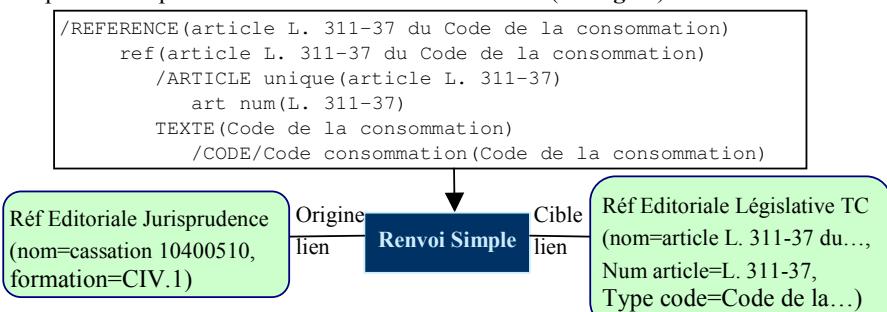


Fig. 5 – Extraction linguistique modélisable en Association « Renvoi simple »

Une fois la mise en correspondance définie, chacune des règles d’acquisition sera formalisée en langage Xpath et ajoutée dans l’ontologie du domaine sur le concept qu’elle va instancier. Par exemple, le concept «Réf Editoriale Législative TC Article» aura dans l’ontologie la règle d’acquisition associée « /REFERENCE DECISION/REFERENCE/ref[ARTICLE and TEXTE] ».

2.2.3 Déclenchement d'une règle d'acquisition

Après analyse linguistique, l'arbre conceptuel du document sélectionné par l'utilisateur est parcouru automatiquement par l'ensemble des règles d'acquisition. A chaque nœud pertinent, l'action d'instanciation de la base de connaissance, associée à toute règle d'acquisition, est déclenchée. Toutefois, afin d'éviter les doublons dans la base de connaissance, un contrôle est effectué avant la création du concept pour vérifier son existence dans la base de connaissance. Une fois le parcours de l'arbre terminé, l'utilisateur peut visualiser toutes les nouvelles instances de la base de connaissance au moyen d'une interface de validation. A partir de cette interface, l'utilisateur peut modifier et/ou supprimer une instance créée, ainsi qu'en ajouter de nouvelles. Grâce à cette interface, l'utilisateur peut contrôler la qualité de la base de connaissance ainsi enrichie.

3 Expérimentations et résultats

Notre corpus d'expérimentation est constitué de 36 comptes rendus de décisions de cour de cassation. Sur ces 36 comptes rendus, quatre seulement ont servi à définir les règles d'acquisition manuellement. Les 32 documents restants ont été utilisés comme corpus de test. Après réception des patrons d'extraction construits et compilés par les linguistes de Temis, nous avons traité l'ensemble du corpus de test et recueilli chaque arbre conceptuel. Nous avons comparé les étiquettes linguistiques avec chaque concept repéré et constaté quels étaient ceux correctement créés, incorrectement créés ou non créés dans la base de connaissance.

Afin d'évaluer quantitativement les résultats de ces traitements, nous avons utilisé les mesures de précision et de rappel, définies pour mesurer soit des résultats en recherche d'information (cf. conférences TREC), soit des résultats d'extraction d'information (Cf. conférences MUC). Dans notre cas, nous avons appliqué ces mesures aux extractions linguistiques étiquetées vis-à-vis des concepts instanciés dans la base de connaissance. La **Précision** mesure le nombre d'instances correctement acquises divisé par le nombre d'instances acquises et le **Rappel** mesure le nombre d'instances correctement acquises divisé par le nombre d'instances existantes dans l'arbre conceptuel.

Suite à l'analyse des 32 documents du corpus de test, et à partir des mêmes règles d'acquisition définies précédemment, le tableau ci-dessous présente les résultats sur l'ensemble des concepts présents dans le corpus des extractions linguistiques. Un ensemble de 1765 concepts de l'ontologie répartis en sujets, attributs (ou occurrences) de ces sujets, associations et rôles sont présents dans les arbres conceptuels du corpus. Parmi ces concepts, 975 ont été correctement instanciés par les règles, 257 incorrectement instanciés et enfin 533 non instanciés. En moyenne, nous obtenons donc un rappel de 0,55 et une précision de 0,79.

En résumé, même si la précision est plutôt satisfaisante pour une première expérimentation, nous constatons qu'un nombre important d'unités textuelles,

pourtant correctement étiquetées dans l’arbre, ne sont pas instanciées par la suite, surtout en ce qui concerne les attributs et les associations. D’autres concepts sont incorrectement instanciés, notamment les sujets. Ceci est principalement dû à un problème de redondance lié à des règles conflictuelles. Ce problème se répercute alors sur les rôles avec le non respect des contraintes modélisées dans l’ontologie, notamment les cardinalités, engendrant pour une même association plusieurs rôles du même type au lieu d’un seul.

Type de concept	Nombre de concepts dans l’arbre (A)	Nombre instanciés corrects (B)	Nombre instanciés incorrects (C)	Nombre non instanciés (D)	Rappel (B/A)	Précision (B/B+C)
Sujets	585	432	139	14	0,74	0,76
Attributs	798	329	0	469	0,41	1
Associations	80	69	0	11	0,93	1
Rôles	302	145	118	39	0,48	0,55
Total	1765	975	257	533	0,55	0,79

Tableau 2 – Résultats des expérimentations sur les 32 documents du corpus de test

Nous constatons également qu’il est nécessaire d’introduire plus de complexité dans le contexte de l’arbre entre les étiquettes générées par l’extraction linguistique. Pour l’instant, nos règles d’acquisition se limitent aux contraintes sur les fils, les pères ou les frères. Or, le contexte des ascendants est particulièrement important pour la création des attributs des sujets. Prenons par exemple l’étiquette « /num » : si le noeud père est « /ARTICLE », l’attribut sera un numéro d’article alors que si ce même noeud est « /POURVOI », l’attribut sera un numéro de pourvoi. Le contexte des nœuds descendants peut également apporter des précisions par rapport à la création d’un sujet ou d’une association. Dans la **Fig. 6**, l’étiquette « /Noms-de-personnes » renvoie au concept « Personne » dans l’ontologie et qui a deux sous-concepts : « Personnalité Juridique » et « Personnalité Politique ». Une analyse des descendants du noeud « /Noms-de-personnes », et notamment la présence d’un nœud « Juridique » ou « Politique », permet de préciser le sous-concept à instancier.

```
/Noms-de-personnes (M. BOUSCHARAIN , président)
  Nom (M. BOUSCHARAIN)
  role(président)
  /Role/Juridique (président)
```

Fig. 6 – Exemple d’une analyse contextuelle

4 Conclusion et discussion

Cette plateforme propose donc une solution innovante d’enrichissement d’une base de connaissance contrainte par l’ontologie du domaine à partir d’extractions linguistiques grâce à la définition de règles d’acquisition. A notre connaissance, il

n'existe pas d'approche similaire dans le cadre d'applications pour le Web Sémantique. Bien sûr, d'autres systèmes (Kyriakov, 2003) s'intéressent au peuplement d'ontologies grâce à des outils linguistiques mais leurs ontologies sont modélisées en concordance avec les résultats de leurs extractions linguistiques, à un niveau général et sans relation sémantique complexe (n-aire). A contrario, notre approche permet de peupler une ontologie donnée à partir de n'importe quel outil linguistique, du moment que celui-ci extrait sous forme d'arbre conceptuel les informations pertinentes au domaine concerné.

A partir des problèmes soulevés dans la première implémentation, nous avons défini les priorités suivantes : vérification du respect des cardinalités, notamment pour les rôles participant à une association ; amélioration des deux parcours de l'arbre conceptuel pour gérer plus de complexité dans les règles grâce à une contextualisation plus riche ; détection des conflits liés au recouvrement entre les règles d'acquisition . La résolution de ces priorités permettrait d'améliorer rapidement les performances actuelles du système, notamment en ce qui concerne les associations et les rôles. Il reste également le problème de cohérence et de maintenance entre des règles d'acquisition qui deviendront de plus en plus nombreuses, surtout si l'ontologie cliente comporte un nombre important de concepts à instancier. La construction manuelle des règles est assez fastidieuse et susceptible de comporter des erreurs. Et si les ressources linguistiques ou si l'ontologie cliente sont modifiées, alors ces règles doivent être vérifiées et mises à jour par l'administrateur de ces règles.

C'est pourquoi nous proposons de développer un langage formel de description des connaissances mobilisées pour peupler une ontologie à partir d'un arbre conceptuel. Ce langage s'inspirera de LangText (Crispino, 2003), développé pour modéliser les connaissances linguistiques dans le cadre de l'exploration contextuelle (Desclès, 1991), (Minel & al., 2001). L'un des apports de ce langage est de formaliser de manière déclarative la notion d'espace de recherche, d'indicateur et d'annotation d'une unité textuelle. Nous devons néanmoins adapter ce langage au parcours d'un arbre conceptuel et non d'un texte. Ce langage permettra une meilleure maintenance des connaissances, une plus grande efficacité dans la construction des règles d'acquisition grâce à la gestion des conflits potentiels, et un gain de productivité pour l'utilisateur. En effet, à partir de ce langage nous pourrons générer automatiquement les règles XPath associées et utiliser une feuille de style afin de transformer l'arbre conceptuel au format TM de la base de connaissance ou au format OWL de l'ontologie, selon les besoins de l'utilisateur. Ce langage est en cours d'élaboration.

Enfin, il faut souligner que ce système doit rester assez générique afin de pouvoir définir et appliquer les règles d'acquisition à n'importe quel domaine d'application. De plus, les implémentations des modules linguistiques d'extraction et de l'ontologie cliente peuvent rester indépendantes et c'est aux règles d'acquisition de permettre la transformation d'une étiquette linguistique à un concept instancié de l'ontologie.

Références

- AMARDEILH F. & FRANCART T. (2004). A Semantic Web Portal with HLT Capabilities, In *Actes du colloque « Veille Stratégique Scientifique et Technologique »* (VSST2004), Toulouse, (Vol.2), p. 481-492.
- BONTCHEVA K. & CUNNINGHAM H. (2003). The Semantic Web: A New Opportunity and Challenge for Human Language Technology, in *Proceedings of the Second International Semantic Web Conference*, Workshop on Human Language Technology for the Semantic Web and Web Services, Florida, 20-23 October 2003, p. 89-96.
- BOURIGAULT D., AUSSENAC-GILLES N. et CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, *Revue d'Intelligence Artificielle*, 18(4), 24 pp.
- CRISPINO G. (2003). Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO). Application au filtrage sémantique de textes. Thèse sous la direction de Jean-Pierre Desclés, Université Paris-Sorbonne, Décembre 2003, 241 pp.
- DESCLES J.-P., JOUIS C., OH H-G., MAIRE REPPERT D. (1991). Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte, *Knowledge modeling and expertise transfer*, Amsterdam, p. 371-400.
- GRISHAM R. & SUNDHEIM B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, p. 466-471.
- GRIVEL L., GUILLEMIN-LANNE S., LAUTIER C. et al. (2001). La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux, In *Filtrage et résumé automatique de l'information sur les réseaux*, 3^{ème} congrès du Chapitre français de l'ISKO International Society for Knowledge Organization, Paris, 5-6 July 2001, 9 pp.
- GRUBER T. (1995). A Translation approach to portable ontology specifications. In *Knowledge Acquisition*, 5(2), p. 199-220.
- HANDSCHUH S., STAAB S., CIRAVEGNA F. (2002). S-CREAM – Semi-automatic CREAtion of Metadata, In *Proceedings of the 13th International Conference on Knowledge Engineering and Management* (EKAW 2002), Espagne, 1-4 Octobre 2002, Springer Verlag, p. 358-372.
- KAHAN J., KOIVUNEN M., PRUD'HOMMEAUX E. et al. (2001). Annotea: An Open RDF Infrastructure for Shared Web Annotations, In *Proceedings of the WWW10 International Conference*, Hong Kong, Mai 2001, p. 623-632.
- KIRYAKOV A., POPOV B., OGNYANOFF D., MANOV D., KIRILOV A., GORANOV M. (2003). Semantic Annotation, Indexing, and Retrieval, In *Proceedings of the 2nd International Semantic Web Conference* (ISWC2003), Florida, 20-23 Octobre 2003, p. 484-499.
- LAUBLET P., REYNAUD C. et CHARLET J. (2002). Sur quelques aspects du Web Sémantique, In *Assises du GDR I3*, Eds Cépadues, Nancy, 20 pp.
- MINEL J.-L., J-P. DESCLES, E. CARTIER, G. CRISPINO, S. BEN HAZEZ, A. JACKIEWICZ. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText, *Revue Technique et Science informatiques*, 20(3), Hermès, p. 369-395.
- PARK J. & HUNTING S. (2003). XTM Topic Maps : Creating and using Topic Maps for the Web, Addison Wesley Eds, Boston, p. 81-101.
- VARGAS-VERA M., MOTTA E., DOMINGUE J. (2002). MnM : Ontology Driven Tool for Semantic Markup, In *Proceedings of the Workshop Semantic Authoring, Annotation & Knowledge Markup* (SAAKM 2002), Lyon (France), 22-23 Juillet 2002, p. 43-47.

Construction d'ontologies médicales à partir de textes : propositions méthodologiques

Audrey Baneyx¹, Jean Charlet^{1,2}

¹ INSERM U729 - Laboratoire SPIM

Faculté de Médecine Broussais-Hôtel-Dieu

15 rue de l'Ecole de médecine, 75006 Paris, France

² STIM - DS/ AP-HP

{Audrey.Baneyx,Jean.Charlet}@spim.jussieu.fr

Résumé : Dans le contexte du codage des activités médicales, il est nécessaire de construire des représentations conceptuelles des connaissances. Cet article apporte des propositions méthodologiques sur la construction d'ontologies médicales, à partir de textes, à l'adresse d'un ingénieur cogniticien. Cette méthodologie est fondée sur la mise en œuvre des principes de la sémantique différentielle et utilise les outils de traitement automatique de la langue. Notre principale hypothèse de recherche concerne l'utilisation conjointe de deux méthodes : une méthode éprouvée qui consiste à construire des ressources termino-ontologiques par analyse distributionnelle et une méthode fondée sur la recherche de relations sémantiques par l'utilisation de patrons lexico-syntactiques.

Mots-clés : Ontologie, ingénierie des connaissances, extraction terminologique, sémantique différentielle, analyse distributionnelle, patrons lexico-syntactiques, pneumologie.

1 Introduction

La réduction des inégalités de ressources entre les établissements de santé figure dans la réforme de l'hospitalisation (ordonnance du 24/04/96). Afin de mesurer l'activité et les ressources des établissements, le gouvernement souhaite disposer d'informations quantifiées et standardisées. Ces informations sont recueillies pour chaque séjour d'un patient sous la forme de résumés standardisés de sortie dans lesquels la codification des diagnostics principaux et secondaires est effectuée à partir de la classification internationale des maladies CIM-10. La procédure de codification appelée Programme de Médicalisation des Systèmes d'Information¹, couramment « codage PMSI », est le plus souvent réalisée ma-

¹<http://www.atih.sante.fr/>

nuellement par les praticiens qui s'aident d'un thésaurus de spécialité (Bensadoun, 2001). Ces thésaurus proposés par les sociétés savantes sont construits pour permettre aux médecins de coder à partir de leur terminologie usuelle mais il est aujourd'hui manifeste que les outils d'aide au codage fondés sur ces thésaurus sont inadaptés aux besoins du praticien (Friedman *et al.*, 2004). En effet, les libellés de ces thésaurus se révèlent ambigus (par exemple, à un même code sont associées plusieurs pathologies) et non exhaustifs, le mode de classification choisi est difficile d'utilisation, et le maintien de la consistance ainsi que de la cohérence du thésaurus est impossible. On constate que le sens des libellés de ces nomenclatures médicales (SNOMED, CIM-10,...) repose sur les facultés d'interprétation du lecteur humain et qu'elles ne sont donc pas adaptées à une exploitation par l'ordinateur. Dans ce contexte, il nous semble indispensable de décrire la sémantique et l'organisation des objets du domaine médical afin de se doter de modélisations conceptuelles (non contextuelles et non ambiguës) dont le sens est inscrit dans la structure même du modèle. Une telle modélisation est appelée ontologie (Staab & Studer, 2003). Notre hypothèse est que le développement de telles ressources ontologiques permettra de mettre au point des outils performants, fiables et maintenables pour l'aide au codage. La problématique à laquelle nous nous attaquons est donc la construction de ces ontologies et le terrain d'expérimentation dans lequel nous nous situons est l'aide au codage en pneumologie.

Le terme « ontologie » est utilisé depuis le début des années 90 dans les domaines de l'intelligence artificielle, en particulier de l'ingénierie des connaissances et de la représentation des connaissances. Son champ d'application s'élargit considérablement et il fait désormais partie des objets de recherche courants, notamment dans le secteur de la modélisation des systèmes d'information (Gomez-Pérez *et al.*, 2004). Une ontologie est un système formel dont l'objectif est de représenter les connaissances d'un domaine spécifique au moyen d'éléments de base, les concepts, définis et organisés les uns par rapport aux autres (Rector, 1998). La représentation ontologique des connaissances garantit le maintien de la cohérence des axiomes et de l'intégrité du système, ainsi que l'extensibilité de la représentation sans modification de la structure. Il est cependant difficile de repérer et de classifier les objets d'un domaine. Les critères de classification dépendent des buts poursuivis et n'ont rien d'immuable (Charlet, 2002). Ainsi, nous ne prétendons en aucun cas construire une ontologie universelle de la médecine mais bien une ontologie régionale de la pneumologie (Bachimont, 2000).

Nous précisons nos objectifs section 2. La section 3 présente le matériel utilisé et la section 4 détaille les différentes étapes de notre méthodologie. Nous donnons les résultats obtenus dans la section 5 et concluons cet article, section 6, en discutant de l'intérêt d'un tel travail et des perspectives qu'il offre.

2 Objectifs

Ce travail de construction d'ontologies s'inscrit dans le cadre, plus large, du projet de recherche PERTOMed², financé par le CNRS. Le but de ce projet est de développer une infrastructure proposant un ensemble de méthodes et d'outils opérationnels pour la production et l'utilisation de ressources terminologiques ou ontologiques dans le domaine médical. Les ontologies médicales sont construites en étroite collaboration avec des groupes d'utilisateurs, avec lesquels seront, en particulier, mises en place des procédures d'évaluation de ces ressources dans leur contexte d'usage. Au sein du projet, notre travail consiste à proposer aux médecins pneumologues un environnement d'aide au codage et à la représentation des connaissances médicales reposant sur le modèle conceptuel d'une ontologie du domaine concerné. Nous travaillons en étroite collaboration avec le service de Pneumologie et de Réanimation du groupe hospitalier de la Pitié-Salpêtrière de Paris³ et avec la Société de pneumologie de langue française⁴.

Nous construisons notre ontologie régionale à partir de ressources textuelles sur lesquelles nous appliquons des techniques appartenant au domaine du traitement automatique du langage dans le but de développer les corpus nécessaires à la structuration de l'ontologie. La méthodologie que nous employons a été mise au point au sein du groupe TIA sur, entre autre, les principes de sémantique différentielle (Bachimont, 2000). Notre principale hypothèse de recherche concerne l'utilisation en parallèle de deux méthodes pour enrichir le travail de construction de l'ontologie : *a)* une méthode éprouvée qui consiste à construire des ressources termino-ontologiques par analyse distributionnelle (Bourigault & Lame, 2002), et *b)* une méthode fondée sur la définition *a priori* d'une relation sémantique, puis sur l'observation de séquences en corpus qui véhiculent la relation souhaitée (Séguela, 2001). Sachant qu'aucune ontologie ne couvre le domaine de la pneumologie, notre objectif est double : il s'agit, d'une part, de construire l'ontologie de la pneumologie et, d'autre part, d'apporter des précisions sur les deux premières étapes de la méthodologie. Nous proposons notre propre expérimentation de construction d'ontologies médicales dans la même optique que le travail de Le Moigno *et al.* (2002). Un point de vue quelque peu différent est adopté ici puisque l'ontologie est construite par un ingénieur cogniticien et non par un expert du domaine médical comme dans ce précédent travail. L'intérêt consiste à mettre au point un processus méthodologique précis destiné à l'ingénieur cogniticien de manière à ne faire appel à l'expert médical que pour des moments précis de validation.

Nous souhaitons, dans ce papier, montrer précisément le déroulement des deux premières étapes de la méthodologie de construction d'une ontologie différentielle de la pneumologie et l'apport conjoint de l'analyse distributionnelle et des patrons lexico-syntactiques pour la mise en oeuvre des principes différentiels.

²Production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine, <http://www.spim.jussieu.fr/Pertomed>

³UPRES EA2397

⁴<http://www.splf.org/>

3 Matériel : Corpus et outils

Dans le but de couvrir, avec le plus d'haustivité possible, l'ensemble de l'activité de la pneumologie, nous avons collecté des comptes rendus d'hospitalisation (corpus intitulé [CRH]) dans six hôpitaux de l'Assistance Publique-Hôpitaux de Paris⁵. Au total, nous disposons de 1 038 CRH. Ce premier corpus [CRH] compte environ 417 000 mots. Sachant qu'il a été établi dans (Le Moigno *et al.*, 2002) que 350 000 mots est un minimum pour obtenir de bons résultats avec nos outils, le corpus [CRH] semble être une bonne base d'expérimentation. Le second corpus, intitulé [LIVRE], est construit d'après un ouvrage pédagogique et correspond environ à 823 000 mots.

Nous utilisons le logiciel SYNTTEX-UPERY comme outils d'analyse de traitement du langage. SYNTTEX est un module d'analyse syntaxique fondée sur l'hypothèse que les mots qui ont un sens proche se caractérisent par des dépendances similaires (Bourigault & Lame, 2002). Ainsi, ce module permet d'obtenir des relations de dépendances syntaxiques entre mots ou syntagmes (noms *vs* syntagmes nominaux, verbes *vs* syntagmes verbaux et adjectifs *vs* syntagmes adj ectivaux). A la fin du traitement nous obtenons un réseau de dépendances syntaxiques – ou réseau terminologique – dont les éléments sont les candidats termes qui vont nous servir pour construire l'ontologie. Le module UPERY met ensuite en œuvre le principe de l'analyse distributionnelle « à la Harris » (Harris, 1968) : il calcule des proximités distributionnelles entre les candidats termes du réseau sur la base des contextes syntaxiques partagés. Nous obtenons un réseau de candidats termes, leurs proximités contextuelles et leurs liens avec le corpus source. Les résultats de l'analyse sont visualisables dans TERMONTO, l'interface d'accès et de traitement des données du logiciel. L'éditeur DOE⁶ permet de construire notre ontologie selon la sémantique différentielle. Ce logiciel permet également de compléter l'ontologie en ajoutant à chaque concept sa traduction en anglais ainsi qu'une définition encyclopédique. Il en va de même pour les relations. L'ontologie est exportée en OWL, un langage de représentation des connaissances préconisé par le consortium W3C.

4 Méthode

La méthodologie mise en œuvre permet de décrire les variations des sens des termes considérés en contexte. C'est pourquoi cette méthode considère que le corpus textuel est la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur est associé. Nous distinguons quatre étapes : 1) la constitution du corpus des connaissances et son analyse par des outils de traitement automatique du langage, 2) la normalisation sémantique des termes du domaine grâce à la mise en œuvre des

⁵Ils se répartissent comme suit : Crêteil : 326 CRH, Hôtel-Dieu : 97 CRH, Kremlin-Bicêtre : 125 CRH, Pitié-Salpêtrière : 57 CRH, Saint Antoine : 372 CRH, Tenon : 61 CRH.

⁶The Differential Ontology Editor, <http://opales.ina.fr/public>

principes différentiels, 3) l'engagement ontologique qui permet de formaliser les concepts, 4) l'opérationnalisation de l'ontologie dans un langage de représentation des connaissances interprétable par l'ordinateur (Bachimont *et al.*, 2002). Notre expérimentation dans le domaine de la pneumologie nous permet d'adapter et de préciser, pour l'ingénieur cogniticien, les deux premières étapes de la méthodologie, ce qui contribue à la réalisation de notre second objectif.

4.1 Traitement des ressources de base

Les deux corpus [CRH] et [LIVRE] nous parviennent sous des formats inexploitables par les outils d'analyse du langage. Ils sont donc prétraités⁷ puis mis sous un format semi structuré par des programmes que nous avons développés. Nous disposons ainsi d'un corpus [CRH] anonyme et d'un corpus [LIVRE] didactique, tous deux au format XML. Le corpus [CRH] est ensuite traité par le logiciel SYNTTEX-UPERY. Le résultat de l'analyse distributionnelle nous permet de construire les éléments de base - *i.e.* primitives - de l'ontologie. Le second corpus [LIVRE] est analysé par le biais de patrons lexico-syntaxiques prédéfinis qui permettent d'extraire des couples d'unités lexicales correspondant au motif de la relation sémantique recherchée (hypéronymie, synonymie...). Les résultats obtenus nous aident à contrôler et enrichir la hiérarchie de l'ontologie.

4.2 Choix des candidats termes

Les résultats fournis sur la base du corpus [CRH] servent de support dans le choix de candidats termes⁸ (CT) représentatifs de la pneumologie en tant qu'activité médicale. Nous distinguons deux étapes dans leur sélection. Le travail est manuel et s'appuie sur les fonctionnalités de TERMonto.

1) Nous parcourons l'ensemble des résultats fournis par l'analyse syntaxique et choisissons d'étudier, en premier lieu, les syntagmes nominaux (SN) dont la fréquence d'apparition en corpus est supérieure à 12 (2 % du corpus). Nous repérons les grands axes conceptuels typiques du corpus et donc du domaine représenté. A chaque CT, nous associons un critère de validité (ainsi nommé dans TERMonto), compris dans un intervalle allant de 1 à 6, correspondant à l'un de ces axes : 1 (CT non pertinent appartenant à l'axe Autres), 2 (réservé aux CT déjà modellisés dans l'ontologie), 3 (CT appartenant à l'axe Symptômes), 4 (CT appartenant à l'axe Pathologies), 5 (CT appartenant à l'axe Signes) et 6 (CT appartenant à l'axe Traitements/Examens). Par exemple, nous fixons à 6 le critère de validité pour tous les CT de ce dernier axe - *e.g. examen, doppler, radiographie, etc.* Au début de la méthodologie, tous les CT ont un critère de validité égale à 1 et à la fin égale à 2 car ils sont, en principe, tous définis dans

⁷ Les fichiers ont été convertis au format texte, « nettoyés », anonymisés, segmentés, associés à des identifiants de section et de phrase, étiquetés et analysés morphosyntaxiquement par Cordial Analyseur de la société Synapse.

⁸ Un candidat terme est un syntagme nominal composé d'une tête et d'une expansion. Par exemple, dans le SN *Opacité dans le poumon gauche*, le terme *Opacité* est la tête du syntagme et *dans le poumon gauche* est son expansion.

Descendants en tête	Descendants en expansion	Voisins en tête	Voisins en expansion
Épanchement pleural droit	Lame d'épanchement pleural	Lésions	Liquide
Épanchement pleural liquidiens	Récidive d'épanchement pleural	Infection	Infiltrets
Épanchement pleural de la grande cavité	Lier la dyspnée à l'épanchement pleural	Signes	Décompensation

TAB. 1 – Exemple de résultats du rapprochement contextuel pour le SN *Epanchement pleural*.

l'ontologie. Les critères de validité 3, 4, 5 et 6 sont utilisés temporairement durant la phase de construction. Ces regroupements permettent une première phase de travail sur les rapprochements des CT par contexte. La sélection par critère de validité laisse 35 % des CT sur lesquels élaborer le cœur de notre ontologie.

2) L'analyse distributionnelle rapproche deux à deux les termes partageant les mêmes contextes (descendants en tête et en expansion). Comme cette analyse est symétrique, elle rapproche également les contextes en fonction des termes qu'ils partagent (voisins en tête et en expansion). Sur le tableau 1 *épanchement* est la tête du SN *épanchement pleural* et *pleural* est son expansion. Les descendants en tête donnent des informations sur ce qui pourrait être des concepts fils ou des concepts définis. Les descendants en expansion donnent des informations sur la place du concept dans la hiérarchie, sur le concept père. Les voisins en tête et en expansion nous permettent de constituer des regroupements de candidats termes sémantiquement proches du candidat terme étudié ici, *épanchement pleural*. Ces regroupements sont d'une grande aide pour élaborer la structure hiérarchique de l'ontologie, aussi bien l'axe horizontal que vertical. L'exemple ci-dessous montre un premier rapprochement possible : nous pouvons mettre en rapport le groupe A *{épanchement, lésion, infection, décompensation}* avec *{signes}*. Les CT du groupe A partagent un même contexte sémantique, la première hypothèse est donc qu'il peut s'agir de concepts frères dont *signes* est possiblement le concept père.

4.3 Mise en œuvre des principes différentiels

Pour élaborer cette hiérarchie, il convient d'articuler les CT choisis dans la précédente étape (*cf.* section 4.2) en précisant les principes différents qui les définissent. Par exemple, le concept *Ultrasonographie* et le concept *ExamenIsotopique* sont des concepts frères dont le concept père est *ImagerieParRayonnement* (*cf.* figure 1). Le principe de communauté avec le concept père est la projection ou l'injection d'une substance artificielle dans le but d'effectuer des mesures. Le principe de communauté entre les concepts frères est lié au média d'injection. Le principe différentiel entre les concepts frères est relatif au type de media artificiel

mis en œuvre : un isotope dans le cas de l'*examen isotopique* (la scintigraphie est un exemple d'examen isotopique) et les ultrasons pour l'*ultrasonographie*. Les candidats termes des 4 axes conceptuels (3, 4, 5 et 6) sont définis selon ces principes.

Les résultats de l'analyse par patrons lexico-syntaxiques sur le corpus [LIVRE] nous aident à définir les principes différentiels. Les patrons lexico-syntaxiques représentent des motifs de relations sémantiques spécifiques. Ils sont construits autour d'un marqueur, également appelé pivot, qui est l'indice d'une relation lexicale, comme le marqueur *entre autres* pour la relation d'hyperonymie. Ainsi, un patron de la forme *DET SN, entre autres SN* permet d'extraire l'unité lexicale *Les méningites, entre autres pathologies ...* et de mettre en relation d'hyperonymie *méningite* et *pathologie*. Cette méthode a été présentée dans (Hearst, 1992) et expérimentée dans plusieurs travaux, notamment dans (Caraballo, 1999). Les patrons lexico-syntaxiques liés à l'hyperonymie mettent en relation des couples père-fils potentiels intéressants pour contrôler et enrichir la structure hiérarchique de l'ontologie. Dans le cadre de la construction d'ontologies différentes, nous appliquons cette méthode à la recherche d'énoncés définitoires en corpus pour aider à renseigner les principes différents. Les patrons que nous employons ont été développés par Malaisé *et al.* (2004). Le corpus [LIVRE], d'un genre pédagogique, à la particularité d'être très structuré et se révèle particulièrement propice à ce type de recherche. Par exemple, les patrons utilisant le marqueur *Il s'agit de* ont pu associer un titre de section (par exemple « Asthme ») à sa description dans la première ligne du texte (« Il s'agit d'une maladie inflammatoire des voies aériennes »). Les unités lexicales extraites sont validées manuellement et les hiérarchies créées sont visualisables sous DOE. Il est alors facile de comparer les deux structures terminologiques obtenues : la hiérarchie issue de l'analyse distributionnelle du corpus [CRH] et celle issue du repérage par patrons lexico-syntaxiques sur le corpus [LIVRE], et d'en tirer les informations nécessaires à l'enrichissement et au raffinement de l'ontologie. La comparaison des structures obtenues est détaillée dans (Baneyx *et al.*, 2005).

Enfin, il paraît important de souligner que l'ingénieur cogniticien doit, tout au long du travail de construction, décider si le concept doit être primitif ou défini, c'est-à-dire s'il est essentiel ou non par rapport aux buts poursuivis. Un concept défini est construit à partir d'un ou plusieurs concepts primitifs et d'une ou plusieurs relations. Le SN *douleur thoracique gauche* est modélisé par le concept défini suivant : *[Douleur](au niveau du)[thorax](localisée à)[gauche]*⁹. A la fin de cette étape, nous avons normalisé le sens des termes du domaine et représenté la hiérarchie des concepts primitifs et des relations avec DOE.

⁹Le crochet indique un concept primitif et la parenthèse indique une relation. Dans notre méthodologie, les relations sont traitées de la même manière que les concepts.

5 Résultats

Après l'utilisation de SYNTTEX, le corpus [CRH] donne 36 881 SN. D'après les résultats de l'analyse syntaxique et de l'analyse distributionnelle, le SN *chimiothérapie de <nom>* a le plus grand nombre de voisins en tête , soit 28, et sa fréquence d'apparition dans le corpus s'élève à 190, le SN *<nom> de chimiothérapie* a le plus grand nombre de voisins en expansion, soit 52, et sa fréquence d'apparition est également la plus haute, soit 454. Nous pouvons vérifier la pertinence des rapprochements par groupe de candidats termes dont les contextes d'apparition sont sémantiquement proches. Pour *cure de chimiothérapie* par exemple : [*Hospitalisation, Examen, Navelbine, Cisplatine, Doxorubicine, Taxotere, Carboplatine, MIP*] sont ses voisins en tête et [*Traitemet, Bilan, Antibiothérapie, Injection, Radiothérapie*] sont ses voisins en expansion. Ces résultats sont examinés et mis sous une forme ontologique visualisable avec DOE. Nous construisons la hiérarchie suivante : *ActionMedicale/Traitemet/TraitemetMedicamenteux/Chimiothérapie*. La chimiothérapie étant considérée comme un traitement médicamenteux, nous retrouvons les principes médicamenteux suivant classés sous *Medicament/PrincipesMedicamenteux/Navelbine, Cisplatine, Doxorubicine...* Les candidats termes *Antibiothérapie* et *Radiothérapie* sont également placés sous *Traitemet*. Cette méthode de regroupement semble donner de bons résultats pour construire l'ontologie et rend la tâche bien plus facile pour un ingénieur cognitien non spécialiste du domaine médical modélisé.

Le repérage de définition par patrons lexico-syntactiques sur le corpus [LIVRE] permet d'extraire 799 unités lexicales. Nous en validons 119, ce qui représente près de 15 % des extractions. La comparaison des deux structures terminologiques obtenues (*cf. tableau 2*) permet de distinguer s'il s'agit de hiérarchies complémentaires, identiques ou comparables, ou bien divergentes. Ces informations permettent de préciser et de corriger la structure.

Notre ontologie (*cf. figure 1*) contient à ce jour 500 concepts primitifs, issus de la première analyse des candidats termes. Les étapes de construction 1 et 2 étant itératives, nous augmentons très rapidement la représentation en examinant les candidats termes dont la fréquence d'apparition dans le corpus [CRH] est inférieure à 12 et dont le critère de validité vaut 1.

La question de la validation en ingénierie ontologique n'est pas résolue. Bien que certaines pistes semblent se faire jour, il n'existe pas de méthode unanime pour évaluer ce type de travail. Nous construisons une ontologie de la pneumologie dans le but de servir de support à un outil d'aide au codage, aussi il nous paraît important d'axer la question de la validation sur l'aide que nous serons en mesure d'apporter aux pneumologues, c'est-à-dire sur l'usage qu'ils feront de l'outil. Cependant, la tache n'étant pas terminée, nous validons des étapes intermédiaires, en terme de qualité et de complétude. Dès à présent, l'avancement de la hiérarchie conceptuelle est corrigé et validé périodiquement par des médecins pneumologues de la Société de pneumologie de langue fran-

Type	Exemple	Commentaire
Identique ou comparable	<i>Broncho pneumopathie/Asthme</i> [CRH] vs <i>Bronchopathie/Asthme à dyspnée continue</i> [LIVRE]	Les deux premières hiérarchies sont classifiées dans le MeSH sous <i>Poumon, maladie</i> , alors qu' <i>Asthme</i> est une notion plus spécifique dans la même branche hiérarchique. Cela valide la cohérence et la compatibilité des deux hiérarchies.
Complémentaire	<i>Signe/[...]/Signe Respiratoire/Insuffisance-Ventriculaire</i> [CRH] vs <i>Signe/Insuffisance ventriculaire droite</i> [LIVRE]	La deuxième arborescence vient confirmer la première et permet de compléter d'un niveau, celui d' <i>Insuffisance ventriculaire droite</i> .
Divergente	<i>EtatPathologique/Maladie-Respiratoire/Bronchite</i> ET <i>Signe/Toux</i> [CRH] vs <i>Toux avec expectoration/Bronchite chronique</i> [LIVRE]	Dans le MeSH, la toux est classée à la fois comme <i>signe-symptôme</i> et comme <i>pathologie</i> : les deux sources textuelles illustrent chacune un de ces aspects.

TAB. 2 – Comparaison des deux structures terminologiques.

çaise avec laquelle nous collaborons. Le soin que nous apportons à la définition des principes différentiels et leur nature normalisatrice sont un gage de qualité et d'évolutivité pour notre ontologie. Ces principes rigoureusement appliqués assurent la cohérence et la robustesse de notre modélisation. Ainsi, la place de chaque concept étant bien définie dans la hiérarchie, il est plus facile de calculer la position de nouveaux concepts venant enrichir l'ontologie. Dans un deuxième temps, l'évaluation se fera également en testant la couverture de l'ontologie par rapport au thésaurus de spécialité. Cette validation garantit, autant que faire se peut, que l'ontologie développée sera adéquate à un outil d'aide au codage fondé sur le thésaurus de spécialité. Pour estimer cette couverture, nous allons vérifier la possibilité de construire une représentation conceptuelle des connaissances médicales en combinant les concepts primitifs et les relations dont nous disposerons dans l'ontologie. Pour l'instant, nous pouvons construire quelques concepts définis comme par exemple celui de *Chimiothérapie intra-pleurale* qui se trouve dans le chapitre « Plèvre » du thésaurus : *[ActionMedicale/Traitement/TraitementMedicamenteux/Chimiotherapie]/(RelationActe/ModeRealisation/RealiseAuNiveau/intra)[Anatomie/AppareilRespiratoire/Pleure]*.

6 Discussion et conclusion

Nous avons présenté un ensemble de principes méthodologiques pour la construction, à partir de textes, d'ontologies médicales différentes. Nous sommes en train d'améliorer l'ontologie de la pneumologie en analysant les termes présents dans les thésaurus de spécialité CIM-10 et CCAM. Cela étoffe les branches et

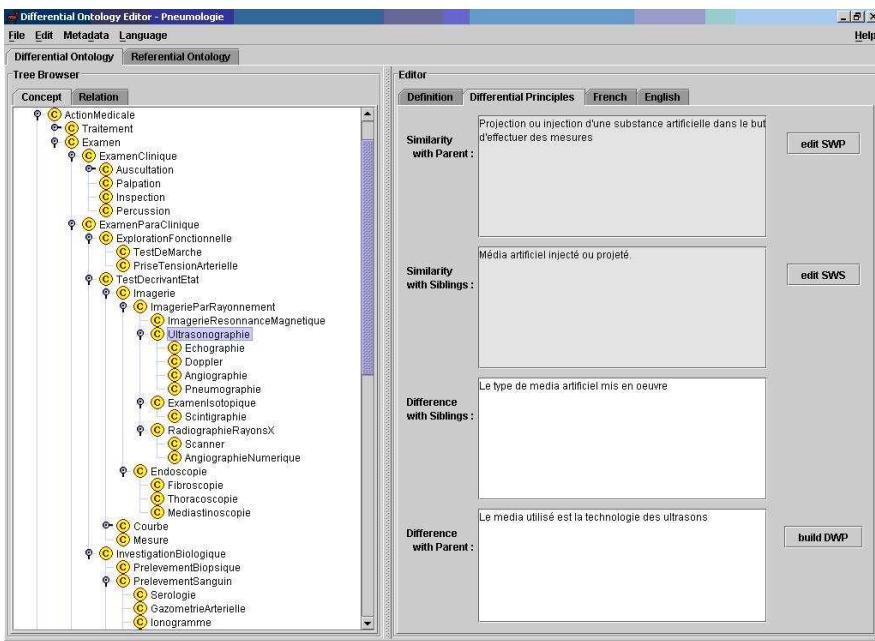


FIG. 1 – Un extrait de l'ontologie de la Pneumologie visualisable sous DOE.

augmente le nombre des feuilles de notre hiérarchie. De même, nous prévoyons de compléter l'ontologie en faisant le lien avec le haut de l'ontologie du projet Ménélas¹⁰. Ce travail nous permettra de vérifier s'il existe un haut niveau conceptuel commun au domaine médical, sachant que notre ontologie est régionale, c'est-à-dire spécifique au domaine de la pneumologie. Nous envisageons d'obtenir rapidement une ontologie de la pneumologie comptant environ 1 200 concepts. D'autre part, un des enjeux de ce travail est de montrer que la méthodologie mise au point permet à un ingénieur cogniticien, non spécialiste du domaine modélisé, de construire une ontologie à partir de textes à l'aide d'outils de traitement automatique des langues. Un travail récent dans le domaine de la réanimation chirurgicale (Le Moigno *et al.*, 2002) ainsi que les premiers résultats de notre recherche permettent de penser que nous allons dans la bonne direction.

Nous avons, en outre, présenté l'utilisation conjointe de deux méthodes adaptées chacune à un genre de corpus particulier. En prenant les textes comme source de connaissances, nous reposons la question des genres textuels développée dans (Aussenac-Gilles & Condamin, 2003). L'utilisation des corpus en Ingénierie des connaissances se veut une réponse au problème de l'accès aux connaissances d'un domaine, pour un objectif particulier, lié à une application informatique, dans notre cas une ontologie. *A priori*, les corpus textuels, comme ceux que nous

¹⁰ <http://www.biomath.jussieu.fr/Menelas/Ontologie>

utilisons, témoignent d'un vocabulaire métier, fixé par l'écrit, consensuel car diffusé et partagé à l'intérieur du corps médical. Cela offre une garantie de fiabilité et de stabilité à notre modélisation. Dans le domaine particulier du traitement automatique de la langue appliquée au domaine médical, Pierre Zweigenbaum propose cinq catégories de mots-clés visant à caractériser les genres textuels (*Op. Cit.*) : dossier patient, enseignement, ressources, publications et oral. Dans ce cadre, notre corpus [CRH] fait partie du genre textuel dit « dossier patient » et notre corpus [LIVRE] de la catégorie « enseignement ». Nous avons également montré qu'il existe une relative compatibilité entre les deux hiérarchies terminologiques obtenues par l'utilisation de l'analyse distributionnelle puis la mise en œuvre des principes différentiels et par l'emploi des patrons lexico-syntactiques. La complémentarité de ces structures est intéressante car elle résulte de l'emploi de méthodes différentes appliquées à des corpus de genres textuels également différents. L'analyse de la divergence des résultats est riche en information et apporte un point de vue critique à l'ingénieur cogniticien et à l'expert sur la manière de modéliser le domaine. De plus, cette expérimentation montre qu'il existe des organisations conceptuelles différentes au sein d'un même domaine, ce qui appuie le fait qu'il s'agit bien de construire des ontologies régionales et non pas universelles car toute modélisation n'est jamais qu'un point de vue sur le monde. Soulignons, qu'il n'est bien évidemment pas question ici de remplacer l'expert du domaine qui intervient à plusieurs moments clés de l'élaboration de l'ontologie : consultations préalables pour cerner aux mieux les besoins de la communauté, phases périodiques de validation des résultats, analyse des divergences, phases de test...

Nous avons également cerné certaines des limites liées à la comparaison de ces deux méthodes : un rapprochement semi-automatisé des hiérarchies nécessiterait, d'une part, la mise en œuvre de techniques plus sophistiquées d'appariement et, d'autre part, d'améliorer la précision des patrons lexico-syntactiques et des marqueurs de relation définitoire en les adaptant spécifiquement au domaine médical. Ainsi, des marqueurs comme *indiquer* et *définir* sont à spécifier plus finement : *indiqué* est souvent utilisé dans le cadre de « traitements *indiqués* pour soigner une pathologie », *définir* intervient surtout dans des phrases telle que « Ces résultats ont permis d'acquérir certaines connaissances et de *définir* les meilleurs traitements pour soigner les 30 patientes ».

Ce travail de modélisation des connaissances à partir de textes nous a également permis de mesurer la nécessité d'utiliser conjointement des outils de traitement automatique du langage (SYNTEX-UPERY) et de modélisation (DOE). Il semblerait intéressant d'intégrer ces deux outils pour faciliter le passage des candidats termes à la représentation des concepts, tout en assurant de pouvoir revenir aux textes (Szulman & Biébow, 2004).

Pour conclure, nous précisons que l'étape finale de la validation de l'ontologie se fera par l'usage et nous tenterons de quantifier et de qualifier l'aide que notre travail apporte aux pneumologues. La méthodologie de construction d'ontologies différentielles utilisée est constructive, elle permet de placer de manière précise chaque concept dans la structure hiérarchique. L'ontologie doit être mise à dis-

position du corps médical au travers d'un environnement d'aide au codage des actes et des diagnostics et à la représentation des connaissances médicales, dans le cadre de la plateforme terminologique du projet PERTOMed.

Références

- AUSSENAC-GILLES N. & CONDAMINES A. (2003). *Rapport de l'action spécifique ASTICCOT*. Rapport interne IRIT/2003-23-R, CNRS. Rapport de l'action spécifique ASTICCOT, « Terminologie et corpus » rattachée au RTP-DOC (RTP-33) du CNRS.
- BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, chapter 19. Paris : Eyrolles.
- BACHIMONT B., ISAAC A. & TRONCY R. (2002). Semantic commitment for designing ontologies : A proposal. In *Proceedings of EKAW*, p. 114–121, Sigüenza, Espagne : Springer.
- BANEYX A., MALAISÉ V., CHARLET J., ZWEIGENBAUM P. & BACHIMONT B. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntactiques pour la construction d'ontologies différentielles. In *Actes de la conférence Terminologie et Intelligence artificielle*, p. 31–42, Rouen, France.
- BENSADOUN H. (2001). Pmsi et chirurgiens : pourquoi les chirurgiens doivent-ils coder, comment bien coder? *Journal de Chirurgie, Masson*, **138**(1).
- BOURIGAULT D. & LAME G. (2002). Analyse distributionnelle et structuration de terminologie, application à la construction d'une ontologie documentaire du droit. *Traitemet automatique des langues*, **43**(1).
- CARABALLO S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Meeting of the Association for Computational Linguistics (ACL'99)*, p. 120–126, Maryland, USA.
- CHARLET J. (2002). *L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Habilitation à diriger des recherches, Université Paris 6.
- FRIEDMAN C., SHANIGA L., LUSSIER Y. & HRIPSACK G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, **11**, 392–402.
- GOMEZ-PÉREZ A., FERNANDEZ-LOPEZ M. & CORCHO O. (2004). Ontological engineering. In *Advanced Information and Knowledge Processing*. Madrid, Spain : Springer.
- HARRIS Z. (1968). *Mathematical Structures of Language*. New-York, USA : John Wiley and Sons.
- HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In A. ZAMPOLLI, Ed., *Proceedings of the 14th COLING*, p. 539–545, Nantes, France.
- LE MOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.-C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. In B. BACHIMONT, Ed., *Actes des 6^{es} Journées Ingénierie des Connaissances*, p. 229–38, Rouen, France.
- MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitemet automatique des langues naturelles)*, p. 269–278, Fès, Maroc : ATALA LPL.
- RECTOR A. (1998). Thesauri and formal classifications : Terminologies for people and machines. *Methods of Information in Medicine*, **37**(4–5), 501–509.
- SÉGUÉLA P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat, Université Toulouse III.
- STAAB S. & STUDER R. (2003). *Handbook on Ontologies*. Berlin, Germany : Springer, 1 edition.
- SZULMAN S. & BIÉBOW B. (2004). OWL et Terminae. In *14^e journées francophones d'Ingénierie des Connaissances (IC 2004)*.

De l'Ingénierie des Connaissances à la Gestion des Compétences

Giuseppe Berio¹, Mounira Harzallah²

¹ Dipartimento di Informatica, Università di Torino,
C.so Svizzera 185, 10149 Torino, Italie

berio@di.unito.it

² LINA, Rue Christian Pauc -La Chantrerie
BP50609-44306 Nantes CEDEX03

mounira.harzallah@iut-nantes.univ-nantes.fr

Résumé. La gestion des compétences porte sur un nombre important de connaissances de l'entreprise et de ses individus. Nous avons proposé un modèle des compétences (CRAI), comme une base pour le développement d'un système d'information pour les compétences. Toutefois, les techniques classiques de l'ingénierie des systèmes d'information ne sont pas suffisantes pour assurer une gestion efficace des compétences : la gestion des compétences comprend plusieurs processus complexes et lourds à mener. Il nous semble intéressant d'investiguer les techniques d'ingénierie des connaissances. Cet article synthétise des travaux de recherche dans la littérature, portant sur les méthodes d'ingénierie des connaissances appliquées à la gestion des compétences, classés selon leur apport aux différents processus de gestion des compétences.

Mots-clés : Ingénierie des connaissances, Gestion des Compétences, Gestion de Connaissance, Ontologies, Modèle CRAI.

1. Introduction

La gestion des compétences est la façon dont l'entreprise organise et gère ses compétences organisationnelles, des groupes et des individus. Son premier objectif est d'identifier et d'entretenir en continu son capital humain. D'autres objectifs concernent l'exploitation des compétences pour assurer la flexibilité de l'entreprise devant des changements du marché, des technologies, du contexte social, etc.

Dans nos travaux précédents, nous avons considéré les compétences individuelles (à savoir, les compétences d'une personne ou d'un employé, distinguées, mais en relation avec, des compétences collectives et des compétences stratégiques). Nous nous sommes basés sur les travaux des gestionnaires (Levy-leboyer, 1996), (Le Boterf, 2000), (Lucia & Lepsinger, 1999), (Pfeffer & Sutton, 2000), (Michel, 1997) pour proposer une base pour le développement d'un système d'information pour leur gestion. Les résultats de nos travaux sont organisés en trois axes (Harzallah & Vernadat, 2002), (Harzallah & Berio, 2004) :

- le modèle CRAI (Competency, Resource, Aspect, Individual) qui offre une représentation formelle des compétences individuelles soit acquises soit requises ;
- une liste de directives pour déployer le modèle CRAI à une organisation spécifique pour concevoir son *système d'information* des compétences et le maintenir ;
- Un ensemble de requêtes pour étudier l'adéquation des compétences requises et acquises.

Le modèle CRAI est basé sur quatre relations principales : une compétence concerne un ou plusieurs aspects de l'entreprise étudiée, une compétence est un ensemble de *c-ressources* (savoir, savoir-faire et savoir-être); une *c-ressource* est liée à un aspect de l'entreprise (partie du modèle d'entreprise); un individu possède une ou plusieurs *c-ressources* (Fig. 1). Par exemple, la compétence «être compétent en machine W21» concerne l'aspect «machine W21» et comprend les *c-ressources* : «connaître les composants de la machine W21», «savoir réparer une panne de la machine W21», «savoir shunter une fonction de la machine W21» et éventuellement d'autres *c-ressources*; la compétence «être compétent en relations avec les clients» concerne l'aspect «relations avec les clients» et comprend les *c-ressources* : «connaître les besoins des clients», «savoir relier les bénéfices d'un produit à ses fonctionnalités», «savoir simplifier l'information», etc. La définition d'une compétence requise est ainsi basée sur un *modèle d'entreprise* (Vernadat 1996) qui représente normalement les aspects organisationnels, décisionnels, fonctionnels, informationnels et des ressources.

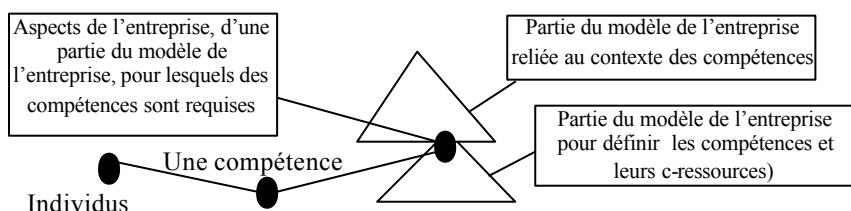


Fig. 1 - La structure du modèle CRAI

Cependant, une modélisation des compétences individuelles à l'aide des techniques classiques de l'ingénierie des systèmes d'information n'est pas suffisante pour assurer une gestion efficace des compétences : plusieurs processus complexes, dépendants et lourds restent à mener. Pour cela, il nous semble intéressant d'investiguer les méthodes et techniques d'ingénierie des connaissances (IC) et de les appliquer à la gestion des compétences (GC). En effet, la compétence est un type de connaissances de l'entreprise, à acquérir, à représenter, et à exploiter. Il est donc important de voir quand et comment ces techniques peuvent être utilisées en support aux processus de gestion des compétences.

Dans cet article, nous exposons notre point de vue sur les différents processus de gestion des compétences. Ensuite, nous utilisons ces processus pour définir les fonctionnalités requises d'un système de GC. Nous utilisons ces mêmes fonctionnalités pour classifier les travaux de recherche dans la littérature, portant sur

la GC utilisant des techniques d'IC. Pour mieux synthétiser ces travaux et situer des techniques d'IC envisageables, nous considérons trois catégories de ces techniques : acquisition de connaissance, extraction de connaissance et raisonnement. Une synthèse et des perspectives sont présentées en conclusions.

2. La gestion des compétences

Nous distinguons quatre types de processus dans la gestion des compétences de l'entreprise (chacun comprend un ou plusieurs sous-processus) :

1. L'identification des compétences requises, à savoir l'identification et la définition des compétences requises par les tâches, les missions, les objectifs, etc. de l'entreprise ;
2. L'acquisition des compétences, à savoir l'acquisition des compétences par les employés de l'entreprise, par exemple, à l'aide de la formation ;
3. L'identification et l'évaluation des compétences acquises, à savoir l'identification des nouvelles compétences acquises par les individus ou l'évaluation de l'acquisition des compétences requises de l'entreprise par les individus ;
4. L'exploitation des compétences, à savoir l'exploitation des données et connaissances déterminées dans les processus d'identification et d'évaluation des compétences.

Les objectifs des trois premiers types de processus sont bien définis. L'objectif de l'exploitation des compétences est très varié, il dépend de l'entreprise. Par exemple, dans nos travaux précédents sur la gestion des compétences, nous avons considéré comme objectif la réorganisation d'entreprise basée sur les compétences (Harzallah & Berio, 2004).

Ces quatre types de processus interagissent. Les processus d'évaluation et d'acquisition des compétences peuvent être basés sur les compétences identifiées dans le premier processus. Des nouvelles compétences peuvent être identifiées suite aux processus d'évaluation ou d'acquisition des compétences. Les compétences à acquérir sont celles dont l'évaluation est insatisfaisante. Enfin, l'exploitation des compétences est basée sur les données concernant les compétences acquises et requises. Des relations itératives entre processus permettent de fiabiliser et améliorer l'efficacité d'un système de GC. Si ces processus sont orientés vers les objectifs de l'entreprise (par exemple : l'identification des compétences requises pour atteindre les objectifs d'une unité d'organisation de l'entreprise) alors l'efficacité d'un système de GC participe dans l'amélioration de la *performance de l'entreprise*, d'où l'intérêt de développer un système efficace de GC en entreprise.

3. Ingénierie des connaissances appliquée à la gestion des compétences

Dans la littérature, des travaux de recherche se sont intéressés à l'application des techniques d'IC à la GC. Ils ont porté sur les différents types de processus de GC cités

dans la Section 2. Dans la suite, nous présentons ces travaux et les classifions selon leur apport, en définissant des fonctionnalités bien précises, à chaque type de processus de GC. Ces travaux sont également classés en trois catégories : 1) *Acquisition manuelle ou Semi-automatique des Connaissances (ASC)* centrée sur les experts du domaine, aussi connue sous le nom « Knowledge Elicitation » (Schreider 1999) ; 2) *Extraction Automatique des Connaissances (EAC)* centrée plutôt sur l'utilisation d'algorithmes de découverte de la connaissance aussi connue sous les noms de « machine learning, data mining, text mining, etc. » (Kodratoff, 2001) ; 3) *Techniques de Raisonnement (TR)* à savoir toutes les techniques qui s'appuient sur la notion d'épreuve, de propriété ou de théorème, qui sont associées à une logique quelconque et qui formalisent l'idée de déduction à partir d'un modèle ou d'une théorie donné(e). Ces trois catégories utilisent des techniques de modélisation des connaissances. En outre, une technique peut utiliser une autre technique : par exemple certaines techniques d'extraction automatique demandent des techniques de raisonnement.

3.1 Ingénierie des connaissances pour l'identification des compétences

Les méthodes les plus connues chez les gestionnaires pour l'identification des compétences requises sont l'observation directe, l'auto-description, les interviews, la méthode des incidents critiques et la méthode de la grille de Kelly. Cependant, il reste difficile d'identifier les compétences requises sans un modèle de référence. Dans nos travaux précédents, nous nous sommes basés sur le modèle CRAI et les modèles d'entreprise pour identifier les compétences requises.

Plusieurs travaux de recherche ont développé des *ontologies des compétences*, comme *modèle de référence pour l'identification des compétences requises*. Il peut s'agir d'une *ontologie de référence* qui représente la structure générique des compétences et qui est ensuite spécialisée pour l'identification des compétences d'une entreprise particulière (Colucci *et al.* 2003), (Vasconcelos, 2003), (Posea & Harzallah, 2004). Il peut s'agir également d'une ontologie spécifique à un domaine (entreprise) donné(e) : par exemple, dans (Colucci *et al.* 2003) et dans le projet KMP (Corby *et al.* 2004). Dans les deux cas, le recours à une ontologie pour les compétences a été fait pour pouvoir partager la connaissance ce qui est très important pour la gestion des compétences ; si l'ontologie fait partie du cœur du système de gestion des compétences, elle est parmi les moyens d'intégration pour les autres systèmes et pour le personnel de l'entreprise qu'y feront référence.

L'ontologie des compétences d'une entreprise peut être définie d'une façon manuelle ou (semi) automatique. La construction manuelle peut se faire en utilisant des documents de référence de l'entreprise et en demandant aux experts quel type de compétence a fallu pour créer ces documents (Ley & Albert, 2003). Il s'agit bien des compétences requises puisqu'elles sont liées aux documents réellement utilisés dans l'entreprise. Cette approche s'applique aux organisations centrées plutôt sur les services que sur la production. Corby *et al.* (2004) utilisent une ontologie spécifique pour annoter les documents électroniques de l'entreprise et identifier ensuite les compétences fournies par l'entreprise.

Toutefois, en ce qui concerne l'utilisation des ontologies, il nous semble que les ontologies des compétences développées dans le cadre de ces travaux ne sont pas

aussi explicites que le modèle CRAI. En effet, ces ontologies ne prennent pas en compte toutes les caractéristiques du concept de compétence, telles que la différence entre les savoir et les savoir-faire, la relation sous-jacente entre les compétences requises et les objectifs et les missions de l'entreprise ; de même, dans Colucci et al. (2003), les auteurs ne différencient pas ni entre compétences acquises et requises ni entre la qualification, la disponibilité des individus, les aspects du modèle d'entreprise (à savoir un produit, un processus etc.) et les compétences proprement dites.

Table 1 - Application des techniques d'IC à l'identification des compétences

Définition des fonctionnalités	Données d'entrée	Données de sortie	Techniques d'IC
Apprentissage (à partir de l'expérience) des compétences nécessaires pour réaliser une tâche ou atteindre un objectif, etc.	Historiques des données et documents sur les tâches, missions, etc., l'entreprise.	Compétences requises par tâche, mission, objectif, etc.	EAC (Annotation automatique des documents de l'entreprise et déduction de ses compétences dans le projet KMP (Corby et al. 2004)).
Identifier les compétences requises par une tâche nécessaire pour accomplir une mission ou atteindre un objectif dans l'entreprise.	Modèles de référence du domaine de l'entreprise, Modèle spécifique de l'entreprise.	Compétences requises par tâche, mission, objectif, etc.	ASC (interview dans CRAI ; interview en utilisant les documents de l'entreprise et validation avec des techniques statistiques issues de la psychologie dans (Ley & Albert, 2003)).
Définir les compétences requises.	Les aspects de l'entreprise concernés par les compétences requises.	C-ressources définissant les compétences requises.	ASC (Ley & Albert (2003) proposent une méthode de définition des compétences basée sur les tâches individuelles).
Mise à jour des compétences requises parce que les objectifs fixés n'ont pas été atteints.	Historiques des données et documents sur les tâches, missions, etc. de l'entreprise.	(Nouvelles) compétences requises par une tâche, une mission, etc.	EAC (extraction à partir des données réelles de l'entreprise).
Mise à jour des compétences requises suite à l'ajout d'un nouvel aspect dans le modèle de l'entreprise.	Modèle de référence et modèle de l'entreprise mis à jour.	(Nouvelles) compétences requises par une tâche, une mission, etc.	TR (Classification de concepts en utilisant la logique de description (Colucci et al. 2003)).
Mise à jour des compétences requises suite à la proposition de l'employée de nouvelles compétences requises pour ses tâches.	Définition des compétences proposées.	Nouvelles compétences requises par une tâche, une mission, etc. validées.	TR (Classification de concepts en utilisant la logique de description (Colucci et al. 2003)).

Pour représenter une ontologie des compétences, la *logique de description* peut être utilisée (Colucci et al. 2003), (Vasconcelos et al. 2003). Ce langage a été choisi pour plusieurs raisons : 1) la disponibilité d'outils de raisonnement, 2) la possibilité d'utiliser des langages pour le web (tel que OWL), 3) l'utilisation du «open world assumption» qui permet de représenter des compétences partiellement connues (Colucci et al. 2003). La logique de description n'est qu'un des langages possibles de modélisation des ontologies : d'autres langages logiques permettant d'exprimer des règles peuvent être utilisés.

La *classification de concepts* (« subsumption ») est appliquée dans (Colucci *et al.* 2003) pour la mise à jour des nouvelles compétences requises suite à l'identification des nouvelles compétences ou l'introduction des nouveaux concepts dans le modèle d'entreprise ou dans une ontologie de compétences.

Le tableau 1 ci-dessus présente l'ensemble des fonctionnalités qui nous semblaient appropriées à un système en support de l'identification des compétences requises. Pour chaque fonctionnalité, nous associons les travaux présentés brièvement ci-dessus et qui fournissent un support, même partiel, à cette fonctionnalité. Les travaux sont aussi classés dans une de trois catégories d'IC (en cas d'absence de travaux, la catégorie est utilisée pour indiquer un ensemble de techniques envisageables pour la fonctionnalité considérée).

3.2 Ingénierie des connaissances pour l'acquisition des compétences

La formation est un moyen classique d'acquisition des compétences. Elle peut être considérée comme un ensemble de modules de cours à acquérir, chaque module est représenté par certaines conditions d'entrée et des résultats de sortie. La formation et plus généralement les processus d'acquisition des compétences peuvent être supportés par un *système de « e-learning »*. Aujourd'hui, ce type de système distingue les fonctionnalités suivantes :

- Définition de la méthode d'acquisition (à savoir la planification de la méthode et des ressources nécessaires),
- Déploiement de la méthode d'acquisition aux individus (adapter la méthode envisagée pour délivrer la connaissance de la meilleure façon, prenant en compte les caractéristiques des individus eux-mêmes et le contenu à acquérir).

Un système de «e-learning» peut se baser sur une *ontologie de compétences* permettant entre autre de décrire les contraintes des formations (à savoir les pré-requis entre cours ou concepts à apprendre) et les combinaisons des formations. Ce même système peut intégrer un module d'évaluation des compétences des individus basé sur les formations reçues, en appliquant des techniques de raisonnement sur le planning des formations et les compétences leur sont associées. Il nous semble donc intéressant d'intégrer un système de GC et un système de «e-learning» qui utilise des techniques ASC, EAC et TR à la fois. Pour cela, nous ne sommes pas intéressés à répertorier toutes les fonctionnalités de ce type de système, mais plutôt, d'y inclure les fonctionnalités concernant comment l'entreprise ou un employé peut décider l'intérêt d'une formation, par rapport à des compétences requises.

Dans l'acquisition des compétences, nous pouvons considérer la recherche et la sélection d'un individu ayant des compétences spécifiques (où le chercher, comment le chercher, etc.). Cette recherche peut être difficile surtout si elle concerne des compétences assez spécifiques à l'entreprise. Elle peut porter sur des individus à l'extérieur de l'entreprise, en utilisant des documents ou information externes. Elle peut porter également sur le personnel de l'entreprise, elle s'inscrit plutôt dans l'identification et l'évaluation des compétences acquises. Dans la sélection d'un individu, nous distinguons l'évaluation de ses compétences de la prise de décision

pour son embauche (ce dernier processus n'est pas considéré dans les fonctionnalités d'un système de GC).

Le tableau 2 est structuré suivant la même philosophie du tableau 1.

Table 2 – Application des techniques d'IC à l'acquisition des compétences

Définition des fonctionnalités	Données d'entrée	Données de sortie	Techniques d'IC
Définition d'un plan de formation pour un ensemble de compétences requises.	Compétences requises, modèle des aspects liés aux compétences requises. Ensemble des modules de formation.	Pré-requis de modules de formation. Modules des cours de formation, leur séquence. Connaissances à apprendre dans chaque module.	TR (méthodes de résolution de problèmes ou de planning (Baldoni <i>et al.</i> 2004)).
Validation d'une formation pour acquérir un ensemble de compétences requises.	Cours de formation décrits en fonction des entrées et sorties en terme des compétences	Formation validée amélioration ou modification recommandée pour la formation	TR ((Baldoni <i>et al.</i> 2004)).
Recherche continue des individus ayant des compétences requises.	Documents externes, CV, classés par compétence requise	Liste des individus	EAC & TR (par exemple, l'utilisation d'un « recommender systems»)

3.3 Ingénierie des connaissances pour l'identification et l'évaluation des compétences acquises

Les méthodes les plus connues chez les gestionnaires pour l'évaluation des compétences acquises sont les tests, l'entretien d'appreciation ou d'évaluation du personnel, les échantillons, les références et les bilans des compétences. L'identification et l'évaluation des compétences s'effectuent d'habitude par un responsable hiérarchique qui utilise ses connaissances sur ses employées (les comportements, les tâches réalisées, la performance atteinte) pour les évaluer. Cette identification et évaluation peuvent se réaliser en interviewant la personne concernée. Ce type d'interview peut être semi-structuré et supporté par un système informatique (« self assessment »). Trichet *et al.* (2002) proposent un système où un individu choisit sa liste de compétences dans une ontologie de référence des compétences. Ces processus restent assez lourds et dont les résultats sont souvent insatisfaisants. Il y a donc un intérêt très fort dans leur automatisation. Par exemple, Lindgren *et al.* (2003) propose un « recommender system » qui utilise les intérêts d'un individu pour déduire ses compétences (suivant les auteurs de l'article, les intérêts d'un individu sont très fortement liés à ses compétences). L'utilisation de la notion d'intérêt est sûrement intéressante. En effet, elle est beaucoup plus souple que la notion de compétence mais elle constitue une représentation simplifiée de la compétence.

Des méthodes d'extraction automatisée d'identification des compétences à partir des documents associés aux individus, basées sur le text-mining sont proposées. Par exemple, Becerra (2000) effectue une mise à jour des compétences acquises, en faisant une extraction des mots clés des documents associés aux individus, les mots clés représentant des domaines des compétences (plutôt que les compétences elles-mêmes). Garro & Palopoli (2003) considèrent un système de « e-learning » qui peut

intégrer des compétences, enregistrer les traces des formations qu'un individu a eu et en déduire ses compétences.

Table 3 – Application des techniques d'IC à l'évaluation des compétences

Définition des fonctionnalités	Données d'entrée	Données de sortie	Techniques d'IC
Identification et définition des compétences acquises par un individu.	Historiques des activités réalisées ou activités d'apprentissage par un employé ; CV.	Liste des compétences acquises par cet individu.	EAC (par exemple, text-mining sur CV). ASC (Trichet <i>et al.</i> 2002) proposent un système pour saisir un CV à l'aide d'une ontologie.
Identification des compétences d'un individu par rapport à certaines compétences requises	Compétences requises pour l'individu Définition d'une méthode de test pour les compétences requises	La liste des compétences d'un individu et leur nouvelle évaluation	EAC (par exemple, à partir de résultats d'un test automatisé ; utilisation des données réelles)
Identification des individus potentiellement associés à certaines compétences requises	Documents et données de l'entreprise disponibles, classés par compétence requise	Liste des individus	EAC ((Becerra, 2000) propose un système pour la recherche des individus, experts dans un domaine d'intérêt de l'entreprise)
Evaluation et mise à jour des compétences acquises	Les compétences acquises y compris leur évaluation ; Définition d'un ensemble de compétences et d'une méthode de test conséquente	Les compétences acquises y compris leur nouvelle évaluation	EAC ((Becerra, 2000) propose d'associer aux individus des mots-clés trouvés dans les documents de l'entreprise ; TR ((Blanchard & Harzallah, 2004), (Sure <i>et al.</i> 2000) proposent des règles logiques pour mettre à jour les compétences acquises) ; (méthodes de test interactives, (Baldoni <i>et al.</i> 2004))

(Blanchard & Harzallah, 2004), (Sure *et al.* 2000) se sont intéressés aux techniques de raisonnement pour déduire les compétences d'un individu à partir d'un sous-ensemble de ses compétences ou de ses projets. Ces règles sont induites dans le premier cas et de type : « si C alors C' » étant C et C' deux compétences. Dans le deuxième cas, elles sont directement définies par des experts et liées aux (meta) données de l'entreprise : par exemple, « si une personne a travaillé sur des activités d'un projet utilisant le langage C, alors elle est compétente en C ».

En ce qui concerne l'évaluation des compétences, un système de e-learning peut intégrer un module de test permettant d'évaluer l'acquisition des compétences, plus au moins sophistiquée : d'un simple ensemble de réponses à donner, jusqu'à l'évaluation de la résolution des problèmes (Baldoni *et al.* 2004).

Toutes les approches ainsi citées sont centrées sur les individus dont il faut identifier et évaluer leurs compétences acquises. Quand il s'agit des compétences spécifiques ou méconnues, il peut être difficile d'expliquer exactement leur acquisition. Dans ce cas, il peut être intéressant de se contenter d'associer les individus à un domaine de compétences (à savoir les aspects associés à la compétence dans le modèle CRAI) en définissant des conditions précises telles que la fréquence

des mots clés (Becerra, 2000) dans les documents associés à ces individus ou en utilisant la notion d'intérêt (Lindgren *et al.*, 2003).

3.4 L'ingénierie connaissances pour l'exploitation des compétences

Les processus d'exploitation des compétences sont basés sur l'étude de l'adéquation des compétences acquises et requises, leurs objectifs visent toujours l'amélioration de la performance de l'entreprise. Ils portent sur la vérification ou la recherche de l'adéquation entre un profil d'individu ou d'un groupe et un profil requis. Il peut s'agir d'une adéquation simple et quantitative (toutes les compétences requises doivent être acquises) (Becerra, 2000), (Harzallah & Berio, 2004), ou d'une adéquation sémantique et approchée. Par exemple, (Colucci *et al.* 2004), (Sure *et al.* 2000), (Garro & Palopoli, 2003) proposent des algorithmes pour définir des différents types de « correspondance sémantique » : potentiel, partiel, pondéré, entre un profil demandé et des profils proposés. Les deux premiers s'appuient sur une description formelle, à l'aide de la logique de description, des compétences requises et des profils et sur la définition de distances sémantiques entre eux. Garro & Palopoli (2003) utilisent XML et notamment celui proposé par l'organisme IMS (www.imsproject.org).

Ce tableau est partiel, les fonctionnalités de l'exploitation des compétences ne sont pas limitées. Par exemple, en relation avec les fonctionnalités typiques d'un système de gestion des ressources humaines, la planification de la carrière nous semble une fonctionnalité intéressante à prendre en compte.

Toutefois, l'efficacité des processus d'exploitation est très liée aux autres types de processus de GC. En effet, si les processus d'identification des compétences requises et d'évaluation des compétences acquises se basent sur une représentation assez complète et précise des compétences les processus d'exploitation des compétences permettront d'obtenir des résultats fiables (moins approchés).

Table 3 – Application des techniques d'IC à l'exploitation des compétences

Définition des fonctionnalités	Données d'entrée	Données de sortie	Techniques d'IC
Etudier l'adéquation entre les compétences requises et les compétences acquises pour réaliser une tâche, accomplir une mission, etc. dans l'entreprise.	Le modèle de l'entreprise, les compétences requises et acquises de l'entreprise.	La liste de compétences à acquérir ou la liste supplémentaire de compétences acquises par les individus.	TR (par exemple, les méthodes de résolution de problèmes).
Identifier les individus capables de réaliser des tâches spécifiques.	Tâches et leurs compétences requises, compétences acquises	Liste d'individus	TR ((Colucci <i>et al.</i> 2004) proposent la définition et l'utilisation d'une similarité entre compétences ; Dans le projet KMP (Corby <i>et al.</i> 2004), il est possible d'utiliser un moteur de recherche, CORESE, permettant d'introduire des liens sémantiques tels que la similarité)

4. Synthèse et Conclusion

La gestion des compétences comprend plusieurs processus complexes et lourds à réaliser. L'application des techniques classiques des systèmes d'information n'est pas suffisante pour bien mener ces processus. L'application des techniques d'ingénierie des connaissances à la gestion des compétences semble intéressante. Nous avons identifié et étudié des travaux de recherche dans la littérature, ayant appliqué des techniques de l'IC à la GC. Ces travaux portent sur les différents processus de gestion des compétences et s'intègrent dans trois catégories : EAC, ASC et TR. Ils utilisent souvent un langage formel (LD, FLogic, théorie des ensemble, XML, etc.) pour définir une ontologie des compétences et effectuer des raisonnements sur ces dernières. Pour simplifier l'évaluation des compétences acquises, ils utilisent des méthodes d'extraction des compétences à partir des documents qui peuvent être associés aux individus, à leur intérêts ou aux tâches réalisées. Enfin, pour établir une correspondance sémantique entre les profils requis et acquis, ils proposent des algorithmes portant sur la distance sémantique.

Dans tous les cas, il nous semble que les travaux répertoriés visent la *performance des processus de gestion des compétences* en essayant d'en automatiser des tâches lourdes.

Dans la plupart de ces travaux, le concept de compétence n'a pas été bien explicité. Il nous semble intéressant d'avoir un système qui permet de bien représenter les compétences requises et acquises qu'aujourd'hui restent confuses et d'y intégrer une véritable ontologie de gestion de compétences, qui en considérant les données réelles de l'entreprise permet une mise à jour souple des compétences acquises et requises.

En outre, ces travaux s'intéressent et se focalisent sur un type de processus de gestion des compétences parmi les quatre introduits. Il nous semble intéressant d'analyser les possibilités d'intégrer ces processus étroitement via une représentation commune des compétences, pouvant être utilisée pour réaliser toutes les fonctionnalités. Ensuite, il s'agit d'intégrer les différentes techniques d'IC répertoriées et donc de proposer une boîte à méthodes pour les différents processus de gestion des compétences. Cela permet d'implanter les liaisons entre les différents processus de GC et, par conséquence, une exploitation efficace et consistante des compétences acquises et requises.

Les processus d'identification des compétences acquises et requises et aussi de leur acquisition restent très difficile dans certains cas, en particulier, s'il faut être réactif à des problèmes spécifiques et non récurrents, innover, donner des solutions à des problèmes méconnus et très spécifiques ou partager des connaissances peu structurées telles que des expériences sur les projets. Ces cas sont caractérisés par une réactivité assez importante qui empêche une véritable phase de représentation et d'évaluation des compétences ; des problèmes non récurrents (jamais ou peu abordés) où les compétences sur un domaine technique ne sont qu'un pré-requis et les compétences de créativité, de raisonnement et de synthèse sont primordiales ; la difficulté de représentation des compétences utilisées ou des solutions mises en place, le recours aux experts directement plutôt que d'essayer d'extraire leur expertise ; etc. Dans ces cas, la notion de compétence et de sa gestion comme ont été définies dans cet article peuvent devenir difficiles à appliquer. En effet, nous nous sommes

intéressés aux compétences des individus, dont nous sommes capables de donner une représentation et une évaluation explicites (Modèle CRAI). Dans certains travaux répertoriés dans cet article, la représentation et l'évaluation des compétences sont approximatives (par exemple, appréhension de la compétence par une liste d'intérêts ou de mots clés, etc.). Toutefois, nous pouvons nous demander sur leur efficacité et leur frontière par rapport à la gestion des compétences.

Dans les cas cités ci-dessus, des nouveaux paradigmes organisationnels tels que les communautés de pratique (Wengler, 1991) s'imposent. En général, on pourrait dire que s'il est difficile de représenter ou d'évaluer la compétence, il peut être intéressant de partager la connaissance sur un domaine en mettant en contact des individus de ce domaine : l'accent devient donc sur comment permettre à un ensemble d'individus (la communauté) de collaborer sur certains problèmes et pas forcément sur comment définir d'une façon plus ou moins précise les compétences de ces individus. En effet, ces paradigmes ne sont pas alternatives à la gestion des compétences telle qu'elle a été définie dans cet article ; plutôt, il s'agit des paradigmes complémentaires (Gongla & Rizzuto, 2001) qui, dans certains cas, sont peut-être mieux adaptés pour l'amélioration de la performance de l'entreprises (Lesser & Storck, 2001).

Références

- Blanchard E., Harzallah M (2004). Reasoning on competencies, In *Proc. of the Workshop Knowledge Management and Organizational Memories* (joint with ECAI2004).
- Baldoni M., Broglia C., Patti V., and Torasso L. (2004) Reasoning about learning object metadata for adapting SCORM courseware. In *Proc. of International Workshop on Engineering the Adaptive Web, EAW'04: Methods and Technologies for personalization and Adaptation in the Semantic Web*, Eindhoven, Pays-Bas.
- Becerra I. (2000). The role of artificial intelligence technologies in the implementation of people-finder knowledge management systems. In (Staab and O'Leary Eds.) *Bringing knowledge to business processes. Workshop in the AAAI Spring Symposium Series*. Stanford.
- Colucci S., Di Noia T., Di Sciascio E., Donini F. M., Mongiello M., and Mottola M. (2003). A formal approach to ontology-based semantic match of skills descriptions. *Journal of Universal Computer Science*, Special issue on Skills Management.
- Corby O., Dieng-Kuntz R., Faron-Zucker C. (2004). Querying the Semantic Web with the CORESE search engine. In (R. Lopez de Mantaras and L. Saitta Eds) *Proc. of the 16th European Conference on Artificial Intelligence (ECAI'2004)*, subconference PAIS'2004, Valencia, Espagne, IOS Press, pp. 705-709.
- Garro A., Palopoli L. (2003). An XML MultiAgent System for e-Learning and Skill Management. In (H.Tianfield R.Kowalczyk, J.P.Muller and R.Unland, Eds) *Agent Technologies, Infrastructures, Tools, and Applications for E-Services*, LNAI 2592. Springer-Verlag.
- Gongla P., Rizzuto C.R. (2001). Evolving communities of practice: IBM Global Services experience. In *IBM Systems Journal*, vol. 40, n. 4.
- Harzallah M., F. Vernadat (2002). IT-based Competency Modeling and Management: from Theory to Practice in Enterprise Engineering and Operations. In *Computers In Industry*, vol. 48, pp. 157-179.

- Harzallah, M. and Berio, G. (2004). Competency Modeling and management: A case study. In *Proceedings of the 6th international conference on Enterprise Information Systems (ICEIS'04)*, University Portucalense, pp. 350-358, Porto, Portugal.
- Kodratoff Y. (2001). Applications de l'apprentissage automatique et de la fouille de données. In *EGC'01*.
- Le Boterf G. (1997). *Construire les compétences individuelles et collectives*, Les Editions d'Organisation, Paris.
- Ley T., Albert D., (2003). Identifying employee competencies in dynamic work domains: methodological considerations and a case. *Journal of Universal Computer Science*, vol. 9 n. 12, pp. 1500-1518.
- Lesser E. L., Storck J. (2001). Communities of practice and organizational performance. In *IBM Systems Journal*, vol. 40, n. 4.
- Levy-Leboyer C. (1996). *Evaluation du personnel : Quelles méthodes choisir?* Les Editions d'Organisation, Paris, France.
- Lindgren R., Stenmark D. and Ljungberg J. (2003). Rethinking competence systems for knowledge-based organisations. *European Journal of Information Systems*, vol.12, n. 1, pp. 18-29.
- Lucia A. D., Lepsinger, R. (1999). *The art and science of competency: Pinpointing critical success factors in organizations*, Edition Hardcover.
- Michel, S. (1997). Le savoir est-il une compétences. In *Actes de la conférence en Compétences & Contextes Professionnels*, Metz, France, pp. 7-13.
- Pfeffer J., Sutton, R.I. (2000). *The Knowing-Doing Gap: How smart companies turn knowledge into action*, Edition Hardcover.
- Posea V., Harzallah M. (2004). Building a competence ontology. In *Proc. of the Workshop Enterprise modelling and Ontology: Ingredients for interoperability* (joint with PAKM 2004) Vienne, Autriche.
- Schreiber G., Hakermans A., Anjewierden A., de Hoog, R., Shadbolt N., Van de Welde W., et Wielinga B. (1999). *Knowledge Engineering and management: The CommonKADS methodology*, The MIT Press.
- Sure, Y., Maedche A., and Staab S: (2000). Leveraging Corporate Skill Knowledge - From ProPer to OntoProPer. In (D. Mahling & U. Reimer Eds) *Proceedings of the Third International Conference on Practical Aspects of Knowledge Management (PAKM 2000)*, Bale, Suisse.
- Trichet F., Bourse M., Harzallah M. and Leclère M (2002). CommOnCV: modeling the competencies underlying a Curriculum Vitae. In *Proc. of the 14th int. conf. on Software Engineering and Knowledge Engineering (SEKE'2002)*. ACM Press, Ischia, Italie, pp. 65-73.
- Vasconcelos J. B. Kimble C., Rocha A., (2003). Ontologies and the Dynamics of Organisational Environments. An example of a Group Memory System for the Management of Group Competencies. In *Proc. of I-KNOW '03 - 3rd International Conference on Knowledge Management*, Graz, Autriche.
- Vernadat F. (1996). *Enterprise Modeling and Integration: Principles and Applications*, London: Chapman & Hall.
- Wenger E. (1991). Communities of practice where learning takes place. *Benchmark Magazine*, Fall Issue.

Dérivation d'un arbre de décision pour la mise en œuvre de stratégies thérapeutiques dans le cas des maladies chroniques

Jacques Bouaud, Brigitte Séroussi et Jean-Jacques Vieillot

STIM, DPA/DSI/AP-HP & INSERM, U729, Paris, France
{jacques.bouaud,brigitte.seroussi}@sap.aphp.fr

Résumé : La prise en charge des maladies chroniques par les médecins généralistes est une tâche difficile. En effet, contrairement aux maladies aigües où le traitement est ponctuel, les maladies chroniques évoluent au cours du temps et la prescription d'un nouveau traitement dépend des traitements précédemment administrés et de la réponse du patient à ces traitements. Dans le cadre du projet ASTI¹ dont l'objectif est d'élaborer un système d'aide à la décision permettant l'implémentation et la diffusion des guides de bonnes pratiques (GBP), nous avions représenté la base de connaissances du mode guidé (construction manuelle) sous la forme d'un arbre de décision à deux niveaux, le niveau clinique pour caractériser la situation clinique et le niveau thérapeutique pour déterminer le nouveau traitement recommandé sous contrainte de l'historique thérapeutique du patient. L'objectif de ce nouveau travail est de proposer une méthode de dérivation automatique du niveau thérapeutique de l'arbre de décision à partir d'une formalisation des séquences thérapeutiques recommandées décrites dans le GBP. Une implémentation a été réalisée avec des ATN (*Automated Transition Network*). Les premiers résultats obtenus sur des séquences additives du type $(A, A + B, A + B + C)$, où A , B , et C sont des médicaments, sont encourageants et la méthode doit être étendue afin de prendre en compte des modèles plus complexes.

Mots-clés : Guides de bonnes pratiques, Construction de bases de connaissances, Stratégies thérapeutiques, Maladies chroniques, Arbre de décision.

1 Introduction

Les maladies chroniques sont de plus en plus l'objet de préoccupations étant donné le vieillissement rapide de la population et l'augmentation de plusieurs facteurs de risque qui y sont associés. À ce titre, la qualité de la prise en charge thérapeutique de ces

¹ ASTI est un projet dont la première phase (1999 – 2001) a reçu une subvention du Ministère de l'Éducation Nationale, de la Recherche, et de la Technologie (MENRT), et la deuxième phase, actuellement en cours, (2003 – 2006) est subventionnée par la CNAM.

maladies constituent aujourd’hui un enjeu majeur de santé publique dans la majorité des pays développés ou en voie de développement.

Comme les maladies aigües, les maladies chroniques posent le problème du choix du traitement médicamenteux car l’arsenal thérapeutique mis à la disposition des médecins est difficile à gérer. La sélection de la monothérapie qui convient, c’est-à-dire le traitement faisant intervenir la classe médicamenteuse adaptée, est une tâche difficile. Par ailleurs, du fait de l’intensification thérapeutique à la recherche d’une meilleure efficacité, les médecins s’orientent vers des schémas de bi et trithérapies, et le choix des bonnes associations de classes médicamenteuses du fait de la combinatoire devient presque aléatoire. Mais, si les maladies aigües se caractérisent par une décision thérapeutique « en un coup » (*one-shot*), la prise en charge des maladies chroniques est une collaboration patient-médecin au long cours dans laquelle les décisions thérapeutiques sont non seulement dépendantes de l’état clinique courant du patient, mais également des traitements qu’il a déjà reçus, et de ses réponses à ces traitements en termes d’efficacité et de tolérance, un ensemble d’informations qui constituent son historique thérapeutique.

Aujourd’hui, le médecin généraliste est supposé équipé pour aborder le problème de la décision thérapeutique posé par un patient atteint d’une maladie chronique. En effet, portés par le mouvement de l'*evidence-based medicine* (EBM), les guides de bonnes pratiques (GBP) ou *guidelines* apparaissent comme la solution à la rationnalisation de la prise en charge médicamenteuse des patients. Ainsi, de nombreux GBP sont actuellement élaborés à l’intention des médecins généralistes par des agences nationales telles que, en France, l’ANAES ou l’AFSSAPS, afin d’harmoniser les pratiques dans des domaines comme la prise en charge de l’hypertension artérielle (HTA), du diabète, ou des cancers. Ces GBP sont habituellement organisés sous la forme d’un catalogue de situations particulières dans lesquelles figure la maladie chronique qui fait l’objet du GBP associée à d’autres pathologies. Pour chacune de ces situations particulières, la conduite à tenir thérapeutique recommandée est décrite. On aura ainsi les recommandations de traitement pour le patient hypertendu et diabétique, le patient hypertendu et insuffisant rénal, le patient hypertendu et insuffisant cardiaque, etc., dans les GBP de prise en charge de l’HTA.

Pourtant ces GBP ne répondent pas au besoin des médecins généralistes pour une aide à l’identification, pour un patient donné, du meilleur traitement. En effet, si le traitement initial est effectivement décrit dans les GBP, les étapes ultérieures de la prise en charge ne le sont pas toujours, et lorsqu’elles le sont, elles apparaissent sous la forme de transitions. Ainsi, on pourra trouver dans le GBP que le traitement recommandé pour un patient traité par Tr_1 et dont la réponse au traitement est partielle est le traitement Tr_2 mais il n’y aura aucune aide à la thérapeutique pour le médecin si le patient est traité par T , avec $T \neq Tr_1$. D’une manière plus générale, il n’existe pas de séquence de prise en charge thérapeutique recommandée pour chacune des situations particulières répertoriées : on ne trouve en général que le traitement initial et certaines transitions.

En pratique, les séquences de prise en charge attachées aux situations cliniques du GBP peuvent être reconstruites à partir du traitement initial et des transitions thérapeutiques sous contrainte de se fixer un cadre d’interprétation du texte des recommandations afin de lever certaines ambiguïtés sémantiques (Bouaud & Sérussi, 2003). Ainsi,

pour une situation clinique donnée, les GBP permettent d'établir la séquence de traitements recommandés (Tr_1, Tr_2, \dots, Tr_n). Quoique adaptées a priori à la situation clinique, ces séquences n'en restent pas moins théoriques et génériques. En effet, elles ne sont en aucun cas centrées patient puisqu'elles ne prennent pas en compte l'historique thérapeutique du patient, c'est-à-dire les traitements précédemment administrés et sa réponse à ces traitements.

Plusieurs situations doivent ainsi être considérées. Lorsqu'on est dans le cas d'un traitement initial, c'est-à-dire lorsque l'historique thérapeutique du patient est vide, le traitement recommandé est le premier traitement de la séquence thérapeutique recommandée, soit Tr_1 , sauf si Tr_1 est contre-indiqué. Sinon, lorsque le patient présente un historique thérapeutique non vide, il faut « positionner » le patient au sein de la séquence thérapeutique recommandée. Deux cas de figure doivent alors être envisagés :

- Le patient possède un historique thérapeutique cohérent avec les premières étapes de la stratégie thérapeutique recommandée, c'est-à-dire (Tr_1, \dots, Tr_i). La nouvelle décision consiste à choisir l'étape suivante dans la séquence recommandée, soit Tr_{i+1} .
- Le patient possède un historique thérapeutique totalement ou partiellement incohérent avec la stratégie recommandée, ce qui peut arriver lorsque les patients sont traités depuis longtemps. Ce dernier cas de figure, pourtant très fréquent en pratique, n'est pas explicité par les GBP, ce qui laisse les médecins extrêmement démunis.

Dans le cadre du projet ASTI (Séroussi *et al.*, 2001b), nous avons développé le « mode guidé » (Bouaud & Séroussi, 2003) conformément aux principes de l'approche documentaire de l'aide à la décision établis dans OncoDoc (Séroussi *et al.*, 2001a). La base de connaissances a été représentée sous la forme d'un arbre de décision, une étape manuelle extrêmement coûteuse. Aussi, l'objectif de ce travail est de proposer une méthode de construction automatique du niveau thérapeutique de l'arbre de décision à partir des séquences génériques associées aux situations cliniques et obtenues sous contrainte d'un modèle d'interprétation du texte des GBP. La méthode proposée passe par une étape intermédiaire basée sur le développement d'une stratégie d'exploitation des séquences thérapeutiques génériques, implémentée au moyen d'ATN (*Augmented Transition Network*).

2 Modèles de représentation des GBP

De nombreux formalismes ont été proposés pour la modélisation des GBP. Les plus récents, regroupés sous l'intitulé de réseaux de tâches ou « Task Network Models » (Peleg *et al.*, 2003) tels que Asbru (Shahar *et al.*, 1998), EON (Musen *et al.*, 1996), GLIF (Peleg *et al.*, 2000), Guide (Quaglini *et al.*, 2001), PROforma (Fox *et al.*, 1998), proposent de représenter les recommandations sous la forme de graphes ou d'organigrammes. Ils permettent de coder les connaissances des GBP textuels et de représenter les séquences d'actions recommandées en reproduisant la chronologie idéale décrite dans le GBP.

Afin de proposer une solution au problème de positionnement qui se pose dès lors que l'historique thérapeutique du patient n'est pas vide, certaines approches (EON avec Prodigy III (Johnson *et al.*, 2000), GLIF (Peleg *et al.*, 2000)) ont introduit des points

d'entrée multiples dans la représentation d'un GBP. Ces points d'entrée, appelés « scénarios » ou « états du patient » (*patient states*) représentent des situations types associant des critères cliniques et thérapeutiques (tels que le niveau d'association thérapeutique, mono, bi ou trithérapie, du traitement courant). Ainsi, le positionnement se fait par la détection automatique du scénario ou de l'état du patient qui convient, sur la base de l'appariement entre les éléments du dossier patient et les critères des points d'entrée. Outre l'identification de la situation clinique, ces critères concernent essentiellement le niveau d'association thérapeutique, ce qui permet de court-circuiter les étapes antérieures de la séquence thérapeutique recommandée, et de proposer directement l'étape suivante de traitement. Ces aménagements aux formalismes d'origine constituent une réponse opérationnelle au problème de l'entrée en des points arbitraires du processus de prise en charge modélisé. Mais la caractérisation de l'état du patient (ou du scénario) se fait à un niveau d'abstraction qui ne permet pas de proposer un traitement qui soit réellement adapté à l'historique thérapeutique. Par exemple, de telles approches, en n'explorant pas les contre-indications, ne garantissent pas que certaines propositions de traitement n'intègrent pas des médicaments non tolérés pour un patient donné. De plus, l'exploration systématique de toutes les configurations d'historiques thérapeutiques n'est pas réalisée. Seules les configurations du patient prévues dans les GBP sont représentées, ce qui pose le problème d'un système « muet » pour certains patients.

Nous avons développé le mode guidé du système ASTI dans un objectif ambitieux d'aide à la décision, plutôt que dans l'objectif simple de la stricte diffusion des GBP. Aussi, nous avons étendu les recommandations des GBP afin d'intégrer les accords professionnels pour pouvoir proposer une solution thérapeutique dans toutes les situations cliniques, quelles soient ou non couvertes par l'EBM. La base de connaissances a été représentée sous la forme d'un arbre de décision à 2 niveaux, le niveau clinique permettant la détermination de la situation clinique et le niveau thérapeutique permettant la détermination du meilleur traitement sous contrainte d'un historique thérapeutique personnalisé. Ainsi, le mode guidé d'ASTI propose une solution au médecin qui cherche le meilleur traitement pour un patient dont l'historique thérapeutique, non vide, est discordant avec la séquence des traitements recommandés : l'arbre de décision explore la séquence recommandée (déterminée par l'instanciation de la situation clinique établie au premier niveau de l'arbre de décision) et compare l'historique thérapeutique du patient aux traitements recommandés. La nouvelle décision thérapeutique consiste en pratique à choisir le premier traitement de la séquence recommandée qui n'appartient pas à l'historique thérapeutique sans être contre-indiqué, et qui n'a donc pas encore été prescrit.

3 Méthode

Dans un travail précédent (Bouaud & Séroussi, 2003), nous avons proposé un cadre méthodologique qui permettait d'interpréter, en adoptant des heuristiques propres au domaine thérapeutique, l'énoncé des stratégies thérapeutiques décrites dans un format textuel au niveau des GBP pour aboutir à leur formalisation. Cette première étape nous a permis de caractériser formellement les stratégies thérapeutiques adaptées à tout profil clinique. À partir de cette représentation, nous avons construit « manuellement » une

base de connaissance BC_m , représentée sous la forme d'un arbre de décision à 2 niveaux, permettant le positionnement systématique au sein du GBP de n'importe quel patient, quel que soit son historique thérapeutique.

L'objectif du travail (figure 1) présenté ici est de produire automatiquement par dérivation à partir de l'énoncé formel d'une stratégie thérapeutique, représentée par une séquence de traitements ainsi qu'une liste de substituants, un arbre de décision BC_d équivalent à celui manuellement construit.

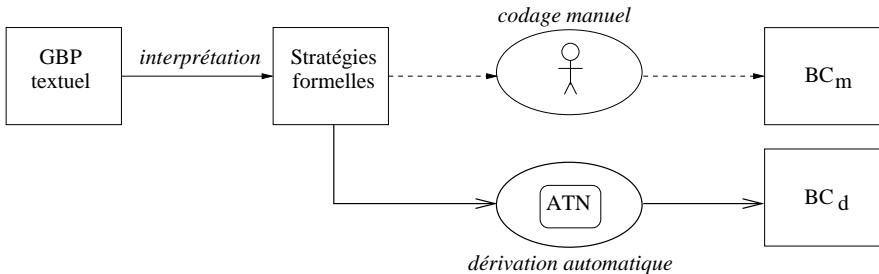


FIG. 1 – Schéma de production des bases de connaissances

3.1 Modèle d'une stratégie thérapeutique

Le modèle de stratégie adoptée ici est une séquence thérapeutique additive. Si le traitement initial consiste à administrer le médicament A et si ce traitement est inefficace, l'étape suivante sera d'ajouter le médicament B au médicament A . Enfin, si cette bithérapie $A + B$ est à nouveau inefficace, on ajoutera C et on prescrira la trithérapie $A + B + C$.

Par ailleurs, en cas de contre-indication à l'un des médicaments recommandés, on dispose de substituants connus. Ainsi, si A est contre-indiqué (allergie) ou si A a été précédemment prescrit et a donné lieu à des effets secondaires inacceptables, A est non toléré, et il peut être remplacé dans le traitement par A' . Puis si A' est à son tour contre-indiqué ou non-toléré, il peut être remplacé par A'' .

Ainsi, une stratégie thérapeutique théorique est modélisée par la séquence $(A, A + B, A + B + C)$ et la liste des substituants des différents composants de la séquence : (A, A', A'') , (B, B', B'') et (C, C', C'') .

3.2 Opérationnalisation d'une stratégie thérapeutique par ATN

L'opérationnalisation d'une stratégie thérapeutique peut se décomposer en plusieurs étapes. On distingue deux cas. Dans le cas d'un traitement initial, le traitement recommandé est la première étape de la séquence thérapeutique (aux contre-indications près). Si le patient a déjà été traité, il faut pouvoir évaluer le traitement courant et, s'il est inefficace ou non toléré, le modifier et proposer un nouveau traitement, sous contrainte des recommandations mais également de l'historique thérapeutique du patient.

On choisit de programmer la mise en œuvre d'une séquence thérapeutique par un ATN dont les sous-graphes permettent de traiter les différentes sous-procédures.

3.2.1 Proposition d'un traitement initial

L'objectif est de proposer le meilleur traitement non contre-indiqué pour le cas clinique identifié. L'automate correspondant à la situation décrite par la Figure 2 peut se décomposer en un graphe principal et un sous-graphe. Le graphe principal est un graphe linéaire, modélisé par le premier automate, qui reprend les étapes nécessaires à la proposition du traitement recommandé. Le traitement initial A , détecté (action 1) par la lecture de la séquence recommandée ($A, A + B, A + B + C$) mémorisée dans une liste ordonnée, est proposé. Ce traitement candidat recommandé par le GBP doit néanmoins, au préalable, être évalué pour le patient afin de vérifier qu'il n'est pas contre-indiqué (Assure_tt_non_ci). Cette deuxième partie est modélisée par un automate qui vérifie si le traitement A est contre-indiqué ou non. Si A est contre-indiqué, les substituants de A , donnés par la liste (A, A', A''), sont évalués (action 4). Dès qu'un substituant est non contre-indiqué, il est retourné au graphe principal, et constitue le traitement recommandé. Si aucun substituant ne convient, le praticien est informé du fait qu'aucun traitement toléré ne peut être proposé.

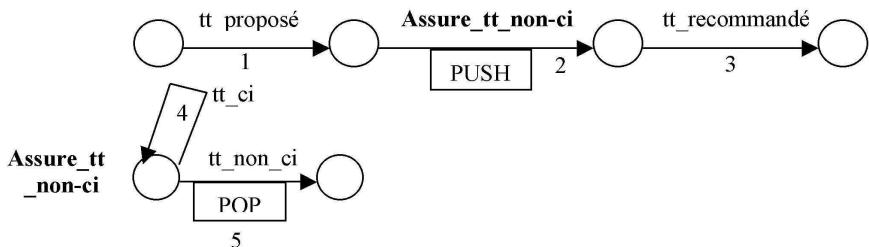


FIG. 2 – Automate de la prescription d'un traitement initial

3.2.2 Proposition d'un traitement non initial

Dans le cas où le patient est déjà traité, si le traitement courant est non toléré ou inefficace, il s'agit de proposer un nouveau traitement qui serait efficace et non contre-indiqué. La proposition d'un traitement non initial nécessite d'être modélisée par un automate plus complexe, représenté par la figure 3. Comme précédemment, un automate principal décrit les étapes d'analyse de la réponse du patient au traitement courant afin de proposer le meilleur traitement recommandé. La première étape est d'évaluer si le traitement courant a été toléré. Ensuite, on teste l'efficacité. Chacun de ces 2 contrôles est assuré par un sous-graphe. Dans chacun des 2 cas, on doit s'assurer que les traitements proposés en remplacement ne sont pas eux-mêmes contre-indiqués, ce qui nécessite d'appeler le sous-graphe de la figure 2 (Assure_tt_non_ci).

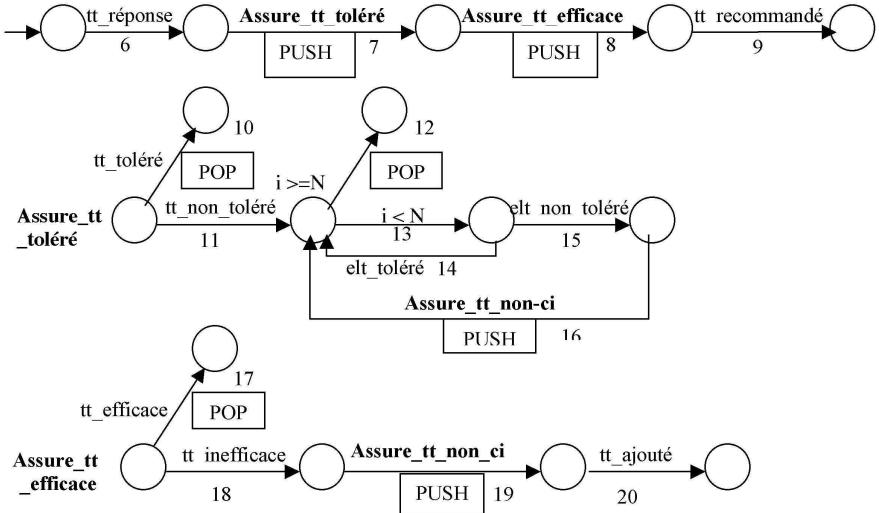


FIG. 3 – Automate de la réponse à un traitement prescrit

Si le traitement courant est toléré, le sous-graphe *Assure_tt_toléré* ne propose aucune modification et c'est le même traitement qui est proposé en sortie.

Si le traitement courant est non toléré, on doit étudier un à un les différents médicaments qui interviennent dans sa composition afin d'établir s'ils sont, chacun, tolérés ou non. Les médicaments non tolérés sont substitués (action 15) en s'assurant que les substituants proposés ne sont pas contre-indiqués. Le sous-graphe produit ainsi un traitement équivalent en terme de niveau d'association médicamenteuse (on reste en mono, bi, trithérapie si le traitement courant était en mono, bi, trithérapie).

Si le traitement courant est inefficace, la recommandation est d'ajouter une classe médicamenteuse et d'augmenter ainsi le niveau d'association du traitement. Ainsi, par l'action 18, on va proposer le premier élément de la classe thérapeutique recommandée en addition au traitement courant. En effet si le traitement courant est une monothérapie de la classe (A, A', A''), on proposera une bithérapie en associant un médicament de la classe (B, B', B''), aux contre-indications près. Ainsi, il faut parcourir le sous-graphe s'assurant que le médicament ajouté est supporté par le patient avant de le prescrire. Une fois que le bon complément est identifié, on l'ajoute à la prescription courante et on recommande ce nouveau traitement.

3.3 L'algorithme de positionnement

L'algorithme de positionnement définit les opérations à effectuer pour pouvoir placer un patient en fonction de son historique thérapeutique au sein de la stratégie recommandée représentée par l'ATN afin de prescrire le meilleur traitement adapté à son état.

Si l'historique thérapeutique est vide, on utilise l'ATN correspondant au traitement initial. Si le patient a un historique thérapeutique non vide, on connaît sa réponse aux

traitements qui ont déjà été administrés, en termes de tolérance et d'efficacité. Ainsi on dispose de la séquence ordonnée des traitements déjà administrés, d'une liste des médicaments contre-indiqués, et d'une liste des traitements inefficaces. Le principe est alors d'explorer la séquence thérapeutique recommandée afin d'identifier le premier traitement recommandé n'appartenant pas à la séquence des traitements déjà administrés de l'historique thérapeutique, sans que cela ne soit le fait de contre-indications ou d'une inefficacité.

Si l'historique thérapeutique est conforme à la séquence recommandée, les étapes antérieures de traitement du patient reprennent exactement celles de la séquence recommandée et il suffit de localiser l'étape courante dans la stratégie thérapeutique recommandée et de simplement proposer l'étape suivante. Si l'historique thérapeutique n'est pas conforme à la séquence recommandée, la séquence recommandée est de la même manière chronologiquement explorée, et pour chaque étape de traitement recommandé, il faut tester si le traitement n'a pas déjà été prescrit, et alors le proposer en traitement recommandé sauf s'il est contre-indiqué (exploration de la liste des médicaments contre-indiqués).

Dans tous les cas, la liste des médicaments contre-indiqués est réactualisée par la réponse du patient au traitement courant. De même, la liste des traitements inefficaces est réactualisée en intégrant les mono, bi, et/ou trithérapies inefficaces, ainsi que les traitements dérivés de moindre niveau de combinaison. Par exemple, si le traitement $B + C$ a été prescrit et est inefficace, alors que ni B , ni C n'ont été prescrits, on peut déduire que B , C , et $B + C$ sont inefficaces.

3.4 La dérivation de l'arbre de décision

L'automate prenant en compte une stratégie autorise un changement du type de représentation produite. En pratique, nous avons utilisé l'ATN pour dériver automatiquement l'arbre de décision représentant la stratégie thérapeutique recommandée et permettant de produire le meilleur traitement quelque soit l'historique thérapeutique du patient. Pour cela on enrichit l'automate de la stratégie d'action pouvant construire l'arbre. Il est important de noter que dans le cas du parcours de l'arbre de décision afin de proposer le traitement adéquat, on effectue des choix dans le franchissement des étapes guidés par l'existence de contre-indications, de non-tolérance, ou d'inefficacité. Il nous faut ainsi explorer toutes les branches du graphe et donc pour chaque transition effectuer une exploration en profondeur.

Le premier niveau de profondeur de l'arbre est représenté par le nœud permettant d'établir s'il s'agit d'un traitement initial ou non. L'arbre de décision doit ensuite présenter l'ensemble des possibilités de prescription pour un traitement donné correspondant à une étape de la stratégie. Ainsi dans le cas d'une réponse, on dérivera le graphe pour chaque niveau de thérapie. Quel que soit le graphe parcouru, l'ensemble des possibilités de parcours de celui-ci est dérivé. Par exemple, dans le cas de l'arbre de décision représentant la stratégie correspondant à une réponse, on étudie la branche « tt_toléré » du graphe, puis successivement les branches « tt_efficace » et dans cette dernière branche, il faudra dériver toutes les branches du sous-graphe qui assure que le traitement n'est pas contre-indiqué « assure_tt_non_ci ». Enfin on dérivera la branche

« tt_non_toléré ».

Dans le cas d'un exemple simplifié de la prise en charge de l'HTA, la séquence recommandée consiste à administrer des inhibiteurs de l'enzyme de conversion (IEC), puis à ajouter d'un diurétique de l'anse (DA) en cas d'inefficacité de la monothérapie. Sachant que les inhibiteurs des récepteurs de l'angiotensine II (IRA2) sont le substituant de référence des IEC, la stratégie thérapeutique est représentée alors par (*IEC, IEC + DA*) avec les classes de substitution (*IEC, IRA2*) et (*DA*). La figure 4 illustre un extrait de l'arbre de décision produit.

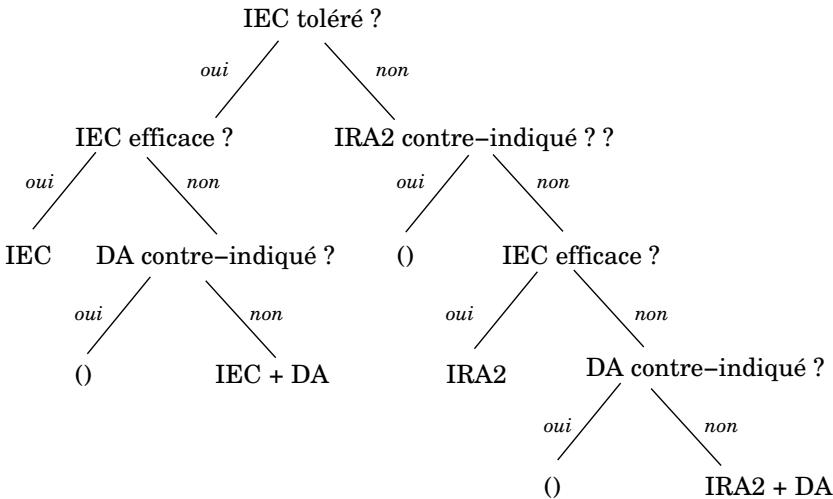


FIG. 4 – Extrait d'un arbre de décision généré

4 Discussion

L'objectif de ce travail est de proposer une méthode permettant d'automatiser la production d'une base de connaissances structurée sous la forme d'un arbre de décision afin d'opérationnaliser la détermination d'un traitement médicamenteux sous contrainte d'un GBP et d'un historique thérapeutique patient dans le cadre de la prise en charge des maladies chroniques. Une formalisation des séquences de traitements recommandées par le GBP, établie sous contrainte d'un modèle d'interprétation du texte, a été utilisée comme représentation de départ pour la production de l'arbre de décision.

4.1 Limites méthodologiques

L'algorithme de positionnement du patient au sein de la stratégie thérapeutique, n'a pas été testé sur un panel suffisamment important d'exemples. En particulier, nous n'avons développé l'algorithme que dans le cas particulier d'une séquence thérapeutique de la forme (*A, A + B, A + B + C*). Or dans la prise en charge de l'HTA, il

existe des séquences du type (A ou B , $A + C$ ou $B + C$, $A + B + C$), ou du type (A , $A + B$ ou $A + C$, $A + B + C$). Les développements permettant le traitement de ce type de séquences n'ont pas été réalisés.

La dérivation automatique de l'arbre de décision, qui doit s'effectuer à partir du modèle des ATN, a du être complétée pour prendre en compte la quasi-totalité des situations possibles. Par exemple, il a fallu introduire des niveaux de profondeur « artificiels » pour la construction de critères correspondant à la caractérisation du niveau de thérapie du traitement courant (mono, bi et trithérapie). Une fois ce niveau identifié, l'objectif était de tester si le traitement recommandé pour ce niveau d'association avait été déjà prescrit, et si oui, s'il était toléré et efficace, sinon pourquoi il n'avait pas été prescrit, et déterminer s'il était contre-indiqué ou s'il avait été oublié.

4.2 Comparaison avec ASTI

Pour des séquences médicalement pertinentes, nous avons comparé l'arbre de décision généré par la méthode automatique avec celui construit manuellement dans le cadre du projet ASTI. Par exemple, dans la situation clinique particulière pour laquelle la stratégie thérapeutique recommandée correspond à la séquence ($IEC, IEC + DA, IEC + DA + BB$), toujours avec les classes de substitution ($IEC, IRA2$), (DA) et ($BB, DILT$), le tableau 1 fournit des éléments de comparaison quantitatifs.

TAB. 1 – Comparaison quantitative de l'arbre ASTI manuel et de son équivalent généré par ATN

	Méthode manuelle	Méthode automatique
Nombre de noeuds	63	131
Nombre de chemins	33	67
Moyenne des niveaux de profondeur	7,03	8,84

Les arbres générés par l'ATN sont plus volumineux que ceux qui sont construits dans ASTI du point de vue de la représentation interne. Ceci provient du fait que la dérivation automatique procède de façon systématique ce qui conduit à la production de nœuds logiques mais non valides sur le plan médical qui n'ont pas été créés au cours de la construction manuelle des arbres thérapeutiques d'ASTI. C'est le cas, par exemple, d'un médicament non toléré ou contre-indiqué pour lequel il n'existe pas de substituant dans le GBP. Mais cela peut être également le cas lorsque par exemple un traitement, composé de N médicaments ($A + B + C$, pour $N = 3$), n'est pas toléré par le patient, mais que lors de la dérivation de l'arbre à partir du graphe si les $N - 1$ composants sont tolérés (A toléré, B toléré) on va de manière automatique chercher à dériver la possibilité que le N^{e} composant soit toléré. La dichotomie des réponses en OUI/NON conduit à construire un chemin impossible (C toléré) puisque $A + B + C$ non toléré, avec A toléré et B toléré implique forcément que C n'est pas toléré.

Cette différence vient de la démarche de dérivation adoptée. Dans ASTI, on ne vérifie pas systématiquement, comme c'est le cas avec la méthode automatique, si les traite-

ments sont efficaces ou tolérés. On ne contrôle ces critères que s'il existe un moyen de résoudre le problème, c'est-à-dire respectivement s'il y a des classes médicamenteuses additionnelles (pour augmenter le niveau d'association médicamenteuse) ou s'il existe des substituants recommandés. De ce fait les arbres dérivés de l'ATN permettent de mieux visualiser les choix qui nous amènent à déterminer le nouveau traitement recommandé mais présentent le désavantage de parfois doubler le nombre de chemins réellement exploitables.

Enfin, il y a des différences dans l'ordre selon lequel les critères de l'arbre de décision sont organisés. Dans l'arbre dérivé à partir de l'ATN, comme la génération s'effectue automatiquement il y a des situations où l'ordre des critères qu'il faut instancier pour accéder à la recommandation thérapeutique ne suit pas la logique du praticien. En effet, lorsque le traitement recommandé $A + B$ n'a pas été prescrit, on vérifie avant de le prescrire que chacun de ses constituants n'est pas contre-indiqué. La logique médicale voudrait que, lorsqu'un constituant A est contre indiqué et donc substitué par A' , le praticien contrôle si ce nouveau traitement recommandé a déjà été prescrit ($A' + B$ déjà prescrit ?), avant de poursuivre. Or, lors de la génération automatique on vérifiera d'abord si les constituants sont non contre-indiqués avant de chercher le nouveau traitement recommandé (A' non contre-indiqué, B non contre-indiqué, $A' + B$ déjà prescrit ?).

5 Conclusion

L'utilisation des GBP dans les systèmes d'aide à la décision passe par une étape de formalisation des textes suivie d'une étape de modélisation.

Les ATN sont des graphes particuliers qui permettent de modéliser une grande variété de stratégies. Ces graphes permettent notamment de modéliser des problèmes récursifs ou de découper un problème en sous-problèmes. C'est d'ailleurs cette dernière propriété des ATN qui nous a intéressés pour la représentation de l'exploration des contre-indications d'un médicament a priori recommandé par le GBP mais pas forcément adapté pour un patient donné, compte tenu de son historique thérapeutique, un problème qui intervient à chaque étape de la stratégie avant de prescrire un nouveau traitement. Les ATN que nous avons programmés se limitent actuellement au cas particulier des stratégies linéaires ($A, A + B, A + B + C$) où il n'y a pas, à une étape donnée, le choix entre des traitements différents qui ne sont pas des substituants l'un de l'autre.

Nous avons également développé une méthode de dérivation automatique d'un arbre de décision mais cette dérivation ne s'effectue actuellement qu'à partir de nos graphes qui ne représentent que partiellement la stratégie décrite dans les GBP.

Il apparaît ainsi qu'il est nécessaire de complexifier l'ATN construit afin d'étendre la modélisation aux stratégies thérapeutiques incluant des OU dans la séquence. De plus, il faudra également pouvoir confronter les arbres de décision engendrés à partir du nouveau graphe avec un nombre important d'arbres thérapeutiques ASTI pour évaluer correctement la dérivation automatique.

Références

- BOUAUD J. & SÉROUSSI B. (2003). Un cadre pour l'interprétation et l'opérationnalisation de l'expression textuelle des stratégies thérapeutiques. In R. DIENG-KUNTZ, Ed., *Actes des 14^{es} Journées Ingénierie des Connaissances*, p. 35–50, Laval, France : Presses universitaires de Grenoble.
- FOX J., JOHNS N. & RAHMANZADEH A. (1998). Disseminating medical knowledge : the PROforma approach. *Artif Intell Med*, **14**(1,2), 157–182.
- JOHNSON P. D., TU S., BOOTH N., SUGDEN B. & PURVES I. N. (2000). Using scenarios in chronic disease management guidelines for primary care. *J Am Med Inform Assoc*, **7**(suppl), 389–393.
- MUSEN M. A., TU S. W., DAS A. K. & SHAHAR Y. (1996). EON : a component-based approach to automation of protocol-directed therapy. *J Am Med Inform Assoc*, **3**(6), 367–388.
- PELEG M., BOXWALA A. A., OGUNYEMI O., ZENG Q., TU S., LACSON R., BERS-TAM E., ASH N., MORK P., OHNO-MACHADO L., SHORTLIFFE E. H. & GREENE R. A. (2000). GLIF3 : The evolution of a guideline representation format. *J Am Med Inform Assoc*, **7**(suppl), 645–649.
- PELEG M., TU S. W., BURY J., CICCARESE P., FOX J., GREENES R. A., HALL R., JOHNSON P. D., JONES N., KUMAR A., MIKSCH S., QUAGLINI S., SEYFANG A., SHORTLIFFE E. H. & STEFANELLI M. (2003). Comparing computer-interpretable guideline models : a case-study approach. *J Am Med Inform Assoc*, **10**(1), 52–68.
- QUAGLINI S., STEFANELLI M., LANZOLA G., CAPORUSSO V. & PANZARASA S. (2001). Flexible guideline-based patient careflow systems. *Artif Intell Med*, **22**(1), 65–80.
- SHAHAR Y., MIKSCH S. & JOHNSON P. (1998). The Asgaard project : a task-specific framework for the application and critiquing of time-oriented guidelines. *Artif Intell Med*, **14**(1,2), 29–52.
- SÉROUSSI B., BOUAUD J. & ANTOINE E.-C. (2001a). OncoDoc, a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artif Intell Med*, **22**(1), 43–64.
- SÉROUSSI B., BOUAUD J., DRÉAU H., FALCOFF H., RIOU C., JOUBERT M., SIMON C., SIMON G. & VENOT A. (2001b). ASTI, a guideline-based drug-ordering system for primary care. In V. L. PATEL, R. ROGERS & R. HAUX, Eds., *Medinfo*, p. 528–532.

Les annotations pour gérer les connaissances du dossier patient

Sandra Bringay¹, Catherine Barry¹, Jean Charlet²

¹ Laboratoire LaRIA, Université d'Amiens,
`{sandra.bringay,catherine.barry}@u-picardie.fr`
² Laboratoire STIM/DSI/AP-HP, Paris,
`Jean.Charlet@spim.jussieu.fr`

Résumé : Les praticiens ne disposent toujours pas d'outils informatiques de gestion du dossier patient hospitalier leur permettant de reproduire toutes les pratiques qu'ils réalisent avec le dossier papier. En nous plaçant dans le paradigme d'une approche documentaire, nous adoptons une vision originale sur les documents en les considérant munis de leurs annotations. L'objectif de cette publication est de montrer l'intérêt d'une sémantique hypertextuelle annotationnelle pour travailler sur les annotations du dossier patient et les fonctionnalités qui en découlent (édition de documents, filtrage, message, etc.).

Mots-clés : Dossier patient, Document, Annotation, Gestion des connaissances

1 Introduction

Les praticiens utilisent traditionnellement un ensemble de documents papier, le dossier patient, pour véhiculer des connaissances médicales. Ce dossier montrant désormais ses limites, de nombreuses équipes ont travaillé sur son informatisation depuis les années 80. Au cœur de ces travaux, on retrouve la problématique de la distribution des connaissances au sein d'une organisation. En effet, de nombreuses catégories de praticiens doivent avoir accès aux connaissances médicales mais leurs objectifs, leurs missions sont très différents. Il est donc difficile de leur proposer les connaissances appropriées, dans le format adéquat, au moment opportun.

Le schéma 1 adapté de (Charlet, 2003), résume nos hypothèses de travail. Il représente les flux de connaissances dans le système de soin. Nous avons placé au centre le dossier patient, véritable noyau de la mémoire médicale. Il s'agit d'une collection de documents semi-structurés et structurés regroupant les connaissances nominatives d'un patient. Des techniques de traitement automatique de la langue TAL (Badr *et al.*, 2003) peuvent être appliquées pour en extraire des connaissances non nominatives et structurées. Dans leur mission de recherche, les praticiens les utilisent pour produire des études épidémiologiques, dont sont issus les protocoles et guides de bonnes pratiques GBP (utilisés pour les soins sous forme de documents semi-structurés Séroussi *et al.*, 2001). Des connaissances non nominatives et structurées sont utilisées pour la gestion médico-économique. Les connaissances nominatives du dossier patient peuvent être présentées à d'autres acteurs (patient, réseaux de soins).

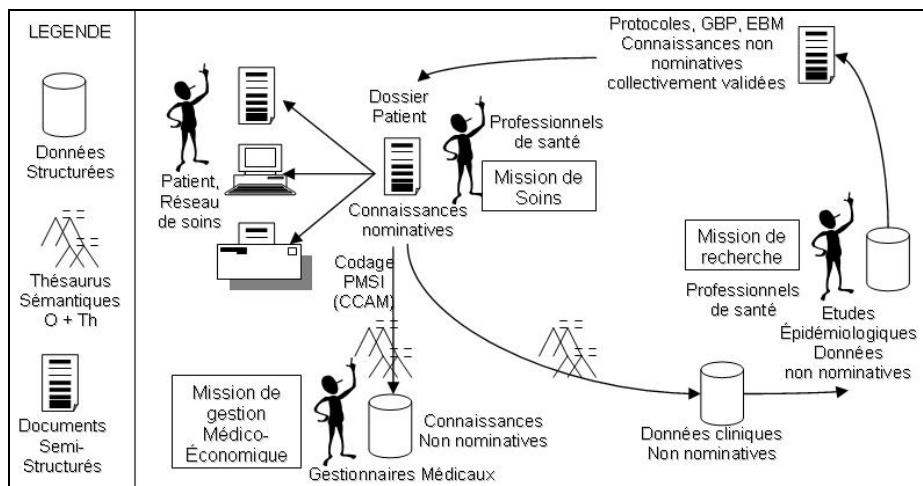


Fig. 1 – Les flux de connaissances médicales dans le système de soins

Ce schéma souligne l'importance des documents du dossier patient pour agir (Zacklad, 2004), pour soigner le patient, pour construire des études épidémiologiques, pour prendre des décisions de gestion médico-économique, etc. Ce schéma met aussi en évidence les difficultés rencontrées pour relier les outils de soins aux outils de management et de recherche. En effet, d'un côté on trouve les praticiens, dans leur mission de soin, qui manipulent des connaissances contextuelles, nominatives et semi-structurées au travers de documents. De l'autre côté, on trouve d'autres acteurs qui manipulent, pour la recherche médicale et la gestion, des connaissances non nominatives et structurées¹. Cette constatation explique d'ailleurs l'échec de certains dossiers patients informatisés. Centrés sur les données, ces outils sont bien adaptés aux problématiques de gestion et de recherche, mais ils restent inadaptés aux problématiques de soins. Or les principaux collecteurs de connaissances sont les praticiens qui, dans le cadre des pratiques médicales, réalisent les saisies. Ces premières remarques justifient le choix d'une approche documentaire pour informatiser les outils de soins. Notre position est confortée dans la littérature par bien d'autres auteurs (Bachimont, 2004, Berg, 1999, Charlet, 2003, Lovis *et al.*, 2003).

Depuis 2002, nous participons au projet DocPatient², dans lequel nous cherchons à informatiser le dossier patient hospitalier selon une approche documentaire. Ce projet est mené en collaboration avec un site pilote³ et un partenaire industriel⁴. Nous travaillons sur les fonctionnalités documentaires facilitant la manipulation des documents électroniques et en particulier sur les annotations.

¹ Les praticiens sont pour la plupart impliqués à la fois dans les activités de soins et de recherche.

² Ce projet de l'université d'Amiens, financé par le programme Homme Technologie et Systèmes Complexes HTSC, regroupe une équipe pluridisciplinaire composée des sciences pour l'ingénieur (informatique) et des sciences humaines et sociales (droit, gestion, psychologie).

³ Le service Réanimation Pédiatrique et de Néonatalogie dirigé par le Docteur G. Krim du CHU d'Amiens.

⁴ La société UNI-MEDICINE <http://www.uni-medecine.com/>

La démarche adoptée dans le projet a été la suivante. Afin de concevoir une vision consensuelle sur la notion de document et en nous basant sur les travaux de Pedauque⁵, nous avons mené une recherche théorique sur la notion de document appliquée aux documents médicaux (Bringay *et al.*, 2004a). Nous avons réalisé une étude multidisciplinaire sur le terrain (dans le service de Pédiatrie) pour construire le cahier des charges du projet. Nous avons particulièrement travaillé sur l'ensemble des supports de connaissances (dossiers patient papier et informatisés, notes personnelles) pour transformer ces ressources en un seul dossier informatisé.

De plus, l'étude des documents papier a montré que ceux-ci sont souvent annotés. Ces annotations contiennent des connaissances pertinentes à conserver dans le dossier. Nous avons alors mené une recherche théorique sur la notion d'annotation, une étude des outils électroniques d'annotations⁶, ainsi qu'une étude pratique des situations d'annotations dans le dossier patient (Bringay *et al.*, 2004b). Au regard de ces travaux, nous affirmons que les annotations sont une des solutions envisageables pour résoudre certains problèmes rencontrés par les praticiens lorsqu'ils manipulent les documents électroniques du dossier patient. Nous ne limitons pas notre étude aux annotations utilisées pour indexer les documents, ni à celles produites lors de processus de conception collaborative.

L'objectif de cet article est de montrer que les commentaires laissés au cours de l'écriture et de la lecture des documents électroniques du dossier patient, peuvent être utilisés pour faciliter la gestion des connaissances. Dans la section 2, nous décrivons les documents médicaux et leurs limites. Ainsi, nous soulignons un certain besoin d'annotation. Dans la section 3, nous exposons des premières réflexions sur une sémantique hypertextuelle annotationnelle.

2 Les documents du dossier patient : nécessité des annotations

2.1 Deux types de documents dans le dossier patient

Dans un service hospitalier, les *situations de transaction*⁷ entre les praticiens sont nombreuses. Elles sont réalisées par un grand nombre de *réaliseurs* pour de multiples *bénéficiaires* (le patient, les médecins, les infirmières, les laborantins). Une majorité de ces transactions sont orales. Toutefois, par tradition, les praticiens ont développé une véritable culture de l'écrit. Afin de conserver des traces d'une grande partie des connaissances élaborées, ils les transcrivent ou les enregistrent sur des

⁵ Nous nous référions ici au travail de fond sur la notion de document mené dans le cadre du réseau thématique pluridisciplinaire STIC 33 « Documents et contenu : création, indexation, navigation », sous la signature collective « Pedauque ». Site du projet : <http://rtp-doc.ensib.fr/> Document http://archivesic.ccsd.cnrs.fr/sic_00000511.html

⁶ iMarkup (<http://www.imarkup.com>), Xlibris (<http://www.fxpal.com/xlibris>), Anchored conversation (FXPAL Palo Alto), Annotea (W3C project), TheBrain (www.mines.inpl-nancy.fr/~tisseran/tsie/02-03/etudes/thebrainssite)

⁷ Nous empruntons les termes de (Zacklad, 2004), qui s'appuie sur les théories transactionnelles pour décrire les documents médicaux comme des documents pour l'action. Ces termes sont en italique dans cette publication.

supports *pérennes*, les documents papier ou électroniques du dossier patient. Ces connaissances sont alors manipulées (complétées, annotées, lues) par les praticiens, réactivées dans différents contextes pour être le support de nouvelles transactions.

Vu le nombre et la complexité des situations de transaction, les unités hospitalières ont mis en œuvre un véritable processus de *documentarisation*, pour faciliter la gestion des connaissances contenues dans les documents. Les chefs de service ont conçu l'architecture de leur dossier et de certains documents. Cela facilite *a*) la gestion des documents : dans le dossier patient papier, on sait où trouver les comptes rendus opératoires ; *b*) leur manipulation physique : grâce au plan préétabli du compte rendu d'hospitalisation CRH, un rédacteur sait où chercher le paragraphe à compléter et un lecteur sait où trouver le paragraphe contenant la connaissance désirée.

Pour organiser les connaissances dans un document, le chef de services (concepteur) le structure et donne des indications dans des intitulés sur les connaissances à y consigner. L'articulation interne⁸ du document est alors explicite. Cette décomposition en fragments hiérarchisés et mis en forme, met en relief les productions sémiotiques et donne du sens à leur ordonnancement. Les praticiens (rédacteurs) analysent les indications du concepteur pour compléter les documents. Dans le dossier patient, on trouve surtout deux grandes sortes de documents :

Les *formulaires rédigés en temps réel*, au pied du lit du patient contiennent essentiellement des connaissances brutes liées à des situations de soins stéréotypées. Le concepteur structure le document finement et donne des indications très précises dans les intitulés des champs sur les connaissances à y consigner. Le rédacteur interprète ces indications pour compléter les champs. Ces formulaires, très structurés, permettent d'identifier rapidement de petites productions sémiotiques et les liens très précis qui les unissent. Cela facilite le travail d'écriture du rédacteur (sa saisie se limite à quelques mots ou segments de phrases) et celui des lecteurs (ils ont appris, pendant leurs études, à rechercher des connaissances dans ces documents). Un exemple de formulaire est la « fiche administrative d'entrée » du patient.

Les *documents de synthèse rédigés a posteriori de l'acte médical* contiennent les interprétations des praticiens. Ici aussi, on retrouve deux réalisateurs. La trame du document, élaborée par le concepteur, structure le document en paragraphes (et non en champs comme dans les formulaires). Le rédacteur dispose alors d'indications sur le contenu des paragraphes, mais il reste libre des connaissances qu'il y consigne. En général, il utilise la langue naturelle. Le concepteur en imposant une trame, veut organiser au maximum l'écriture du rédacteur pour qu'elle soit exploitable. Cette trame ne peut pas être aussi précise que celles des formulaires car il est impossible de prédefinir les connaissances issues d'une réflexion. Ces documents sont donc semi-structurés. Ils permettent d'identifier de grandes productions sémiotiques et les liens qui les unissent. Un exemple de document de synthèse est le CRH.

Bien sûr, il existe aussi des documents rédigés sans « modèle » prédéfini par un concepteur, tel que le schéma du chirurgien improvisé pour expliquer son opération au patient, mais ces documents sont rares. Finalement, nous pouvons dire que ces deux

⁸ Nous avons décrit dans (Bringay *et al.*, 2004a) l'articulation interne du document (la structure logique Bachimont 2004), c'est-à-dire la décomposition en fragments (en sous productions sémiotiques) et leur ordonnancement qui est lui-même source de signification.

catégories de documents s'opposent par le type de connaissances que l'on y consigne (des connaissances prédefinies *vs* des connaissances imprévisibles), par le type d'écriture réalisée par le rédacteur (une écriture contrainte *vs* une écriture plus libre) et par le niveau de structure du document (des documents structurés *vs* des documents semi structurés). (Bachimont, 2004) affirme que ce sont ces genres textuels prédefinis qui fixent les règles d'écriture et de lecture. Ce sont eux qui permettent des lectures dans des contextes distants dans le temps et l'espace de la rédaction.

2.2 Nécessité des annotations

Malgré l'effort de documentarisation des autorités médicales pour simplifier l'écriture et la lecture des documents, ceux-ci ne sont pas suffisants pour permettre aux praticiens de réellement travailler sur toutes les connaissances qu'ils élaborent. Ces difficultés sont accentuées par le passage au numérique, mais certaines existent déjà sur papier. Les annotations sont alors une solution pour résoudre une partie de ces problèmes. D'ailleurs, les praticiens les utilisent déjà dans les documents papier.

L'étude des documents papier a montré que même si le concepteur laisse des champs texte pour les connaissances non prévisibles, les rédacteurs vont préférer annoter les formulaires rigides pour l'écriture. Avec un moyen graphique (une flèche, une partie surlignée), ils relient le commentaire et la partie du document ayant suscité le commentaire. Les documents de synthèse sont aussi annotés. Il ne s'agit pas de commentaires laissés par le rédacteur car celui-ci est libre de rédiger dans les paragraphes. Ce sont des annotations, laissées par les lecteurs pour garder des traces de leur interprétation. Les annotations permettent donc de *contextualiser* les connaissances non envisagées par le concepteur des documents, produites pendant l'écriture et la lecture. À la manière de Pedauque, nous étendons l'équation donnée dans (Bringay *et al.*, 2004a) :

Document annoté (formulaire ou synthèse)= structure et connaissances du concepteur (dans les intitulés des champs ou des paragraphes) + connaissances du rédacteur (dans les annotations s'il s'agit d'un formulaire et dans le document) + connaissances du lecteur (dans les annotations)

Cette équation rend compte de l'insertion du lecteur dans le processus constitutif du document. Comme l'explique (Bachimont, 2004), la pratique d'annotation est donc une manière, pour le lecteur, de se réapproprier le document, de le réécrire selon l'usage désiré. Il devient ainsi l'« auteur de sa lecture » (Bachimont, 2004).

Les annotations permettent aussi aux praticiens de lier plusieurs documents, de guider la lecture d'un document vers un autre. Par exemple, pour justifier son argumentation dans un compte rendu d'imagerie, l'expert ajoute « cf. radio thorax 2 ».

Par ailleurs, l'étude du dossier patient papier a montré que certains documents résultait de la combinaison de plusieurs ressources (des annotations ou des parties de documents). Prenons l'exemple d'un médecin qui désire rédiger la partie « Motif d'hospitalisation » et « Histoire de la maladie » du dossier d'entrée. Il lit le dossier de la maternité et y recherche des connaissances relatives aux antécédents familiaux du nouveau-né. Il annote les points importants et les rassemble dans les paragraphes concernés. Il réécrit le tout pour construire quelque chose de compréhensible. Dans un

tel scénario, le lecteur devient un lecteur-scripteur (Stiegler, 2000) car il réalise deux tâches : la lecture des documents qu'il utilise pour l'écriture du nouveau document.

Pour finir, les annotations sont souvent le support de communications informelles. (Hardstone *et al.*, 2004) a montré que les praticiens les utilisent pour consigner les connaissances souvent partielles, provisoires, incrémentales qu'ils ne veulent pas écrire dans les documents du dossier car ils considèrent ceux-ci comme trop publics (distribués à une large audience) et trop formels (règles d'écriture du concepteur).

De ces exemples, nous pouvons conclure qu'une personne annote car :

- *elle ne peut pas* sans cela ajouter sa production sémiotique au document. C'est le cas des formulaires trop figés pour que le rédacteur puisse y ajouter des connaissances non prévues par le concepteur. Les annotations sont donc une solution si l'on ne dispose pas de méthode courante pour étendre les formulaires.
- *elle ne veut pas* ajouter sa production sémiotique au document car celle-ci est écrite avec une intention de communication différente de l'intention initiale du document annoté. Dans l'annotation, elle ajoute une méta-information, i.e. une information à propos du document plutôt qu'une information dans le document. C'est le cas lorsqu'un lecteur annote pour garder des traces de sa lecture ou bien lorsqu'une personne cherche à construire un nouveau document à partir de ses lectures. Les annotations sont donc une solution si l'on ne dispose pas de méthode courante pour coder les commentaires à propos des documents.

Ces exemples montrent en quoi les praticiens utilisent les annotations pour agir : soit pour enrichir le document annoté, soit pour être le support transitoire de connaissances utilisées pour créer de nouvelles connaissances inscrites ou non dans un nouveau document. Annoter c'est donc déjà une action en soi.

2.3 Qu'est ce qu'une annotation ?

Notre définition d'une annotation élaborée à partir de (Denoue *et al.*, 2003, Soubrié, 2001, Stiegler, 2000, Zacklad, 2004) est :

Une annotation est une note particulière attachée à une cible par une ancre. La cible peut être une collection de documents, un document, un segment de document (un paragraphe, un groupe de mots, une image, une partie d'image, etc.) ou bien une autre annotation. Chaque annotation possède un contenu matérialisé par une inscription. Cette dernière est une trace de la représentation mentale élaborée par l'annotateur à propos de la cible. Le contenu de l'annotation peut être interprété par un autre lecteur. L'ancre lie l'annotation à la cible (une ligne, une phrase surlignée, etc.)

Le contenu de l'annotation permet à l'annotateur de transmettre un message. Il est source de signification pour lui et pour le lecteur. Il peut apparaître dans le document (dans la marge, entre deux lignes), en dehors du document (un post-it collé sur un document) ou bien être fusionné avec l'ancre (un passage surligné). Dans les documents papier, le contenu d'une annotation peut être présenté sous forme textuelle

(commentaire) ou sous forme typographique (passage surligné). Avec un support électronique, les contenus peuvent être autres : audio, avec une image fixe ou animée, multimédia, sous forme d'un lien hypertexte vers un autre document, etc.

Comme précisé dans la section 2.1, l'orientation du lecteur dans un document est liée à l'articulation des fragments. La seule articulation perceptible qui permet de guider le lecteur de la cible (zone annotée) vers le contenu de l'annotation est l'ancre. Elle établit le contexte en liant le contenu à la cible. Par exemple, si un annotateur souligne un mot dans un texte et ajoute un commentaire dans la marge, nous savons quelle partie du document est liée au commentaire.

3 Une sémantique hypertextuelle annotationnelle

Ayant justifié l'intérêt d'une fonctionnalité d'annotation dans le dossier patient, nous nous intéressons maintenant aux traitements que l'on peut leur appliquer, pour améliorer les fonctionnalités de navigation hypertextuelle et la manipulation (création, lecture) des documents électroniques du dossier patient informatisé. L'objectif d'une telle spécification est de fournir à l'utilisateur un composant logiciel, un système d'annotation, complétant un système de gestion des dossiers patient informatisé. Pour cela, nous définissons une sémantique hypertextuelle annotationnelle. Cela consiste à étudier la signification des annotations, pourquoi sont-elles utilisées, ce qu'elles permettent d'exprimer, leurs propriétés et les traitements que l'on peut leur appliquer en fonction de ces propriétés (combinaisons potentiellement valides, actions, etc.)

Pour tester la liste des traitements et des propriétés des annotations élaborée, nous avons réutilisé une première maquette informatique mise en oeuvre par notre partenaire industriel pour présenter l'approche documentaire dans le projet DocPatient (maquette du dossier papier toujours utilisé dans notre site pilote). Nous avons complété cette maquette en implémentant des traitements à réaliser sur les annotations et nous avons réalisé des tests préliminaires de cette maquette dans notre site pilote.

Dans cette section, nous indiquons les traitements réalisables avec les annotations. Nous énumérons ensuite les propriétés nécessaires à ces traitements et donnons un premier retour des praticiens.

3.1 Trois types de traitements

3.1.1 Combiner des annotations

Les annotations se combinent pour former des documents de navigation (en lecture seule) et des documents éditables (modifiables).

3.1.1.1 Document de navigation

Un *document de navigation* est un document ajouté aux documents du dossier qui permet d'accéder à ces documents (sommaire, table des matières). Un tel document construit à partir des annotations correspond à une liste de points d'entrée vers des

annotations sélectionnées par l'utilisateur selon certains critères. Depuis les annotations, les lecteurs ont accès aux documents annotés. On leur offre ainsi de nouveaux parcours de lecture. Les points d'entrée correspondent aux titres des annotations pouvant être complétés par des informations comme le nom de l'annotateur ou la date qui résultent des critères de sélection des annotations et qui vont aider le lecteur à choisir les parcours de lecture.

La liste peut être *plate*. Par exemple, un patient souffre d'un problème cardiovasculaire. Un praticien génère un nouveau document lui permettant de visualiser toutes les annotations produites ce jour là et traitant du système cardiovasculaire. Il cherche ainsi à reconstruire en partie l'histoire de cet événement.

La liste peut être *hiérarchisée*. Par exemple, un médecin cherche à se remémorer les échanges de connaissances qu'il a eus avec un confrère. Il visualise la liste des annotations utilisées pour se répondre l'un l'autre. Elles forment des files de discussion pouvant être présentées sous la forme de listes hiérarchisées par le lien « Répondre » (comme pour les forums).

La liste peut être représentée par *un graphe*. Avec les annotations reliant les documents les uns aux autres (et/ou les liens prédefinis entre documents), on génère un graphe⁹. À partir des nœuds, l'utilisateur accède aux documents. Par exemple, un médecin utilise un nouveau protocole. Il recherche tous les documents, dans un ensemble de dossiers, ayant conduit ses confrères à utiliser ce protocole. Ainsi, il peut connaître la manière dont ses collègues l'ont utilisé, dans quelles situations, etc.

3.1.1.2 Document éditable

Pour construire un *nouveau document éditable*, on place les contenus des annotations sélectionnées par l'utilisateur les uns à la suite des autres. L'utilisateur peut ensuite retravailler le document généré, y ajouter des connaissances et le remettre en forme. Par exemple, un médecin rédige le CRH. Il parcourt le dossier, sélectionne et commente des parties avec des annotations destinées au CRH. Le regroupement de ces annotations dans un nouveau document lui donne une base pour rédiger ce CRH.

3.1.2 Filtrer des annotations

Lors de la création d'une annotation, l'annotateur peut spécifier les destinataires en imposant des droits d'accès : lui-même, un groupe de personnes, tous les lecteurs. Lors de la consultation du dossier, il y a un filtrage automatique des annotations, qui permet à un lecteur de ne visualiser que les annotations qu'il a le droit de visualiser.

Par ailleurs, au cours de sa lecture du dossier, le lecteur lui-même peut choisir de ne visualiser que certaines annotations sélectionnées en fonction d'un ou plusieurs critère(s) (le nom de l'annotateur, la date). Par exemple, un médecin veut visualiser les annotations portant uniquement sur le système cardiovasculaire ou bien seulement les annotations qu'il a lui-même rédigées.

3.1.3 Envoyer un message

⁹ Logiciel Nestor <http://ausweb.scu.edu.au/aw99/papers/eklund2/paper.html>

Un utilisateur peut envoyer une annotation à un ou plusieurs destinataire(s). Il y a plusieurs sortes de messages :

- un message peut être produit en rapport avec un élément d'un dossier. Par exemple, un praticien lit une analyse et y décèle une anomalie. Il la commente et décide de l'indiquer à toutes les personnes concernées. Un destinataire recevant ce message doit retrouver le document annoté.
- un message peut être produit en rapport avec un patient (i.e. un dossier). Par exemple, un praticien demande l'avis d'un confrère pour élaborer un diagnostic. Il lui envoie un message en liant l'annotation au dossier du patient. Le destinataire recevant ce message, doit retrouver ce dossier.
- un message peut être produit en réponse d'un autre message. Lorsque le destinataire reçoit ce message, il doit pouvoir retrouver le message précédent, ainsi que la source, si elle existe, à l'origine du premier message.

Un composant logiciel d'annotation complétant une application de gestion des dossiers patient doit pouvoir gérer tous ces messages plus ceux qui n'ont pas de rapport avec un patient particulier.

3.1.4 Impacts des traitements sur le contenu du dossier.

Le composant logiciel d'annotation permet d'agir sur les annotations, en les filtrant et en les envoyant comme des messages. Il peut aussi modifier le contenu du dossier en ajoutant de nouveaux documents. Ces derniers sont issus uniquement de calculs sur les annotations combinés ou non avec l'intervention d'un humain (pour choisir les critères de sélection des annotations).

3.2 Propriétés des annotations

Nous énumérons maintenant les propriétés nécessaires aux traitements précédemment décrits. Une première liste a été élaborée à partir de la liste des traitements et de la liste classique des métadonnées du Dublin core. Cette liste comprend des propriétés liées à l'événement à l'origine de l'annotation, des propriétés liées à l'action d'annotation, des propriétés liées à l'audience de l'annotation et des propriétés liées au contenu sémantique de l'annotation.

Les propriétés liées à l'événement qui est à l'origine de la création d'une annotation (ou, qui, quand, etc.) sont : *l'annotateur*, la *date* (de création ou de modification), le *document*, la *cible*. La cible est utile pour attribuer un identifiant à l'annotation. Cet identifiant résulte d'une interprétation de la cible. Suivant l'objet annoté, l'identifiant correspond à du texte (si la cible est textuelle), au nom de l'image (si la cible est une image), etc. Comme précisé par (Lewkowicz *et al.*, 2005) ces connaissances correspondent à la dimension organisationnelle de l'annotation qui permet de déterminer la place et le rôle de l'annotation dans l'organisation. Ces propriétés sont utilisées pour combiner, filtrer et envoyer des annotations.

Les propriétés liées à l'action d'annotation déterminent sa *catégorie* à savoir si celle-ci est un commentaire, un lien entre deux documents (un document du dossier ou

un document externe au dossier), une annotation créée en vue de rédiger une synthèse, une réponse à une annotation, un message pour un destinataire précis. Ces propriétés sont utilisées pour combiner des annotations (pour créer une liste hiérarchisée -resp. un graphe-, seules les annotations message -resp. lien- sont traitées) et envoyer des annotations (seules les annotations message sont envoyées).

Les propriétés liées à l'audience de l'annotation correspondent aux destinataires de l'annotation. Quelle est la *sphère* de l'annotation : l'annotateur lui-même (sphère privée), tous les lecteurs du document (sphère publique), un (groupe de) destinataire(s) précis (sphère du groupe) (Zacklad *et al.*, 2003) ? Ces propriétés sont utilisées pour filtrer les annotations (pour le filtrage automatique par rapport au droit d'accès fixé par l'annotateur) et envoyer des annotations (pour n'envoyer un message qu'aux destinataires situés dans la sphère).

Les propriétés liées au contenu sémantique de l'annotation permettent de connaître le domaine auquel se rapporte l'annotation. Le domaine correspond à un ensemble de connaissances formant un référentiel commun pour un ensemble de personnes. L'utilisateur peut alors choisir un *thème* relatif à ce domaine pour qualifier l'annotation. (Lewkowicz *et al.*, 2005) parle de dimension spécifique au domaine. Attribuer un thème à une annotation consiste donc à la typer avec des connaissances spécifiques du domaine des utilisateurs (des mots clés pour une spécialité médicale). Ces propriétés sont utilisées pour combiner et filtrer des annotations.

3.3 Premiers retours des médecins et travaux futurs

Notre maquette a été présentée aux praticiens pour des tests d'utilité. Pendant des entretiens individuels, nous leur avons montré comment utiliser les interfaces et nous avons récolté leurs premières remarques. Ils ont particulièrement apprécié la possibilité de créer des synthèses (combinaison d'annotations dans des documents éditables) car cette activité est très importante pour eux. Grâce à ces discussions, nous avons validé les traitements et spécifié les propriétés des annotations.

Bien sûr de nombreuses pistes restent encore à explorer avant d'obtenir une représentation stabilisée. Notre maquette doit être testée avec de vrais dossiers, dans de véritables situations de soins. Pour cela, nous avons besoin d'un outil plus robuste. Malheureusement, le contexte économique a obligé notre partenaire industriel à réorienter son travail sur la conception d'une « visionneuse » de documents médicaux électroniques standardisés et stockés dans un entrepôt de données. Cette visionneuse inclut bien une fonctionnalité d'annotation, mais comme cet outil ne sera pas utilisé pour la production des documents médicaux mais seulement pour leur consultation, nous allons seulement retrouver les annotations produites au cours de la lecture des dossiers. Notre partenaire industriel prévoit de tester cet outil au cours de l'été 2005. De notre côté, nous continuerons à étudier les pratiques d'annotations dans le contexte de production du dossier en faisant évoluer notre première maquette.

La validation finale sera de toute manière liée à la manière dont les praticiens utilisent les annotations et les possibilités offertes par les calculs. Nous nous posons toujours des questions sur cette pratique et nous devons notamment étudier son impact sur le contenu du dossier patient informatisé. Existe-t-il un risque d'appauvrissement du

dossier si les praticiens préfèrent annoter plutôt que d'écrire dans les documents ? Comment les autres acteurs utilisant le dossier (chercheurs, gestionnaires) vont utiliser ces annotations ? Comment motiver les utilisateurs à annoter les documents électroniques aussi intuitivement que sur le papier ? Dans ce contexte, il faudra être attentif aux réactions suscitées et à la façon d'y répondre : des travaux plus anciens (projet Dome Séroussi *et al.* 1996) ont montré qu'il fallait être réactif à des demandes que l'on ne peut anticiper et qu'il faut y répondre par des réflexions et finalement des développements logiciels s'insérant parfaitement dans l'activité de soin.

4 Conclusion

Le dossier patient est le partenaire privilégié de la pratique médicale. Son informatisation a de nombreuses conséquences sur les acteurs médicaux et leur organisation. Ayant constaté que les documents sont le support le plus approprié pour manipuler les connaissances médicales pendant les soins, nous affirmons qu'une approche documentaire est adaptée. Nous proposons l'intégration d'une fonctionnalité documentaire particulière : les annotations. En effet, les outils d'annotation sont désormais courants (la plupart des logiciels de traitement de texte et de collaboration permettent d'annoter). Les praticiens annotent déjà les documents papier. L'originalité de notre travail vient alors de la manière dont nous exploitons ces annotations pour faciliter la manipulation des documents électroniques. Les tests préliminaires réalisés par les praticiens sur l'outil développé par notre industriel, valident notre hypothèse qu'un outil d'annotation est utile pour leur mission de soin. Sur le plan des perspectives théoriques, ces travaux ont mis en avant un besoin, consistant à définir le sens des manipulations réalisées sur les annotations. La définition de cette sémantique hypertextuelle annotationnelle reste à compléter, en généralisant nos définitions à des contextes plus larges que le dossier patient car les annotations sont utilisées dans de nombreux autres domaines comme la génétique, l'architecture, etc. Finalement, le changement du support qui bouleverse le travail des praticiens, influence la manière dont ils lisent et écrivent. Comme le Petit Poucet (Stiegler, 2000), ils peuvent laisser des traces de leurs actions sur les documents pour construire leur propre vision du dossier. Les annotations, en tant qu'objets d'action, rendent le lecteur de plus en plus actif au cours du processus de conception des documents. L'écriture rejoint la lecture sous la forme d'une nouvelle activité l'« écrilecture » (Soubrié, 2001).

5 Remerciements

Nous tenons à remercier le docteur G. Krim, responsable du service Réanimation Pédiatrique et de Néonatalogie du CHU de Amiens pour ses précieuses compétences.

Références

- BACHIMONT B. (2004). Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle. *Habilitation à diriger des recherches*. http://www.utc.fr/~bachimon/Livresettheses_attachments/HabilitationBB.pdf
- BADR Y. & LAFOREST F. & FLORY A. (2003). DRUID: coupling user written documents and databases. *Actes de ICEIS International Conference on Enterprise Information Systems*, 23-26 avril 2003, Angers, 191-6. http://lisi.insa-lyon.fr/~flafores/articles/iceis03_16-01-2003.pdf
- BERG M. (1999). The contextual nature of medical information. *International Journal of Medical Informatics* 56, 51-60.
- BRINGAY S. & BARRY C. & CHARLET J. (2004a). Les documents et les annotations du dossier patient hospitalier. *Numéro spécial de la revue I3 Information, Interaction, Intelligence*, Vol. 4, Num. 1, 191-211.
- BRINGAY S. & BARRY C. & CHARLET J. (2004b). Annotations: A new type of document in the Electronic Health Record. *Actes de DOCAM "Document Academy"*, 22-24 Octobre 2004, San Francisco (USA). <http://thedocumentacademy.hum.uit.no/events/docam/04/program.html>
- CHARLET J. & BACHIMONT B. & BRUNIE V. & EL KASSAR S. & ZWEIGENBAUM P. & BOISVIEUX J.F. (1998). Hospitexte : towards a document-based hypertextual electronic medical record. *JAMIA Journal of American Medical Informatics Association*, 5(suppl), 713-717.
- CHARLET J. (2003). L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. *Habilitation à diriger des recherches*. http://tel.ccsd.cnrs.fr/documents/archives0/00/00/69/20/index_fr.html
- DENOUE L. & CHIU P. & FUSE T. (2003). Shared Freeform Input for Note Tacking across Devices. *Actes de Human Factors in Computing Systems*, 5-10 Avril 2003, Fort Lauderdale (Florida), 794-795.
- HARDSTONE, G. & HARTSWOOD, M. & PROCTER, R. & SLACK, R. & VOSS, A. & REES, G. (2004). Supporting Informativity: Team Working and Integrated Care Records. *Acte de ACM Conference on Computer-Supported Cooperative Work*, 6-10 Novembre 2004, Chicago (USA), 142-151.
- LEWKOWICZ M. & LORTAL G. & TODIRASCU A. & ZACKLAD M. & SRITI& M.F. (2004). A Web-based annotation system for improving cooperation in a care network. *Matera, M., Comai, S., Engineering Advanced Web Applications, Rinton Press 2004*. <http://www.ii.uam.es/~rcarro/AHCW04/Lewkowicz.pdf>
- LOVIS C. & LAMB A. & BAUD R. & RASSINOUX A. & FABRY P. & GEISSBÜHLER A. (2003). Clinical Documents: Attribute-Values Entity Representation, Context, Page Layout and Communication. *Actes de l'AMIA, America Medical Informatics Association*, 8-12 Novembre 2003, Washington (USA), 254-258.
- SEROUSSI B. & BAUD R. & MOENS M. & MIKHEEV A. & SPYNS P. & CEUSTERS W. & ZWEIGENBAUM P. (1996). *Rapport final Dome, Delivrable MLAP-Dome 8, DIAM-SIM/AP-HP*.
- SEROUSSI B. & BOUAUD J. & ANTOINE E.C. (2001). ONCODOC: A successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artificial Intelligence in Medicine*, 22(1), 43–64.
- SOUBRIE T. (2001). Apprendre à lire grâce à l'hypertext. *Mémoire de thèse*. http://marg.lng2.free.fr/documents/the0010_soubrie_t/the0010.pdf
- STIEGLER B. (2000). Annotation, navigation, édition électronique : vers une géographie des connaissances. *Ecarts*, n°2.
- ZACKLAD M. & M. LEWKOWICZ & M. BOJUT & J.F DARSE F. & DETIENNE F. (2003). Forme et gestion des annotations numériques collectives en ingénierie collaborative. *Actes de IC Ingénierie des connaissances*, Laval (France), 207-225.
- ZACKLAD M. (2004). Documents for Action (DofA): infrastructures for Distributed Collective Practices. Actes du workshop “Distributed Collbective Practice: Building new Directions for Infrastructural Studies”, associé à la conférence (CSCW) 2004 Computer-Supported Cooperative Work, 6-10 Novembre 2004, Chicago (USA).

Construction d'une ontologie du droit communautaire

Sylvie Despres (1), Sylvie Szulman (2)

(1) Université Paris 5, CRIP5 Equipe IAA

(2) Université Paris 13, LIPN - UMR 7030

sd@math-info.univ-paris5.fr - ss@lipn.univ-paris13.fr

Résumé : Ce papier décrit une méthode de construction d'une ontologie du droit communautaire. La méthode repose sur la construction de micro-ontologies à partir de directives du droit communautaire en utilisant la méthode TERMINAE augmentée d'un processus d'alignement sur une ontologie générique du droit. L'ontologie est ensuite élaborée à l'aide d'un processus de fusion de micro-ontologies.

Mots-clés : construction d'ontologie à partir de textes, alignement et fusion d'ontologie, droit communautaire.

1 Introduction

Ce papier fait suite à un travail effectué dans le cadre de l'Action Spécifique "Ontologies du droit et langage juridique". Une méthode de construction d'une ontologie du domaine du droit communautaire y est présentée. La finalité de cette ontologie est la création d'un modèle formel utilisable dans des systèmes à base de connaissances sur le droit communautaire. La vérification de la cohérence entre les différents concepts utilisés dans les diverses directives européennes est également visée. Une directive est un texte produit par la Communauté Européenne qui traite d'un point de droit.

Le droit communautaire régit les rapports entre les institutions européennes et les gouvernements des pays membres et définit les procédures décisionnelles. Supérieur au droit national, le droit communautaire a des effets contraignants à l'égard des Etats membres et de leurs ressortissants et apporte une protection juridique unifiée à tous les citoyens européens. L'élaboration du droit communautaire nécessite de travailler sur des principes généraux qui, bien que non écrits, s'imposent lors de la rédaction de tous les textes de droit communautaire : l'Etat de droit, la protection des droits fondamentaux, le non cumul des sanctions.

La transposition des directives communautaires exige également la maîtrise des concepts afin de rester en conformité avec l'esprit et éviter les incohérences entre les textes des états membres et ceux de la communauté européenne. La compréhension des concepts définis dans les textes par l'ensemble des états membres

est par conséquent primordiale et constitue un préalable à cette activité.

La méthode proposée consiste en la construction d'un ensemble de "micro-ontologies" à partir des directives éditées par la Communauté Européenne. La méthode TERMINAE Aussénac-Gilles *et al.* (2000) enrichie par un processus d'alignement avec une ontologie générique du droit CLO Gangemi *et al.* (2003) est utilisée pour concevoir les micro-ontologies. Un processus de fusion de ces micro-ontologies permet d'obtenir l'ontologie du droit communautaire.

Nous présentons les ontologies les plus significatives en droit dans la section 2. La section 3 décrit les techniques de combinaison d'ontologies. Puis nous détaillons la méthode dans la section 4 et le travail effectué dans la section 5.

2 Les ontologies dans le domaine du droit

Dans le domaine du droit, les ontologies constituent un axe de recherche important dans des domaines divers comme les systèmes d'information, les services Web, les agents et l'ingénierie des connaissances. Dans ce paragraphe, nous présentons un état de l'art restreint des ontologies juridiques, en commençant par des ontologies génériques, puis en donnant un aperçu de quelques ontologies spécialisées.

Parmi les ontologies génériques du droit figurent : FOLaw (Functionnal Ontology of Law) Valente (1995) Valente *et al.* (1989) qui a ensuite conduit au développement de LRI-Core Breuker & Winckels (2003) ; Frame-Based Ontology Kralingen (1995) Kralingen *et al.* (1999) ; CLO (Core Legal Ontology) Gangemi *et al.* (2003) ; l'ontologie documentaire du droit français pour la reformulation de requêtes sur le Web Lame (2002).

FOLaw se voulait une ontologie de "haut niveau" dont une des finalités était l'identification des différents types de connaissances utilisés dans le raisonnement juridique et en particulier ceux qui lui sont propres. L'essentiel des connaissances représentées dans FOLaw sont les connaissances normatives, les connaissances du monde, les connaissances de responsabilité, les connaissances réactives, les connaissances créatives et les métaconnaissances sur le droit. FOLaw est maintenant considérée par Breuker & Winckels (2003) comme une structure d'inférences de CommonKADS. Les insuffisances de FOLaw ont conduit Breuker & Winckels (2003) à développer LRI-Core qui est également une ontologie de "haut niveau". LRI-Core est construite sur deux niveaux. Les principales catégories sur lesquelles repose LRI-Core sont les entités primaires du monde physique, les entités physiques, les entités mentales. L'organisation sociale et les communications sont composés de rôles qui sont exécutés par des agents qui sont identifiés comme des processus individuels.

La finalité de Frame-Based Ontology est d'améliorer les techniques de développement des systèmes à base de connaissances en droit et de réduire le problème de l'interaction. A l'origine Van Kralingen et Visser ont élaboré deux ontologies distinctes. Si la première était conceptuelle et la seconde formelle, leurs simili-

tudes ont permis aux auteurs de les traiter comme une unique ontologie. Une des idées qui sous-tendait leur travail était que les ontologies constituaient un moyen de réduire le problème de l'interaction dans le contexte de la spécification des connaissances juridiques. Les concepts de normes, d'actions et de concepts légaux sont utilisés comme point de départ à la construction de la Frame-Based Ontology. Une norme contrôle et restreint des actes en terme de modalités juridiques. Les actes sont des descriptions complexes des actions Stuckenschmidt *et al.* (2001).

CLO est fondée sur l'ontologie de haut-niveau DOLCE +, une extension de DOLCE Gangemi & Mika (2003). Le processus de développement de CLO prend en compte des méthodologies de développement des ontologies "de haut niveau" et des travaux sur les ontologies dans le domaine juridique Gangemi *et al.* (2003). CLO organise les concepts juridiques et les relations sur la base de métapropriétés formelles définies dans DOLCE + Masolo *et al.* (2002). Les types de base des entités du domaine du droit sont supposés être clairement identifiables et expri-mables par un ensemble minimal de propriétés et de relations issues de DOLCE+. Les choix méthodologiques, tout comme l'exploitation des propriétés pertinentes pour le domaine juridique sont fondés sur la théorie du droit et la philosophie de la Loi.

L'ontologie documentaire du droit français pour la reformulation de requêtes sur le Web est destinée à faciliter l'accès au site juridique www.droit.org. Dans ce travail, le terme ontologie est envisagé comme un ensemble de termes et de concepts structurés entre eux par des liens de divers types. Chaque concept peut présenter plusieurs sens thématiques. L'ontologie est construite automatiquement à partir d'un corpus constitué de douze codes du droit français, jugés essentiels (Code civil, Code du travail). L'analyseur syntaxique de corpus syntax Bourigault & Fabre (2000) est utilisé pour extraire de ce corpus une liste de noms et de syntagmes nominaux, structurée par des relations de dépendance syntaxique. Les résultats de l'analyse syntaxique sont ensuite exploités par le module d'analyse distributionnelle upery Bourigault (2002) pour enrichir cette structuration initiale avec des liens de type distributionnel (coordination et co-occurrence statistique). Ce réseau structuré constitue la première version de l'ontologie. Il est intégré dans une interface d'accès aux documents du site droit.org dans lequel il est utilisé comme un index thématique au sein duquel l'utilisateur peut naviguer pour définir ou préciser sa requête, et comme ressource pour un module d'expansion de requêtes Bourigault & Lame (2001).

Le processus de développement des trois premières ontologies génériques décrites s'appuient sur des théories du droit Hohfeld (1996) Kelsen (1991) pour la modélisation des concepts, les relations qu'ils entretiennent et le raisonnement dans le domaine juridique. L'approche de construction adoptée par Guiraud Lame est effectuée automatiquement à partir de textes à l'aide d'outils de TAL. Ces ontologies sont peu formalisées sauf CLO qui est écrite en OWL.

Il existe également de nombreuses ontologies de domaine ou d'application. Le terme "ontologie" est alors à prendre au sens large. Il peut s'agir de bases de données terminologiques contenant des informations terminologiques sur des

éléments de compréhension du domaine de la tva Melz & Valente (2004) ou des ontologies exprimées en rdf comme la propriété intellectuelle (IPR) Sagri *et al.* (2004) ou des ontologies plus formelles comme dans le domaine de la fraude financière Zhao *et al.* (2004) ou OPLK conçue pour aider les juges dans leurs activités Benjamins *et al.* (2003) ou IPROnto dans le domaine des droits à la propriété intellectuelle Delgado *et al.* (2003). L'utilisation d'ontologies au format XML pour comparer et harmoniser des textes législatifs est également étudié dans le projet E-POWER Boer *et al.* (2003).

3 Combinaison d'ontologies

L'existence de nombreuses ontologies créées dans le domaine du droit ou dans d'autres domaines conduit inévitablement à leur réutilisation. Cette réutilisation implique une combinaison d'ontologies réalisée avec des opérations d'alignement ou de fusion. L'alignement consiste à rechercher des relations entre des concepts appartenant à des ontologies différentes. La fusion crée une nouvelle ontologie à partir de deux ou plusieurs ontologies ayant une partie commune Klein (2001). Les disparités entre deux ontologies peuvent apparaître à deux niveaux logique et ontologique :

- au niveau logique, les disparités sont liées au langage utilisé pour décrire chaque ontologie. La syntaxe, la représentation logique, les primitives sémantiques et l'expressivité du langage peuvent être différentes et rendent l'appariement difficile ;
- au niveau ontologique, des différences peuvent apparaître sur le plan de la conceptualisation, l'explication et la terminologie des ontologies.

Actuellement, les travaux réalisés dans les domaines des graphes, des bases de données et de l'apprentissage contribuent largement à la combinaison d'ontologies. Les travaux sur les mesures de similarités sont également exploitables dans ce contexte Thieu *et al.* (2004). Euzenat & Valtchev (2004) ont proposé une mesure pour comparer les entités de deux ontologies écrites en OWL-Lite. La méthode décrite dans le paragraphe suivant exploite certains des éléments décrits dans ces travaux.

4 La méthode

4.1 Architecture de la méthode

La méthode de construction de l'ontologie d'une partie du droit communautaire comporte deux étapes :

- la construction de "micro-ontologies" à partir de directives en utilisant la méthode TERMINAE enrichie par un alignement avec l'ontologie générique CLO ;
- la fusion des "micro-ontologies" obtenues à l'étape précédente.

La figure 1 présente l'architecture de la méthode.

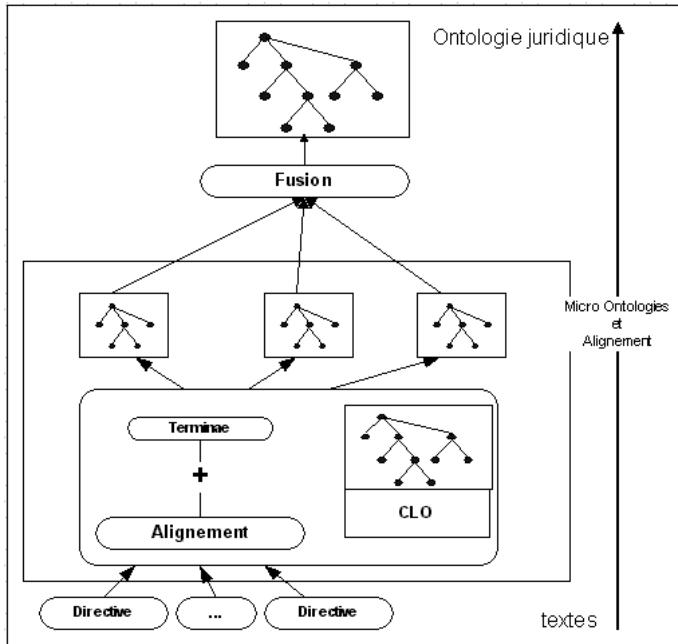


FIG. 1 – Architecture de la construction de la méthode

4.2 Construction des "micro-ontologies"

Cette méthode s'appuie sur la méthode TERMINAE enrichie par un processus d'alignement avec l'ontologie générique CLO. Une première expérience d'alignement avec DOLCE a montré l'intérêt de partir d'une ontologie générique du droit Després & Szulman (2004). Les concepts sont spécifiques au domaine ce qui permet de les réutiliser plus facilement. Dans ce travail, CLO a été choisie pour réaliser un alignement à partir des définitions terminologiques des concepts. La méthode TERMINAE prend en compte des occurrences de termes en corpus pour décrire des concepts dans une ontologie formelle. Les concepts créés à l'aide de ce processus sont dits terminologiques et le lien texte/concept est conservé dans des fiches terminologiques. Les concepts non terminologiques dits de structuration vont être trouvés en partie dans l'ontologie générique CLO.

4.3 Fusion de micro-ontologies

La fusion des micro-ontologies est facilitée car le langage de représentation utilisé est le même pour les différentes micro-ontologies. Elle est réalisée avec l'API d'alignement Euzeunat (2004) sur les ressources construites avec Terminae. Elle permet de retrouver automatiquement les concepts ayant la même

étiquette et en déduire les concepts différents. L'identité des étiquettes ne suffit pas à identifier les concepts. Il faut vérifier s'il y a une correspondance sémantique en s'appuyant sur la définition et les propriétés associées décrites dans les fiches terminologiques. Pour les concepts différents, il faut identifier les concepts proches sémantiquement et comparer leur sens en utilisant les définitions en langage naturel consignées, à défaut les occurrences des termes, dans les fiches terminologiques.

5 Les résultats

Nous présentons dans une première section l'étude linguistique et la construction ontologique sur la première directive que nous nommons "D-travailleur", puis nous donnons le résultat de la deuxième directive nommée "D-citoyen" avant de présenter le résultat de la fusion de ces deux micro-ontologies.

5.1 Construction des micro-ontologies

5.1.1 Le corpus

Nous avons travaillé sur deux directives, la Directive 2001/23/CE du *"Conseil du 12 mars 2001 concernant le rapprochement des législations des États membres relatives au maintien des droits des travailleurs en cas de transfert d'entreprises, d'établissements ou de parties d'entreprises ou d'établissements"* (D-travailleur) et la Directive *"2004/38/CE du Parlement européen et du Conseil du 29 avril 2004 relative au droit des citoyens de l'Union et des membres de leurs familles de circuler et de séjourner librement sur le territoire des États membres"* (D-citoyen).

5.1.2 Etude linguistique

Cette étape de la méthode consiste à sélectionner les termes et les relations lexicales qui doivent être modélisés à partir de la directive. Nous avons utilisé plusieurs outils de TAL : SYNTEX Bourigault & Fabre (2000) a fourni 900 candidat-termes dont les plus utilisés sont *travailleur*, *transfert*, *cédant*, *cessionnaire*. LINGUAE qui est un concordancier inclus dans l'outil TERMINAE Szulman *et al.* (2002) et MFD qui est un module de fouille textuelle Ceausu & Després (2004) nous ont permis d'étudier les relations lexicales. La figure 2 présente un exemple d'étude de la relation entre *cédant* et *cessionnaire* avec LINGUAE.

(lemme : cédant) (*) (lemme :cessionnaire)
cédant notifie au cessionnaire

FIG. 2 – Etude de relation lexicale avec Linguae

5.1.3 Normalisation sémantique

La normalisation sémantique est l'étape qui permet de passer de l'étude lexicale et syntaxique à l'étude sémantique et à la modélisation. La méthode propose de partir des concepts centraux du modèle à réaliser. Ces concepts sont donnés par l'objectif de la micro-ontologie, soient TRAVAILLEUR, TRANSFERT, LICENCEMENT.

A. Amorce de la modélisation

Chaque concept central est décrit par une fiche terminologique qui contient l'ensemble des occurrences du ou des termes (synonymes) correspondants. Une définition en langage naturel est soit trouvée dans le texte, soit établie à partir des occurrences du terme en contexte. A partir des occurrences du terme considéré, un repérage des propriétés qui lui sont associées est effectué. Nous avons défini deux types de propriétés, structurelle et fonctionnelle. Une propriété structurelle indique une relation avec un concept ancêtre. Une propriété fonctionnelle décrit une relation entre concepts qui ne sont pas en relation de subsumption. La syntaxe de ces propriétés est proche du langage de description de l'ontologie. Nous avons trouvé dans le texte la définition du terme *travailleur* : *toute personne qui, dans l'Etat membre concerné, est protégée en tant que travailleur dans le cadre de la législation nationale sur l'emploi*.

Dans cette définition trois idées sont à retenir : (i) la Communauté Européenne est constituée d'Etats Membres qui ont chacun des lois nationales ; (ii) un travailleur est protégé par le droit du travail du pays dont il est citoyen ; (iii) il n'y a pas de liens explicites entre droit communautaire et droit national. Les propriétés du concept TRAVAILLEUR sont établies à partir de ces éléments (voir figure 3).

PROPRIÉTÉS STRUCTURELLES	PROPRIÉTÉS FONCTIONNELLES
est une personne appartient à une entreprise appartient à un Etat membre	possédant des droits et des intérêts ayant des représentants ayant des conditions de travail est l'objet du transfert de son entreprise est protégé par une législation nationale

FIG. 3 – Définition du concept TRAVAILLEUR

B. Consolidation de la modélisation

La création de la micro-ontologie est obtenue en intégrant les définitions des concepts élaborées à l'étape de normalisation enrichie par un processus d'alignement avec l'ontologie CLO. Les propriétés structurelles sont utilisées pour rechercher les concepts ancêtres (axe ascendant). Le travail est ensuite poursuivi sur un axe descendant pour spécialiser ou différencier des concepts Bachimont (2000) déjà définis. Les rôles décrivent les propriétés fonctionnelles. Certains

Nom	Concept domaine	Concept valeur
estRégi	licencement	législationNationalSurEmploi
estRégi	Social-Object	objet juridique
régit	objet juridique	Social-Object
estProtégé	travailleur	législationNationalSurEmploi
appartient	travailleur	états membres

FIG. 4 – Un extrait de l'ensemble des rôles génériques

restreignent des rôles hérités.

* Généralisation du concept TRAVAILLEUR

Le travail sur l'axe ascendant consiste à trouver les concepts pères d'un concept. Les propriétés structurelles sont les expressions linguistiques à partir desquelles la recherche des concepts pères est effectuée.

La propriété structurelle *est une personne* conduit à étudier le terme *personne* dans le domaine juridique. Dans la directive, le terme *personne physique ou morale* est présent et entre dans la définition d'un *cédant*. Avec l'aide de juristes et du texte, nous avons attaché trois sens différents à ce terme. Chaque sens est décrit par un concept. Nous distinguons les concepts PERSONNEPHYSIQUE, PERSONNEMORALE et PERSONNEPHYSIQUEOUMORALE qui sont intégrés dans la micro-ontologie. L'expression linguistique *le travailleur est une personne* est traduite par un lien de subsumption entre le concept TRAVAILLEUR et le concept PERSONNEPHYSIQUE. La propriété structurelle *appartient à* conduit à modéliser les termes *entreprise* et *états membres* et à exprimer la relation par un rôle APPARTIENT.

* Etude des propriétés fonctionnelles

Les propriétés fonctionnelles interviennent dans la définition des rôles. Un modèle des termes *droit, intérêts, représentant, conditions de travail, législation nationale* est établi afin de relier ces concepts au concept TRAVAILLEUR par un rôle exprimant la relation linguistique.

* Définition de rôles

Nous avons créé le concept de structuration ascendante OBJETJURIDIQUE qui va regrouper tous les concepts décrits par un texte juridique et qui est défini par le rôle RÉGIT à valeur dans le concept SOCIAL-OBJECT de CLO. Le rôle inverse ESTRÉGI est explicité. Les relations lexicales comme *couvert par* ou *protégé par* constituent une spécialisation de la relation *est régi*. Il y a donc une hiérarchie entre ces rôles. La figure 4 présente quelques rôles de la micro-ontologie.

* Alignement avec CLO

A chaque concept de structuration ascendante créé comme les concepts PERSONNEPHYSIQUE, OBJETJURIDIQUE un alignement avec CLO est étudié.

CLO contient le concept NATURAL PERSON qui est décrit par *Cognitive objects have a specific dependence on agentive physical objects (e.g. a natural person)*. Le concept PERSONNEPHYSIQUE est identifié au concept NATURAL PERSON.

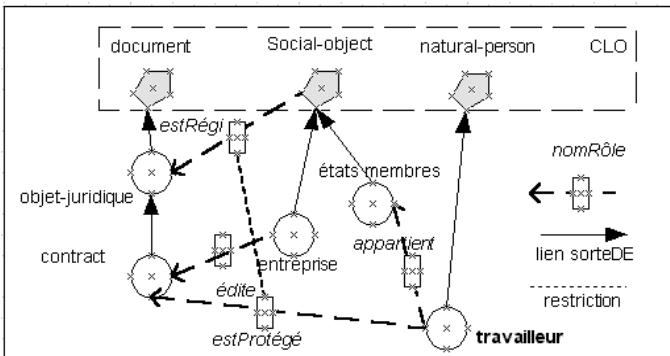


FIG. 5 – Un extrait de la micro-ontologie "D-travailleur"

Le concept CLO DOCUMENT est décrit par *An information realization that realizes (at least) a text*. Le concept de structuration ascendante OBJETJURIDIQUE regroupe tous les concepts décrits par un texte juridique est lié au concept DOCUMENT de CLO par un lien de subsomption.

La figure 5 présente un extrait de la micro-ontologie "D-travailleur". L'extrait focalise sur le concept TRAVAILLEUR.

5.2 Fusion d'ontologies

Un travail équivalent a été effectué pour construire la micro-ontologie "D-citoyen" à partir de directive "D-citoyen". Les concepts centraux sont CITOYEN, ETATS MEMBRES, LIBRE CIRCULATION. Bien qu'il ne soit pas un concept central de cette directive, le terme *travailleur* est présent et est modélisé car il est l'objet de l'étude. La figure 6 donne un extrait de la micro-ontologie "D-Citoyen".

Chaque micro-ontologie représente un modèle d'un point du droit communautaire. Pour obtenir un modèle du travailleur, il est nécessaire de fusionner les micro-ontologies décrites précédemment. Le processus de fusion n'est pas achevé mais un début de solution est présenté ci-après et doit être finalisé. La micro-ontologie "D-citoyen" sert d'ontologie de référence dans ce processus de fusion. Les deux micro-ontologies sont écrites dans le même langage, il n'y a donc pas de disparités à ce niveau. Les disparités se situent au niveau terminologique et ontologique. Le concept TRAVAILLEUR est présent dans les deux micro-ontologies. Le recours aux fiches terminologiques permet de mettre en évidence que TRAVAILLEUR dans la micro-ontologie D-travailleur est identique à TRAVAILLEUR SALARIÉ de la micro-ontologie D-citoyen et qu'il faut fusionner ces deux concepts. Le rôle APPARTIENT attaché à TRAVAILLEUR est supprimé car hérité de CITOYEN. La fiche terminologique associée à TRAVAILLEUR est mise à jour.

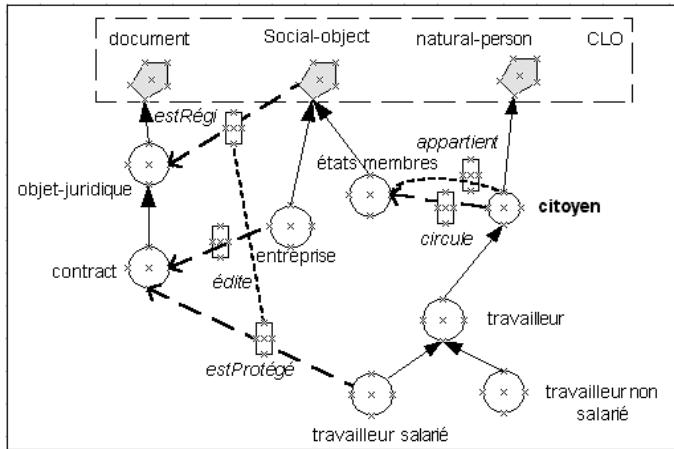


FIG. 6 – Un extrait de la micro-ontologie "D-Citoyen"

6 Conclusion

Ce travail a été initié par des juristes. Ils souhaitaient établir des relations entre la législation européenne et la législation nationale autour du concept de travailleur. Ils ont validé l'ontologie résultat. A ce stade de l'étude, nous proposons d'étudier d'autres directives dans lesquelles le concept de travailleur est présent afin de vérifier l'adéquation à notre modèle en utilisant la même méthode. L'application de la méthode aux mêmes directives écrites en anglais montre des différences terminologiques et mériterait d'être approfondie.

La méthode décrite peut être appliquée à d'autres domaines. Elle contient des processus de combinaisons d'ontologies qui doivent être encore améliorés ; le processus d'alignement avec CLO nécessite une connaissance approfondie de cette ontologie générique mais les concepts d'ancre comme (NATURAL-PERSON et DOCUMENT) facilitent le processus de fusion.

Il reste un travail important à faire dans le processus de fusion mais nous avons montré que le recours au texte est nécessaire pour établir une identité des concepts dans des domaines particuliers comme le droit communautaire.

L'ontologie du droit communautaire obtenue à la fin de ce processus est écrite en OWL.

Remerciements

Cette recherche a été effectuée dans le cadre de l'action spécifique "Ontologie du droit et langages juridiques" (G.Lame) subventionnée par le RTP "Droit et systèmes d'information" (Danièle Bourcier). Nous remercions Nicolas Moisard (IDEA, Centre Juridique de Poitiers) pour avoir formulé les problèmes juridiques et pour son travail de validation.

Références

- AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Revisiting ontology design : a methodology based on corpus analysis. In R. DIENG & O. CORBY, Eds., *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, p. 172–188 : Springer-Verlag.
- BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, p. 305–323, Paris : Eyrolles.
- BENJAMINS V., CONTRERAS J., CASANOVAS P., AYUSO M., BECUE M., LEMUS L. & URIOS C. (2003). Ontologies of professional legal knowledge as the basis for intelligent it support for judges. In *ICAIL Workshop on Legal Ontologies and Web based legal information management*.
- BOER A., VAN ENGERS T. & WINCKELS R. (2003). Using ontologies for comparing and harmonizing legislation. In *ICAIL 2003*, p. 60–69.
- BOURIGAULT D. (2002). Upéry : un outil d'analyse distributionnelle étendu pour la construction d'ontologies à partir de corpus. In *Actes de la 9ième conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, p. 75–84, Nancy, France.
- BOURIGAULT D. & FABRE C. (2000). *Approche linguistique pour l'analyse de corpus*, In *Cahiers de Grammaires*, volume 25, p. 131–151. Université Toulouse Le Mirail.
- BOURIGAULT D. & LAME G. (2001). Analyse distributionnelle et structuration de terminologie. *Revue TAL*, 10.
- BREUKER B. & WINCKELS R. (2003). Use and reuse of legal ontologies in knowledge engineering and information management. In *ICAIL 2003 Workshop on Legal Ontologies & Web based Legal Information Management*.
- CEAUSU V. & DESPRÉS S. (2004). Une approche mixte pour la construction d'une ressource terminologique. In *IC 2004*, p. 211–223.
- DELGADO J., GALLEGOS I., LLORENTE S. & GARCIA R. (2003). Ipronto : An ontology for digital rights management. In *JURIX 2003*.
- DESPRÉS S. & SZULMAN S. (2004). Construction of a legal ontology from a european community legislative text. In T. F. GORDON, Ed., *Jurix 2004*, p. 79–88 : IOS press.
- EUZENAT J. & VALTCHEV P. (2004). Similarity-based ontology alignment in owl-lite. In *ECAI 2004*, Valencia(Espagne).
- EUZEUNAT J. (2004). *An API for ontology alignment*. INRIA Rhône-Aples, <http://co4.inrialpes.fr/align/>.
- GANGEMI A. & MIKA P. (2003). Understanding the semantic webthrough descriptions and situation. In M. R. . AL., Ed., *ODBASE03*, Berlin : Springer Verlag.
- GANGEMI A., PRISCO A., SAGRI M., STEVE G. & TISCORNIA D. (2003). Some ontological tools to support legal regulatory compliance, with a case study. In *Workshop WORM Core : LNCS*, Springer Verlag.

- HOHFELD W. (1996). *Fundamental Legal Conceptions as Applied in Legal Reasoning*. W.W. Cook - Yale University Press 1919.
- KELSEN H. (1991). *General Theory of Norms*. Oxford : Clarendon Press.
- KLEIN M. (2001). Combining and relating ontologies : an analysis of problems solutions. In A. GOMEZ-PEREZ, M. GRUNINGER, H. STUCKENSCHMIDT & M. USCHOLD, Eds., *Workshop on Ontologies and Information Sharing, IJCAI'01*, p. 309–327, Seattle, USA.
- KRALINGEN R. V. (1995). *Frame-based Conceptual Models of Statute Law*. PhD thesis, University of Leiden, The Hague The Netherlands.
- KRALINGEN R. V., VISSER P., BENCH-CAPON & DEN HERICK H. V. (1999). A principled approach to developing legal knowledge systems. *International Journal of Human Computer Studies*, **51**, 1127–1154.
- LAME G. (2002). *Construction d'ontologie à partir de textes. Une ontologie du droit dédiée à la recherche d'information sur le Web*. PhD thesis, Thèse d'université. Ecole des Mines de Paris.
- MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. & SCHNEIDER L. (2002). *The Wonder Web Library of Foundational Ontologies*. Rapport interne, Laboratory for Applied Ontology, <http://wonder-web.semanticweb.org>.
- MELZ E. & VALENTE A. (2004). Modeling the tax code. In S. LNCS, Ed., *WORM 2004*, Larnaca(Cyprus).
- SAGRI M., TISCORNIA D. & GANGEMI A. (2004). An ontology-based model for representing "bundle-of-rights". In S. LNCS, Ed., *WORM 2004*, Larnaca(Cyprus).
- STUCKENSCHMIDT H., STUBKJÆR E. & SCHLIEDER C. (2001). Modeling land transactions : legal ontologies in context. In *Second international Workshop on Legal Ontologies*.
- SZULMAN S., BIÉBOW B. & AUSSENAC-GILLES N. (2002). Structuration de terminologies à l'aide d'outils de tal avec TERMINAE. In A. NAZARENKO & T. HAMON, Eds., *Traitement automatique des langues. Structuration de terminologie*, volume 43, p. 103–128 : Hermès.
- THIEU M., STEICHEN O., ZAPLETAL E., JAULENT M. & BOZEC C. L. (2004). Mesures de similarité pour l'aide au consensus en anatomie pathologique. In *IC 2004*, p. 225–236.
- VALENTE A. (1995). *Legal knowledge engineering : A modelling approach*. Amsterdam, The Netherlands : IOS Press.
- VALENTE A., BREUKER J. & BROUWER P. (1989). Legal modelling and automate reasoning on-line. *International Journal of Human Computer Studies*, **51**, 1079–1126.
- ZHAO G., KINGSTON J., KERREMANS K., COPPENS F., VERLINDEN R., TEMMERMAN R. & MEERSMAN R. (2004). Towards an ontology of forensics covering financial securities fraud. In S. LNCS, Ed., *WORM 2004 : The second International Workshop on Regulatory Ontologies*, Larnaca(Cyprus).

Trois méthodes d'analyse pour conceptualiser le contenu de différentes sections des monographies des médicaments

Catherine Duclos, Jérôme Nobécourt et Alain Venot

Laboratoire d'Informatique Médicale et de BIOinformatique (Lim&Bio), Université Paris 13
catherine.duclos@vc.aphp.fr, {j.nobecourt, avenot}@smbh.univ-paris13.fr

Résumé : A partir de l'expérience issue de travaux de modélisation conceptuelle des connaissances contenues dans trois sections différentes des monographies des médicaments (indication, pharmacodynamie, pharmacocinétique), une analyse des méthodes de modélisation est proposée. Les différentes méthodes (pattern matching, modélisation ascendante et approche mixte) et les modalités de leur choix sont analysées en mettant en lumière des différences de nature entre les textes et l'existence de connaissances sur le domaine. Ceci nous conduit à proposer plusieurs indicateurs descriptifs de la nature du texte qui nous semblent susceptibles d'aider au choix d'une des trois méthodes proposées. Nous proposons aussi plusieurs méthodologies d'évaluation des modèles obtenus, elles aussi étant liées aux caractéristiques des textes initiaux.

Mots-clés : Médicament, Résumé des Caractéristiques Produit, TAL, Modélisation, Evaluation

1 Introduction

Pour tout médicament ayant une autorisation de mise sur le marché, un résumé des caractéristiques produit (RCP)¹ est élaboré. Ce document textuel, validé par l'autorité de régulation des produits de santé, décrit, en utilisant un vocabulaire contrôlé, les propriétés pharmaceutiques, cliniques, pharmacologiques et administratives du médicament. Le RCP est le document de référence pour le médecin qui cherche l'information légale sur le médicament.

Les RCP constituent ainsi le support essentiel pour la construction de bases de connaissances électroniques sur le médicament.

Outre un accès au RCP en simple consultation dans des sites Web ou des CDRom, les différents éditeurs de bases de connaissances sur le médicament², proposent des fonctionnalités de sécurisation de la prescription comme la détection des interactions médicamenteuses ou celle des contre indications. Ceci est possible grâce à un effort de structuration de certains éléments d'information du RCP.

¹ Le RCP est aussi appelé monographie du médicament

² Vidal® <http://www.vidalpro.net/>, Banque Claude Bernard® <http://www.resip.fr/>

Pour les éditeurs de bases de connaissances sur le médicament, le choix des éléments à structurer est guidé par l'utilisation qui sera faite de l'information et par la facilité du travail de structuration. Par exemple, pour développer la fonctionnalité de détection des contre-indications qui implique une interopérabilité entre le dossier patient et la base de connaissances sur le médicament, les éditeurs français se sont limités à transcoder la pathologie décrite dans la contre indication à l'aide de la Classification Internationale des Maladies 10^{ème} édition, alors que l'information y est beaucoup plus complexe comme l'ont montré (Liu *et al.*, 1998).

Cette sélection des éléments d'information du RCP à structurer conduit à une perte d'information ou à des inexactitudes. Par exemple, le Cibadrex® comprimé (*hydrochlorothiazide 10mg/bénazépril 12,5 mg*) a pour indication « traitement de l'hypertension artérielle, en cas d'échec d'une monothérapie par un inhibiteur de conversion » ; le fait de restreindre la structuration de cette indication à la seule pathologie « hypertension artérielle » dénature le sens de l'indication, entraîne une perte d'information et peut conduire à un mésusage de ce médicament.

Il existe peu de travaux qui, à partir d'une analyse des textes des monographies des médicaments, ont abouti à une modélisation des propriétés de ceux ci .

(Liu *et al.*, 1998) ont modélisé les contre indications de 150 médicaments en réalisant une analyse sémantique manuelle de leur RCP. Si leur modèle n'a pas été évalué, il a permis de décrire avec précision les contre indications et de montrer que la plupart des terminologies médicales susceptibles d'être utilisées pour décrire le contexte clinique des contre indications étaient insuffisantes pour exprimer correctement cette information.

Dans le projet Drug ontology³, la globalité des propriétés des médicaments ont été décrites afin d'obtenir une base de connaissances ayant une dénotation formelle basée sur le Common Reference Model d'OpenGalen⁴ . Pour construire cette ontologie, les textes du British National Formulary⁵ ont été étudiés manuellement, et les concepts utiles pour certaines fonctionnalités d'aide à la décision ont été retenus. Si certains concepts sont précisément décrits (comme le concept de forme pharmaceutique (Cimino *et al*, 1999)), d'autres propriétés comme la pharmacologie sont analysées très superficiellement (identification d'un unique concept « pharmacological feature »)(Solomon *et al*, 1999)).

Il existe actuellement environ 5000 spécialités pharmaceutiques, l'ensemble de leurs monographies représentent un corpus de texte d'environ neuf millions de mots. Les expériences d'analyse manuelle des textes des monographies des médicaments montrent des limites liées au volume du corpus à analyser :

- Pour décrire avec précision les concepts contenus dans une section de la monographie, (Liu *et al.*, 1998) se sont restreints à un corpus de texte « humainement » traitable. Pour cela, ils ont sélectionné les RCP à analyser selon des critères objectifs, mais en aucun cas ils n'ont fait mention d'une

³ DRUG ONTOLOGY. (2005). <http://www.cs.man.ac.uk/mig/projects/old/drugontology/>.

⁴ OPEN GALEN (2005) <http://opengalen.org/>.

⁵ British National Formulary <http://www.bnf.org/bnf/>

représentativité du corpus échantillon traité par rapport au corpus global. L'universalité des concepts identifiés n'a donc pas été démontrée.

- En traitant la globalité du corpus, (Solomon *et al.*, 1999) ont perdu en précision dans la description des concepts. Ainsi le concept de pharmacocinétique se limite à un « processus pharmacocinétique », alors que la lecture du texte semble véhiculer de plus riches informations qui permettraient de répondre, par exemple, aux questions suivantes : « quels antibiotiques diffusent dans l'os ? », « pour quels médicaments antiasthmatiques y a-t-il des données chez l'enfant ? ». L'exhaustivité des concepts identifiés dans cette expérience n'a donc pas été démontrée. Cet *a priori* sur le choix des concepts conduit de nouveau à une limitation des fonctionnalités pouvant être développées à partir de la structuration qui sera faite de la connaissance.

Il serait ainsi important de pouvoir disposer d'un modèle conceptuel susceptible de représenter précisément l'ensemble des propriétés des médicaments et qui aurait ainsi un caractère universel. Il faut, à la fois, pouvoir analyser la totalité des 5000 textes de RCP, trouver une méthodologie permettant un accès facilité aux éléments d'informations contenus dans ces textes et élaborer une méthode de validation des modèles conceptuels générés.

La communauté Ingénierie des Connaissances à partir de Textes (ICT) s'intéresse aux problèmes de traitement automatique du langage et de modélisation des connaissances textuelles. A partir des nombreux travaux réalisés, il est possible de proposer 3 points de vue méthodologiques.

- Les documents ont une structuration interne, ou des formats de reconnaissance (pattern) existent pour ces textes : il est possible, dans ce cas d'utiliser des outils de fouille de texte et d'utiliser la structuration du document. On se contente ici de retrouver le contexte d'une expression (groupe de mots) dans le texte (Alphonse *et al.*, 2004) (Georg *et al.*, 2004).
- Les documents ont été pré-traités et peuvent être passés à des outils de Traitement Automatique du Langage (TAL) pour extraire des candidats termes et des relations. Le modèle est constitué petit à petit via des regroupements conceptuels. On utilise ici des méthodes ascendantes basées sur la sémantique différentielle (Le Moigno *et al.*, 2002).
- Des ressources existent pour le domaine (bases de connaissances sur le domaine, UMLS, GALEN (Trombert-Paviot *et al.*, 2000)...). Des textes ont été analysés et des outils de TAL sont disponibles pour une analyse plus poussée. On peut alors adopter des méthodes permettant de raffiner successivement le modèle en réutilisant les outils de TAL et en spécifiant/généralisant les concepts/propriétés (Ceausu & Després, 2004).

Quel que soit le point de vue adopté, la méthode d'analyse repose sur des outils de TAL (étiqueteur, extracteur, analyseur ...) (Zweigenbaum, 1998) (Nazarenko, 2004) et sur des méthodes et/ou outils d'ingénierie des connaissances (utilisation de ressources textuelles, bases de connaissances, ontologie, création de modèles, validation, représentation formelle, opérationnalisation) (Charlet, 2002) (Biébow, 2004). Pour utiliser tous ces outils, il faut, néanmoins avoir la/les connaissance(s) d'un expert.

L'expertise fait partie intégrante de la méthode. Dans le domaine du médicament, des experts sont disponibles. Ils sont représentés par les producteurs (pharmacologue, toxicologues, cliniciens,...) et les utilisateurs (pharmacien, médecins) de l'information.

En fonction de la nature du texte à traiter, il n'existe cependant pas de recommandations pour le choix d'une méthode particulière.

L'objectif de notre travail est de présenter différentes approches méthodologiques basées sur le traitement automatique du langage pour modéliser efficacement la totalité de l'information contenue dans les RCP des médicaments. En prenant comme illustration trois sections du RCP : la pharmacodynamie, les indications et la pharmacocinétique, nous montrons que chacune des méthodes connues en ingénierie des connaissances est applicable aux textes des RCP et que le choix de la méthode doit être guidé à la fois par la nature du texte et par la connaissance du domaine. Nous montrons également que pour évaluer la validité des modèles générés, il faut à la fois tenir compte de la nature du texte mais aussi de la complexité du modèle.

Dans la suite de cet article nous présenterons successivement les domaines de connaissance et les textes étudiés et les choix faits en terme de méthodologie de conceptualisation et d'évaluation.

2 Les domaines de connaissance

Les domaines relatifs à chaque section du RCP sont plus ou moins précisément définis : il peut exister une définition du domaine de connaissance spécifique de l'information textuelle contenue dans une section du RCP (cas de la pharmacodynamie des antibiotiques), une définition à l'échelle du thème de la section (cas de la pharmacocinétique), voire l'absence de définition précise du domaine (cas de l'indication qui recouvre le domaine de la médecine).

2.1 Connaissances spécifiques au texte d'une section du RCP

La rédaction du RCP suit obligatoirement la recommandation III/9163/89 qui définit les différentes parties du document et les informations attendues. Pour certaines classes pharmaco-thérapeutiques comme les benzodiazépines anxiolytiques et hypnotiques ou les antibactériens, il existe des instructions spécifiques de rédaction de certains points particuliers du RCP. Le but de ces recommandations est de présenter l'information qui nécessite une attention particulière selon un format commun. La section pharmacodynamie des antibiotiques est une section qui décrit l'action et l'activité des antibiotiques sur les bactéries. Elle présente la caractéristique d'être rédigée selon une de ces recommandations⁶. Le guide n°3BC5a donne une définition précise des différents chapitres que la section doit contenir et la façon dont l'information doit être présentée.

⁶ European Commission, Pharmaceuticals <http://pharmacos.eudra.org/F2/eudralex/Vol-3/home.htm>

2.2 Connaissances spécifiques au thème d'une section du RCP

La pharmacocinétique décrit le devenir du médicament dans l'organisme. Elle correspond à une discipline scientifique à part entière (Wagner, 1993) et il existe des conceptualisations du domaine (modèle compartimental de la pharmacocinétique, modèle basé sur la physiologie). Les principaux concepts qui doivent être retrouvés dans la section pharmacocinétique du RCP sont définis dans la recommandation III/9163/89. Il n'existe pas, par contre, de description précise de la façon dont la connaissance doit être exprimée et présentée dans le texte du RCP.

2.3 Connaissances non spécifiques

La section « indication » est particulièrement importante car elle définit pour quelles pathologies le médicament peut être prescrit. Il n'existe pas de description précise de ce qui est attendu dans cette section du RCP. Son domaine couvre à la fois la pathologie et la stratégie thérapeutique.

3 Les textes issus des RCP

Les textes trouvés dans les différentes sections du RCP présentent des niveaux de langage différents : soit proches de l'énumération car il existe une terminologie contrôlée (cas des textes de pharmacodynamie) ou des expressions nominales ayant valeur de primitives conceptuelles (cas des textes d'indication), soit utilisant la langue générale car les expressions nominales isolées n'ont de sens que dans un contexte exprimé par une phrase ou un paragraphe (cas des textes de pharmacocinétique).

3.1 Natures des textes issus de trois sections du RCP

Cents trois textes distincts de pharmacodynamie d'antibiotiques ont été identifiés dans la base des monographies de médicaments du Vidal®. Un exemple d'une section de pharmacodynamie d'un antibiotique est présentée en figure 1. L'ensemble des 103 textes constitue un corpus de 29789 mots.

SPECTRE D'ACTIVITE ANTIBACTERIENNE
Les concentrations critiques séparent les souches sensibles des souches de sensibilité intermédiaire et ces dernières, des résistantes : S < 4 mg/l et R > 16 mg/l
CMI pneumocoque : S < 0,5 mg/l et R > 2 mg/l (à titre provisoire)...
ESPÈCES SENSIBLES
Aérobies à Gram positif : <i>Corynebacterium diphtheriae</i> , ..., <i>Streptococcus pneumoniae</i> (15-35 %),
Aérobies à Gram négatif : <i>Escherichia coli</i> (10 -30 %), ...
Anaérobies : <i>Actinomyces</i> , <i>Bacteroides</i> , <i>Clostridium</i> , <i>Eubacterium</i> , ...
Autres : <i>Bartonella</i> , <i>Borrelia</i> , <i>Leptospira</i> , <i>Treponema</i>
ESPÈCES MODEREMENT SENSIBLES (in vitro de sensibilité intermédiaire)
Aérobies à Gram positif : <i>Enterococcus faecium</i> (40-80%)
ESPÈCES RESISTANTES
Aérobies à Gram positif : <i>Staphylococcus Méti-R</i>

Fig. 1 - Extrait de la section pharmacodynamie de l'Augmentin® 500mg/62,5mg comprimé

Dans la base des monographies de médicament du Vidal®, 3046 textes distincts d'indication ont été identifiés (corpus de 143078 mots). Un exemple d'une section des indications est présentée en figure 2.

De même, 1935 textes distincts de pharmacocinétique ont été identifiés dans la base des monographies de médicament du Vidal® (corpus de 318532 mots). Un exemple d'une section de pharmacocinétique est présentée en figure 3.

Elles sont limitées aux infections dues aux germes définis comme sensibles :

- chez l'adulte et l'enfant :
 - en traitement initial des :
 - pneumopathies aiguës
 - surinfections de bronchites aiguës et exacerbation de bronchites chroniques
 - infections ORL (otite, sinusite, angine) et stomatologiques
 - maladie de Lyme : traitement de la phase primaire (érythème chronique migrant) et de la phase primo-secondaire (érythème chronique migrant associé à des signes généraux : asthénies, céphalées, fièvre, arthralgies...)
 - en traitement de relais de la voie injectable des endocardites, septicémies

Fig. 2 - Extrait de la section indication du Clamoxyl® 500mg gélule

Absorption : Administré par voie orale, l'acébutolol est rapidement et presque complètement résorbé ; toutefois, l'effet de premier passage hépatique est important et la biodisponibilité est de 40% ; .. ; Métabolisme : La majorité de l'acébutolol est transformée au niveau hépatique en un dérivé N-acétylé, le diacétolol, qui est un métabolite actif ; le pic de concentration plasmatique de ce métabolite est atteint au bout de 4 heures environ, et les concentrations plasmatiques de diacétolol représentent le double de celles de l'acébutolol.
Distribution : Liaison aux protéines plasmatiques : la liaison aux protéines est faible : 9 à 11 % pour l'acébutolol, 6 à 9 % pour le diacétolol.
Elimination : L'acébutolol et le diacétolol circulants sont excrétés en majorité par le rein.
Insuffisance rénale : L'élimination urinaire est diminuée et les demi-vies de l'acébutolol, et plus encore du diacétolol, augmentent. ..

Fig. 3 - Extrait de la section pharmacocinétique du Sectral® 200mg comprimé

3.2 Spécificité des différents textes

En utilisant les outils de statistiques disponibles dans Cordial⁷, il nous a été possible de décrire ces corpus de texte en terme de nombre de phrases, nombre de phrases verbales, nombre de formes, nombre d'occurrences, pourcentage de mots inconnus. Des traitements supplémentaires sur le corpus étiqueté ont permis de faire des statistiques à l'échelle du texte. Il a ainsi été possible de décrire la nature des textes d'indication, de pharmacodynamie et de pharmacocinétique. Les principaux résultats sont présentés dans le tableau 1.

Les textes des indications sont plus courts que les textes de pharmacodynamie ou de pharmacocinétique (moins de phrases, moins de mots). Les textes des indications et de pharmacodynamie sont plus proches de l'énumération que les textes de pharmacocinétique (part faible des phrases verbales). Le vocabulaire utilisé dans la pharmacocinétique est assez contrôlé puisqu'il y a une utilisation très fréquente d'un nombre limité de mots. Les textes de pharmacodynamie ont une part importante de

⁷ Logiciel Cordial® <http://www.synapse-fr.com/>

vocabulaire inconnu (nom des bactéries écrit en latin). Le contenu des textes des indications est assez variable (peu de mots se retrouvent dans beaucoup de textes) alors qu'il existe une certaine régularité pour les textes de pharmacocinétique et de pharmacodynamie.

Table 1. Indicateurs statistiques calculés sur les 3 corpus de textes des indications, de la pharmacodynamie et de la pharmacocinétique

	Pharmacodynamie	Indication	Pharmacocinétique
Nombre de textes	103	3 046	1 935
Nombre de phrases	4 818	12 177	20 747
Nombre de phrases par texte	47 (19-456)	4 (2-68)	11 (2-108)
Part des phrases verbales	20%	24%	78%
Nombre de mots	29 789	143 078	318 532
Nombre moyen de mots par textes (minimum-maximum)	289 (185-2766)	47 (2-578)	194 (4 -1383)
Nombre de formes (noms,verbes,adverbes,adjectifs) (Part dans le corpus de leur total d'occurrences)	1135 (61%)	6215 (61%)	5838(54%)
Nombre de formes apparaissant plus de 100 fois (Part dans le corpus de leur total d'occurrence)	40 (27%)	142 (26%)	269 (38%)
Pourcentage de mots inconnus	16%	3%	5%
Nombre de formes apparaissant dans 25% des textes	69	2	32

4 Les méthodes de conceptualisation

Les méthodes utilisées sont d'une part fondées sur l'existence d'un modèle conceptuel initial complet ou partiel. Ensuite elles reposent sur l'analyse du format de présentation des données. Si le document est structuré, le pattern matching est applicable. Si le document n'est pas structuré, il faut pouvoir accéder aux éléments d'information du texte et les outils de TAL peuvent être utiles. L'étude des candidats termes produits par de tels outils peut être quasi exhaustive (cas des indications) ou être filtrée par la connaissance du domaine (cas de la pharmacocinétique). L'étude sémantique seule de ces candidats termes est efficace lorsque le texte n'est pas complexe (proche de l'énumération comme dans les indications), mais lorsque la part des phrases verbales devient importante, le sens porte sur les unités textuelles ; l'environnement contextuel des candidats termes doit alors être étudié.

4.1 Le pattern matching appliqué à la modélisation de la pharmacodynamie

Le guide n°3BC5a peut être vu comme un « gold standard » : les informations contenues dans ce standard sont toujours présentes dans le texte du RCP. Chaque chapitre, sous chapitre, définition contenue dans ce gold standard permet de proposer un modèle conceptuel de la connaissance supposée être trouvée dans la section pharmacodynamie des antibiotiques. Ce dernier est décrit dans (Duclos *et al*, 2004).

4.2 L'approche ascendante appliquée à la modélisation de l'indication

La découverte des concepts contenus dans le textes des indications s'est appuyée sur l'analyse des entités lexicales (noms, adjectifs, verbes, adverbes et unités complexes nominales (UCN) produites par Nomino⁸. Les 10543 entités lexicales ont été étudiées par un expert pharmacien et rassemblées dans des groupes sémantiques distincts. L'analyse particulière des UCN par l'expert a permis de définir de nouveaux groupes sémantiques mais aussi les relations entre ces groupes (par exemple l'UCN « folliculite à *Trichophyton rubrum* » permet de définir la relation « étiologie » entre la pathologie « folliculite » et le champignon « *trichophyton rubrum* »).

Une fois découvert l'ensemble des groupes sémantiques et leurs relations, un travail de rapprochement des groupes, de déduction de concepts et d'abstraction a conduit à la production d'un modèle conceptuel (par exemple le concept de « puissance d'action du traitement » qui définit si le médicament est suffisant pour traiter seul la pathologie visée par l'indication ou s'il doit être associé à une autre thérapeutique, regroupe des groupes sémantiques proches comme « traitement d'appoint », « utilisé en association », « traitement complémentaire » ; le concept « puissance d'action du traitement » est une propriété du concept plus abstrait « degré d'efficacité du médicament »). Le modèle développé est décrit dans (Duclos et Venot, 2000).

4.3 L'approche mixte appliquée à la pharmacocinétique

Un noyau de concepts a été initialement établi. Ce noyau contient les concepts fondamentaux de la pharmacocinétique que l'on retrouve dans la recommandation III/9163/89 et dans des supports pédagogiques de pharmacocinétique (par exemple les concepts d'absorption, de distribution, de métabolisme et d'élimination).

La recherche de nouveaux concepts contenus dans le corpus de texte a servi à enrichir le modèle initial. La découverte de ces concepts s'est appuyée sur une analyse sélective portant sur 1127 des candidats termes (CT) parmi les 17520 produits par Lester (Bourrigault, 1995). Ces CT ont été sélectionnés manuellement par un expert pharmacien car ils décrivaient spécifiquement un concept de pharmacocinétique (par exemple « métabolisme ») et ou semblaient importants (par exemple « insuffisance rénale »). Ces CT, étudiés hors contexte, ne permettaient pas de décrire correctement le domaine, par exemple les phrases « La [liaison aux protéines plasmatiques du principe actif Y] est de [10%]. », et « L'[administration du principe actif X] diminue la [liaison aux protéines plasmatiques du principe actif Y]. » contiennent des CT identiques pourtant elles n'ont pas le même sens, la première quantifie la liaison plasmatique, la deuxième explique un mécanisme d'interaction entre 2 principes actifs.

Pour découvrir les relations entre les CT, l'environnement lexical des CT ayant un sens similaire (par exemple « fixation », « liaison ») a été exploré. Le réseau terminologique de ces CT et les unités textuelles dans lesquelles ils apparaissaient a permis de découvrir les CT co-occurents. Ces CT co-occurents ont été listés selon

⁸ Logiciel Nomino® <http://www/ling.uqam.ca/nomino>

leur fréquence de co-occurrence. Pour des CT de sens conceptuellement voisin, les listes des CT co-occurrents ont été comparées et des concepts sous-jacents ont été déduits (par exemple les CT « absorption » et « métabolisme » qui représentent tous les deux un processus peuvent être comparés. « Absorption » apparaît avec « estomac », métabolisme apparaît avec « hépatique », et on peut déduire que le concept « processus » a une propriété qui est « localisation »). Les concepts et propriétés ainsi découverts ont été organisés pour enrichir le modèle initial par spécialisation ou par généralisation. Le modèle développé est décrit dans (Duclos-Cartolano et Venot, 2003).

5 Les méthodes d'évaluation des modèles conceptuels

L'évaluation tend à qualifier la validité du modèle conceptuel (totalité de la connaissance représentée, non déformation de la connaissance, précision des concepts, utilité des concepts, non redondance des concepts). Elle a pour principe commun de présenter à un ou plusieurs experts la connaissance structurée en utilisant les concepts du modèle et le texte initial pour qu'il(s) effectue(nt) une comparaison entre les deux représentations.

L'étape de production des données est plus ou moins longue, ceci étant conditionné à la possibilité de réaliser une extraction automatique, et si ce n'est pas le cas, à la longueur et à la complexité des textes à étudier.

L'évaluation nécessite de recourir à un nombre plus ou moins grand d'experts. Le nombre d'experts sollicités dépend de la lourdeur de la tâche d'évaluation. Plus le modèle est complexe et plus les textes sont longs, plus le travail de l'expert est fastidieux.

5.1 Cas du modèle de la pharmacodynamie

La méthode d'évaluation du modèle a tiré profit de la présence de mots, de phrases ou d'éléments de ponctuation caractéristiques délimitant des sections d'information. Un algorithme de pattern matching fondé sur le repérage de 40 structures caractéristiques a permis d'extraire les éléments d'information décrits dans le modèle conceptuel et de générer automatiquement une base de connaissances structurée de la pharmacodynamie des antibiotiques (la production des données a durée moins d'une heure). Les informations automatiquement extraites ont été confrontées aux informations manuellement saisies par un expert du domaine lisant les textes des monographies. Les taux de rappel et de précision ont été respectivement de 97,9% et 96,2%. L'ensemble de ce processus d'évaluation a pris moins d'une journée.

5.2 Cas du modèle des indications

Pour réaliser cette évaluation, un échantillon de 100 textes d'indication a été tiré aléatoirement parmi les 3046 disponibles. Deux experts travaillant indépendamment ont typé les informations contenues dans ces textes en utilisant les concepts contenus dans le modèle des indications. Ce travail a duré 2 jours. L'évaluation de la variabilité

entre ces experts dans l'utilisation du modèle pour transcrire l'indication a permis de qualifier la précision des définitions de concepts, l'intervalle de confiance (IC) à 95% du pourcentage de concepts précisément défini est de 89,3% à 94,7%. Ce travail de production des données a aussi permis d'identifier les concepts inutiles (non utilisés) et les concepts redondants (répétition d'éléments structurés dans des concepts différents). Un troisième expert a été sollicité pour définir si le modèle était capable de couvrir l'information dans l'indication. Pour cela il a donné, à chaque texte d'indication, un score compris entre 0 et 2 (0= pas de correspondance raisonnable, 2= correspondance totale). Ce critère d'évaluation développé par Chute (*Chute et al.*, 1996) a qualifié à la fois la complétude du modèle et la déformation du sens de l'information que le modèle est capable de produire. Le score final obtenu a été de 1,95. Il a fallu moins d'une journée pour réaliser cette évaluation

5.2.1 La pharmacocinétique

Pour réaliser cette évaluation, un échantillon de 100 textes de pharmacocinétique a été tiré aléatoirement parmi les 1935 disponibles. Un expert a typé les informations contenues dans ces textes en utilisant les concepts contenus dans le modèle de la pharmacocinétique. Ce travail a duré 3 mois. Devant la lourdeur de la production des données, il n'était pas envisageable de solliciter un 2^{ème} expert pour renouveler cette tâche comme dans le cas de l'évaluation du modèle des indications.

Pour pallier à ce défaut de double production des données, il a été décidé de réaliser une double évaluation en aveugle de chacun des textes. Le volume du corpus à évaluer étant trop important, il a été décidé qu'un expert n'évaluerait que 25 textes attribués aléatoirement. Au total 8 experts ont été sollicités pour cette évaluation.

L'utilisation du modèle pour décrire entièrement le texte produisant en moyenne plus de 100 informations à enregistrer par un expert, il a été décidé de faire l'évaluation au niveau de la phrase. Ainsi, pour conduire cette évaluation, chaque texte a été découpé en phrases qui ont été évaluées selon 2 critères : un critère de complétude (complètement, presque complètement, peu, pas structuré) et un critère de distorsion (entièrement déformé, déformé de façon importante, peu, pas déformé). A l'issue de cette analyse par phrase, l'expert était alors capable d'affecter pour un texte, une valeur globale pour chacun de ces critères. Comme chaque texte était évalué 2 fois, si l'appréciation des experts divergeait, une recherche de consensus était tentée en utilisant la méthode Delphi (elle consiste à refaire une évaluation à la lumière de l'ensemble des résultats obtenus). Sur les 100 textes, 93 évaluations ont été consensuelles dès le 1^{er} tour d'évaluation, et les 7 autres dès le 2^{ème}. Pour 95 à 100% des textes (IC à 95%) l'information n'était pas déformée et pour 83 à 95% des textes (IC à 95%), l'ensemble des concepts était représenté par le modèle. Ce travail d'évaluation a pris une semaine.

6 Discussion et conclusion

L'analyse comparative des méthodes utilisées pour modéliser l'information contenue dans différentes sections du RCP laisse entrevoir un possible scénario

méthodologique pour choisir un mode de traitement des textes en fonction de leur nature et de la connaissances du domaine. Quand le domaine de connaissance est défini à l'échelle du texte du RCP, la terminologie est contrôlée ainsi que l'organisation du texte. L'application du pattern matching permet la production automatique d'une base de connaissance mais est sensible à la qualité du texte. Toute faute d'orthographe, abréviation non prévue, variance d'expression peut altérer les performances de cette méthode. Lorsque le domaine est défini à l'échelle du thème d'une section du RCP, il est possible d'utiliser efficacement les CT issus des techniques de TAL en s'affranchissant du bruit lié à l'extraction terminologique par un filtrage des CT spécifiques du domaine. Ces CT sont le point d'ancrage de l'analyse des autres CT qui lui sont rattachés dans des contextes d'occurrence. Ce sont ici les relations entre CT qui sont importantes pour la construction du modèle. Enfin, lorsque le domaine n'est pas précisément défini, les techniques de TAL peuvent encore être utilisées efficacement si le texte est proche de l'énumération car les CT suffisent pour la construction du modèle ; par contre il n'y a pas moyen de s'affranchir du bruit associé à l'extraction terminologique.

Il existe une difficulté à trouver une méthode parfaitement adaptée à l'évaluation de modèles conceptuels : les critères développés par (Guarino & Christopher 2004, Gomez-Perez, 2004) sont plus destinés à valider des ontologies, alors que les critères définis en informatique médicale par le Canon Group (Evans *et al*, 1994) sont plus destinés à valider des systèmes terminologiques. Les méthodes d'évaluation doivent être adaptées en fonction de la complexité du texte mais plus les textes se complexifient plus le coût de l'expertise augmente en terme d'organisation, de durée, de financement et d'analyse.

Essayer d'identifier la totalité des concepts contenus dans une section de RCP permet de disposer d'une ressource pour développer de nouvelles fonctionnalités. La structuration de la pharmacodynamie des antibiotiques nous conduit actuellement à développer un outil inédit de comparaison des spectres bactériens avec visualisation des prévalences de résistance, destiné au médecin généraliste, avec le soutien financier de la Haute Autorité de Santé.

Références

- ALPHONSE E., AUBIN, S. BESSIERES P., BISSON G., HAMON T., LAGARRIGUE S., NAZARENKO A., MANINE A.P., NEDELLEC C., VETAH M., POIBEAU T. & WEISSENBACHER D. (2004). Event-based Information Extraction for the biomedical domain: the Caderige project. In *Proceedings of the International Workshop on Natural language Processing in Biomedicine and its Applications (JNLPBA)*. p. 43-49. Geneva. Suisse.
- BIEBOW B. (2004). De Daseart à Terminae : Voyage au pays de la modélisation des connaissances à partir de textes. Habilitation à diriger des recherches. Université Paris-Nord.
- BOURRIGAULT D (1995). LEXTER, a terminology extraction software for knowledge acquisition from texts. In *9th Knowledge Acquisition for Knowledge Based System Workshop* , Banff, Canada.
- CEAUSU V. & DESPRES S. (2004). Une approche mixte pour la construction d'une ressource terminologique (Tome 1). In *Actes des 15^{èmes} journées francophones d'Ingénierie des Connaissances*. p. 211-223. Lyon.

- CHARLET J. (2002). L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Habilitation à diriger des recherches - CHU Pitié-Salpêtrière. Université Pierre et Marie Curie.
- CHUTE, C., COHN, S., CAMPBELL, K., OLIVER, D., & CAMPBELL, J. (1996). The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *Journal of the American Medical Association*, 276(3), p.224 - 233.
- CIMINO, J., McNAMARA, T., MEREDITH, T., BROVERMAN, C., ECKERT, K., MOORE, M., & TYREE, D. (1999). Evaluation of a proposed method for representing drug terminology. *Proc AMIA annu Fall Symp*, p.47-51.
- DUCLOS, C., & VENOT, A. (2000). Structured representation of drug indications: lexical and semantic analysis of drug indications. *Methods of Information in Medicine*, 39(1), p.83 - 87.
- DUCLOS-CARTOLANO, C., & VENOT, A. (2003). Building and evaluation of a structured representation of pharmacokinetics information presented in SPCs: from existing conceptual views of pharmacokinetics associated with natural language processing to object-oriented design. *JAMIA*, 16(3), p. 271-280.
- DUCLOS C, CARTOLANO GL, GHEZ M, VENOT A. (2004). Structured representation of the pharmacodynamics section of the Summary of Product Characteristics for antibiotics : Application for automated extraction and visualization of their antimicrobial activity spectra. *JAMIA*, 11(4) , p.285 – 293.
- EVANS D, CIMINO J, HERSCHE W, HUFF S, BELL D. (1996) : Toward a medical concept representation language. The Canon Group. *JAMIA*, 1(3), p.207-217.
- GEORG G., SEROUSSI B., BOUAUD J. (2004). Synthesis of Elementary Single-Disease Recommandations to Support Guideline-Based Therapeutic Decision for Complex Polythological Patients. In *Proceedings of MEDINFO 2004*, p. 38-42. Amsterdam.
- GOMEZ-PEREZ A (2004). Ontology Evaluation.An overview of OntoClean. In *Handbooks on ontologies*. S. Staab, R. Studer eds. Springer, Berlin , p 251-273.
- GUARINO N, CHRISTOPHER AW (2004). An overview of OntoClean. In *Handbooks on ontologies*. S. Staab, R. Studer eds. Springer, Berlin , p. 151-171.
- LE MOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. In *Actes des 13^{ème} journées francophones d'Ingénierie des Connaissances*. p. 229-238. Rouen.
- LIU, J. H., MILSTEIN, C., SENE, B., & VENOT, A. (1998). Object-oriented modeling and terminologies for drug contraindications. *Methods of Information in Medicine*, 37(1), p. 45-52.
- NAZARENKO A. (2004). Donner accès au contenu des documents textuels : Acquisition de connaissances et analyse de corpus spécialisés. Habilitation à diriger des recherches. Université Paris-Nord.
- SOLOMON, W. D., WROE, C. J., RECTOR, A. L., ROGERS, J. E., FISTEIN, J. L., & JOHNSON, P. (1999). A reference terminology for drugs. *Proc AMIA Symp*, p. 152-156.
- TROMBERT-PAVIOT B., RODRIGUES J.M., ROGERS J.E., BAUD R., VAN DER HARING E., RASSINOUX A.M., ABRIAL, V., CLAVEL L. & IDIR H. (2000). GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *International Journal of Medical Informatics*. 58-59, p. 71-85.
- WAGNER J. (1993) Pharmacokinetics for the pharmaceutical Scientist. Basel, Technomic Publishing Company.
- ZWEIGENBAUM P. (1998). Traitement automatique de la langue médicale. Habilitation à diriger des recherches - CHU Pitié-Salpêtrière. Université Paris-Nord.

Construction guidée de graphes de transducteurs pour l'extraction d'évènements spatio-temporellement localisés

Manal EL Zant¹, Liliane Pellegrin¹, Michel Roux¹, Hervé Chaudet^{1,2}

¹Laboratoire d'Informatique Fondamentale, UMR CNRS 6166

Équipe BIM, Faculté de médecine, 27 Bd Jean – Moulin,

13005 Marseille

²Unité de Recherche Epidémiologique, Département de Santé Publique,

Institut de Médecine Tropicale du Service de Santé des Armées,

13998 Marseille Armées

{el.zant, liliane.pellegrin, michel.roux}@medecine.univ-mrs.fr
lhcp@acm.org

Résumé : Dans le cadre du traitement des dépêches épidémiologiques pour un système d'aide à la décision, nous présentons notre approche de l'extraction de ces informations. Celle-ci repose sur des graphes de transducteurs construits par le logiciel INTEX qui sont bâtis à partir d'une théorie de représentation spatio-temporelle des évènements obtenue lors d'un travail préalable, et non pas d'une grille a priori. Ces graphes ont été appliqués sur un corpus de 100 dépêches ProMed pour extraire les évènements et leurs caractéristiques spatio-temporelles. Si ces graphes permettent effectivement d'extraire les informations spatio-temporelles, des difficultés d'application sur des macro-événements propres au domaine épidémiologique restent à résoudre.

Mots-clés : Analyse de corpus textuel, extraction des connaissances, concept spatio-temporel, INTEX, Représentation des événements (STEEL).

1 Introduction

Dans le cadre de la pratique de la médecine des voyages et de l'épidémiologie, de nombreuses ressources Internet mettent régulièrement des informations concernant les phénomènes épidémiologiques dans le monde à disposition des experts. Parmi celles-ci la liste de diffusion internationale ProMED-Mail (<http://www.promedmail.org>) transmet régulièrement des informations sous la forme de dépêches, qui sont des textes spécialisés de longueur variable décrivant l'apparition, l'évolution et les caractéristiques épidémiologiques de tels phénomènes. Les informations épidémiologiques sont utilisées par des médecins, dans le cadre d'une activité générale de veille épidémiologique, qui correspond à la collecte et l'évaluation systématiques d'informations d'intérêt médical dans l'objectif de pouvoir déterminer les risques associés à un triplet <population,lieu,date>. Cela permet de déduire les conduites préventives associées à un déplacement individuel ou de population, ou de surveiller les émergences de maladies et leurs conséquences. Ce

raisonnement de haut niveau, qui s'effectue généralement dans un contexte où les contraintes temporelles sont fortes, est orienté par les connaissances et conduit à la construction d'une représentation adéquate de la situation afin de prendre la mesure des risques et d'en déduire les décisions adéquates (Chaudet, Pellegrin et Rech, 2000). Les représentations utilisées pour décrire les phénomènes épidémiologiques, rapportés dans le cas présent par les dépêches ProMED-Mail, sont donc complexes car elles doivent prendre en compte des informations de nature diverse, aussi bien des connaissances médicales, zoologiques, que socio-économiques ou encore géographiques et temporelles.

Le besoin de systèmes de question-réponse utilisant ce type de ressources textuelles semblerait une réponse « naturelle » en proposant à ces médecins d'alléger leur tâche de recherche d'information. De tels systèmes leur permettraient d'obtenir des réponses adéquates sur les phénomènes épidémiques sans avoir à parcourir un nombre croissant de dépêches disponibles. En effet, le poids de plus en plus important de la recherche et de l'acquisition de nouvelles informations est une constante dans les différentes spécialités médicales qui justifie ainsi le développement de ce type de système (Jacquemart et Zweigenbaum, 2003 Zweigenbaum, 2003). Dans ce cadre, l'objectif du projet EpidemIA est de bâtir un système d'aide à la décision utilisant toutes les caractéristiques des épidémies décrites dans les dépêches pour assister l'utilisateur dans son activité de gestion des risques sanitaires (figure 1).

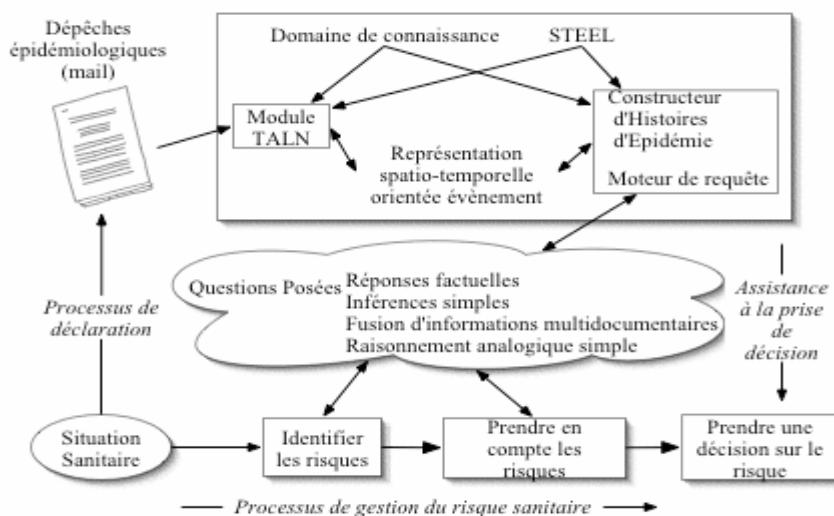


Fig.1 : Organisation générale du système EpidemIA

Ce projet comporte actuellement plusieurs volets :

- La mise au point d'un langage logique de représentation des concepts liés aux épidémies et de l'ontologie qui décrit ces concepts. Le résultat est le modèle général d'une épidémie.
- L'extraction des informations à partir des dépêches dans le module TALN

- La représentation de ces informations dans le langage logique.

Ces différents volets s'inséreront dans un système de question-réponse, qui comportera plus classiquement des modules d'analyse de questions et de génération de réponses aux utilisateurs. Dans le cas présent, il s'agira d'un système QR de domaine fermé puisqu'il est « uniquement » centré sur les phénomènes épidémiologiques (Doan-Nguyen et Kosseim, 2004).

L'hypothèse sous-jacente à ce projet est que les représentations adéquates des événements épidémiques décrits dans ces textes ne peuvent se construire que dans le cadre d'un système fortement orienté par les connaissances du domaine. Nous avons voulu nous appuyer sur l'ingénierie des connaissances pour construire le système en commençant par mettre au point le formalisme de représentation logique des connaissances afin de s'appuyer sur ce dernier pour bâtir la solution de TALN (traitement automatique de la langue naturelle). Un travail préalable nous a permis de développer un langage formel de représentation des connaissances (STEEL) adapté à la problématique des informations épidémiologiques, tenant compte à la fois de l'orientation évènementielle des récits, de la compositionnalité des évènements et de leur localisation spatio-temporelle (Chaudet, 2004). Le formalisme de représentation logique ayant ainsi été défini préalablement, il restait à mettre au point le mécanisme de traitement des textes pour obtenir leur représentation.

L'objectif de cet article est de présenter la solution que nous avons choisie en TALN pour traiter des dépêches épidémiologiques afin de construire une représentation du contenu de ces textes dans le langage formel. Que ce soit pour le traitement des réponses, celui des documents sélectionnés, ou encore la construction d'une réponse appropriée, le choix entre divers outils d'extraction d'informations ou de TALN est un enjeu majeur auquel se trouvent confrontés les systèmes de question-réponse. Dans notre cas, nous avons choisi d'utiliser des transducteurs à état finis lors de l'analyse des dépêches, ici le système INTEX (Silberstein, 1993). La construction de ces transducteurs a été guidée par le langage de représentation STEEL pour "identifier" ou "valider" les structures sémantiques, aboutissant ainsi à une construction du module de TALN guidée par l'ingénierie des connaissances. Dans ces structures, les informations syntaxico-sémantiques sont les éléments sémantiques nécessaires à la construction de la représentation spatio-temporelle des évènements épidémiques par STEEL.

2 Méthodes et contexte

2.1 STEEL : le langage de représentation des connaissances

Il existe de nombreux formalismes permettant de raisonner sur le temps en fonction des concepts de base (situation, événements, actions, chroniques) et la façon de représenter le temps (instants ou intervalles, temps linéaire ou non). Si le calcul des événements (Kowalski et Sergot, 1986) est un modèle déjà bien utilisé en médecine, il ne l'a été que dans le cadre de la clinique, et jamais pour la veille épidémiologique. Par ailleurs, la localisation spatiale n'a été que peu abordée et uniquement dans le cas

de la localisation de lésions. STEEL (Spatio-Temporal Extended Event Language) (Chaudet, 2004) est un langage typé en logique du premier ordre qui est une extension du calcul des événements spécialement développée pour la modélisation qualitative d'événements épidémiologiques. Ses principales caractéristiques sont de réifier simultanément les coordonnées spatiales et temporelles des événements sous la forme de localisations spatio-temporelle basée sur des régions pouvant être référencées par des noms, et de pouvoir représenter les regroupements d'événements sous la forme d'agrégats résultant de l'application d'opérateurs de combinaisons tout en maintenant la validité spatio-temporelle de l'agrégat. À l'occasion de la création de ces agrégats, il sait créer les regroupements corrects des temps et des lieux. Au lieu de la forme traditionnelle *happens(action, time)*, STEEL utilise la représentation *happens(macroevent, <time, place>)* où *macroevent* est un regroupement d'événements élémentaires. Cette extension permet d'obtenir une représentation très proche du récit, centrée sur les événements et conservant l'organisation de ces derniers en réseaux de dépendance et d'inclusion, ce qui permet de représenter une épidémie comme un événement complexe (mécanisme d'abstraction) correspondant à l'agrégation des événements qui la compose. L'objectif alors est de pouvoir interpréter naturellement, et désigner des événements complexes faisant intervenir le temps ainsi que le lieu, dans un langage de programmation en logique du premier ordre. Trois composantes du discours doivent donc être identifiées et représentées de façon coordonnée dans le langage de représentation : l'évènement (simple ou complexe), le temps et le lieu.

2.2 INTEX : un système de transducteurs à états finis

INTEX est un outil d'analyse, qui inclut des dictionnaires de type DELA et des grammaires, permettant l'exploration des textes à des fins d'études sur corpus, tout en utilisant une interface graphique pour construire des automates à états finis à types de transducteurs pour faire une analyse de texte. Il autorise une approche de traitement partiel qui recherche dans les textes des motifs particuliers au moyen d'automates et de relations entre ces motifs. INTEX est adapté à la construction d'extracteurs et de « shallow-parsers ». Il a la particularité de fournir aux utilisateurs un ensemble complet d'outils permettant de construire rapidement, et de gérer des centaines de grammaires locales, ainsi qu'une douzaine d'outils originaux permettant la vérification de chaque grammaire, la vérification de la cohérence entre plusieurs grammaires, leur débogage incrémental, etc. De plus sa capacité à créer des transducteurs permet de générer des informations qui modifient le texte d'entrée (p.e. tagging).

2.3 Annotation temporelle et spatiale des événements

L'annotation précise et détaillée des expressions temporelles a commencé avec les conférences MUC 5-7 (Message Understanding Conferences) pour l'identification et la classification des entités nommées EN (Chinchor 1999). Dans la même vision Ferro et al. (2001) décrivent un ensemble de guidelines pour l'annotation des

expressions temporelles, à partir des plusieurs langues, et leur associent une représentation canonique du temps auxquels elles se réfèrent. Cependant, une autre approche d'annotation a aussi été utilisée. C'est le marquage temporel, qui vise à associer un temps du calendrier à certains ou tous les événements du texte. Filatova et Hovy (2001) décrivent une méthode pour fractionner des phrases en leurs événements constitutifs et leur assigner des marqueurs temporels. Le marquage utilise deux temps principaux : le temps de l'article et le dernier temps indiqué dans la même phrase. Dans cette approche Schilder et Habel (2001) ont développé un système d'étiquetage sémantique des expressions temporelles. Elles sont classifiées selon deux types : celles qui se rapportent à un temps du calendrier ou d'horloge et celles qui se rapportent à des événements. L'ensemble des relations temporelles proposées est équivalent aux relations d'Allen (1983). Une troisième approche (Setzer et Gaizauskas, 2000) se centre sur les relations temporelles entre les événements et le temps ou entre les événements. Cette approche prend l'identification des relations temporelles comme but, et repose sur la façon dont l'information temporelle se présente ainsi que sa relation avec le texte. Leur schéma permet de déterminer l'ordre relatif ou le temps absolu des événements. Katz et Arosio (2001) à leur tour ont proposé une annotation des informations des relations temporelles en se basant sur les relations entre événement. Notre approche se rattache à la deuxième catégorie de travaux. L'annotation est pratiquée en tenant compte du temps de la dépêche et de celui signalé dans le récit. L'association entre l'événement et le marquage temporel se fait ultérieurement au niveau de la représentation logique.

Nous employons le terme localisation spatiale pour désigner les lieux géographiques par leurs noms de lieu (toponymes). Cette localisation spatiale peut être désignée par : des noms des villes, de villages, de provinces, des hôpitaux, des laboratoires, des aéroports, etc. Toute identification de la localisation spatiale nous intéresse puisque selon STEEL la localisation est un symbole attaché à une région de l'espace. L'annotation de la localisation spatiale, comme la temporelle relève du traitement du langage naturel MUC. Chinchor (1999) décrit une annotation, selon MUC, des différentes entités nommées où la localisation spatiale est identifiée par le type *Location*.

3 Application d'INTEX sur le corpus et résultats obtenus

Notre approche permet d'analyser des membres de phrases qui peuvent comporter une séquence longue relative à un concept donné. Nous avons procédé en 6 étapes :

- Identification des mots caractérisant les trois groupes de concepts décrivant les évènements épidémiques : type, localisation géographique et localisation temporelle de l'évènement.
- Construction des dictionnaires spécifiques pour les mots caractérisant chaque groupe, en ajoutant aux informations flexionnelles et lexicales, une information sémantique. Ex Géographie, Pathologie, Biologie,...
- Pour chaque groupe de concepts, sélection par INTEX dans le corpus des membres de phrases comportant un élément du groupe.

- Analyse de la configuration syntaxique et sémantique qui entoure l'élément choisi et identification le sous langage associé à l'élément choisi. INTEX construit les graphes correspondant aux membres de phrases sélectionnés, mais donne ainsi plusieurs solutions. Les ambiguïtés sont nombreuses, dues aux dictionnaires non spécifiques du domaine. Il est donc nécessaire de construire ces graphes après une analyse humaine apportant des précisions aux lexiques.
- Construction des graphes correspondant au sous langage caractérisant chaque concept tout en testant leurs performances.

3.1 Les différents types de graphes et leurs applications sur le corpus

Le calcul de la référence spatio-temporelle joue un rôle important dans la construction de leur représentation. Notre conception du raisonnement spatio-temporel permettrait de répondre automatiquement à deux questions principales suivantes : « Quand l'événement est-il survenu? », « Où l'événement est-il survenu? ». Il s'agit d'accorder à cette orientation deux fonctions. La première consiste en l'analyse du sous-langage, vue comme l'étape initiale qui permet d'accéder aux formes spatio-temporelles de notre corpus. La seconde, qui est conditionnée à la possibilité de représenter formellement la signification des expressions, définit et formalise la sémantique d'un certain nombre d'expressions du langage des dépêches en termes spatio-temporels. Finalement, il devrait être possible de valider ces représentations en répondant à des questions portant sur le contenu de la représentation des évènements (comme les localisations spatiales et temporelles).

3.2 Le graphe de localisation temporelle

Pour créer ce graphe, nous avons étudié les différentes formes d'expressions temporelles rencontrées dans les 100 dépêches. Elles ont été regroupées en deux catégories principales que nous détaillons ici.

Premièrement, des formes langagières spécifiques aux dépêches ont été identifiées. Il s'agit de plusieurs formats non littéraires, de style rédactionnel abrégé comme les expressions suivantes : *10 Apr 2003, 10/Apr/2003, Friday [10 Apr 2003], Friday, 10/04/03, 10 Apr 2003, 2003/04/10, 10-Apr-2003 etc...*

Un graphe de transducteurs a été construit pour identifier ces cas d'expressions temporelles (Fig.2). En particulier, l'étiquette **Month(\$MoisNum)** entraîne l'écriture du prédicat **Month** dans le texte d'origine, avec en argument la valeur de la variable **MoisNum** qui a été retrouvée. Dans l'échantillon de 100 dépêches, 6284 séquences sont reconnues par ce graphe. Nous présentons ci dessous quelques séquences reconnues, ainsi que leurs équivalents en mode de remplacement. L'inconvénient majeur de ce graphe est que nous ne pouvons supprimer quelques cas restants d'ambiguïtés comme, par exemple, le prénom d'une personne identique au nom d'un mois (3 cas identifiés, de type Jun Wang).

Deuxièmement, les cas d'expressions temporelles présentant des formes plus classiquement littéraires ont été identifiées dans notre corpus comme *In mid February, after several days, during the second week of February, last Friday night.*

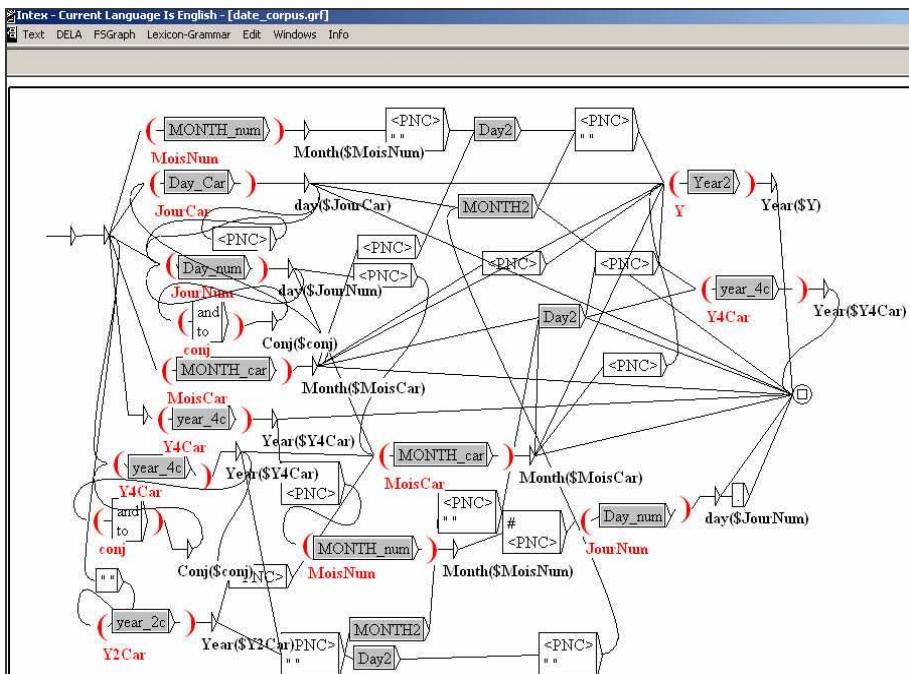


Fig. 2- Annotation des expressions temporelles

Séquence reconnues	Application du Mode de remplacement
Acute Respiratory Syndrome - Worldwide 20030315.0637 Acute respiratory ...	RespiratorySyndrome – Worldwide Year(2003)Month(03)day(15) 0637Acute...
...an outbreak, as happened in February [2003]. However, the government	seed an outbreak, as happened in Month(February)Year(2003)]. However, the....
SARS patients still in the hospital on Friday [18 Apr 2003], he said, 235 of them	patients still in the hospital on day(Friday) (day(18)Month(Apr)Year(2003)], he said,
...acute respiratory syndrome, and by 11-Mar-2003 similar outbreaks had beenacute respiratory syndrome, and by day(11)Month(Mar)Year(2003) similar...

Table 1. Exemples de séquences reconnues de résultats en mode de remplacement pour la localisation temporelle

Pour cela, nous avons bénéficié de la bibliothèque de graphes mise au point par Maurice Gross et disponible sur le site d'INTEX. Appliqués sur ces dépêches, ces graphes identifient toutes les formes littéraires de la langue anglaise présentes dans ce corpus. Cependant, il reconnaît à tort certaines expressions. Comme par exemple: « *that may have identified, from 16 countries, in a second Hong Kong hospital, in 65 cases, in terms of industries affected, on the 9th floor of the Metropole Hotel.* »

Pour réduire ces cas d'erreurs, nous avons fait des modifications dans les graphes concernés qui, ainsi modifiés, s'adaptent mieux au langage professionnel utilisé dans les dépêches.

Séquence Reconnues par INTEX	Résultat en mode de remplacement
in the evening on 10 Apr 2003	<u>ExpTemp(in the evening)day(10)Month(Apr))</u> <u>Year(2003),</u>
, Monday evening [24 Mar 2003]	<u>day(Monday)ExpTemp(Monday evening)day(24)</u> <u>Month(Mar) Year(2003),</u>
since Monday [21 Apr 2003]	<u>day(Monday)ExpTemp(since Monday)day(21)</u> <u>Month(Apr)Year(2003),</u>
, the same day (14 Apr 2003),	<u>ExpTemp(the same day)day(14)Month(Apr) Year(2003),</u>

Table 2. Exemples de séquences mixtes reconnues de résultats en mode de remplacement pour la localisation temporelle

Le graphe final qui englobe l'ensemble des formes associées aux expressions temporelles des dépêches est composé des deux graphes cités ci-dessus. 7087 cas ont pu être identifiés dans notre corpus par l'application de ce graphe, qui permet de traiter les cas incluant des expressions littéraires et non littéraires.

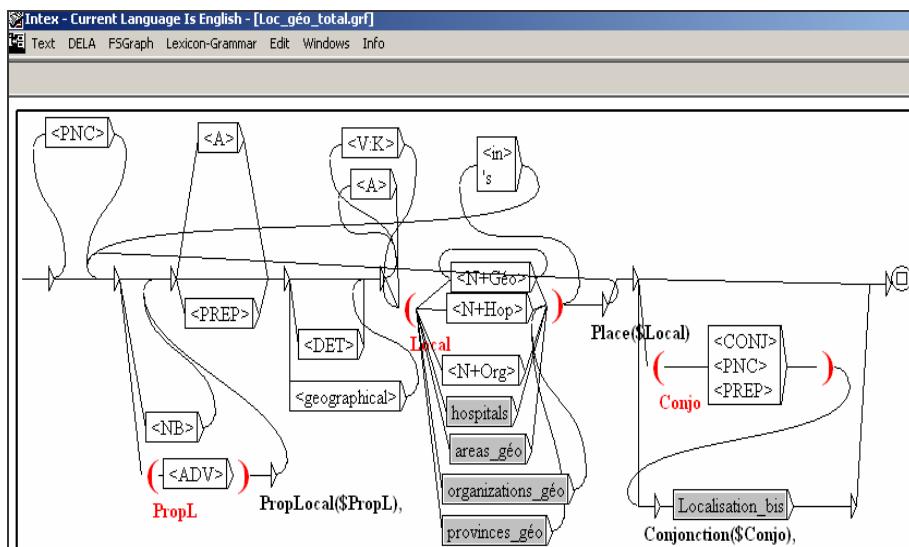


Fig.3 – Extraction des expressions spatiales.

3.3 Le graphe de localisation spatiale

Pour la création de ce graphe nous avons étudié les différentes indications géographiques existantes dans les dépêches. L'analyse nous a révélé que trois éléments sont regroupés :

1. le site géographique au sens classique comme le Nom de ville, de région, d'un pays, etc. (*Hong Kong, China, Hanoi, Canada, Toronto...*);
2. le nom d'un hôpital, d'une clinique, d'une laboratoire, etc. (*Kwong Wah Hospital, Alice Ho Miu Ling Nethersole Hospital (AHNH)*);
3. le nom d'un organisme (*Department of Health(DH), Ministry of Health (MOH), Centers for Disease Control and Prevention (CDC)*).

Nous avons typé les mots correspondant avec un attribut sémantique dans les dictionnaires. L'analyse du sous langage correspondant (grammaire locale) met en évidence la position des déterminants, des adverbes, des adjectifs ainsi que la possible association de plusieurs localisations spatiales, reliées ou non par des conjonctions ou des ponctuations (Fig.4). Par exemple : « *7 patients died among Alice Ho Miu Ling Nethersole Hospital (AHNH), Kwong Wah Hospital, Princess Margaret Hospital (PMH), Queen Elizabeth Hospital (QEHD), Tseung Kwan O Hospital (TKO)and Tuen Mun Hospital (TMH)* ».

L'étiquette **Place (\$Local)** provoque l'écriture du prédicat **Place** avec un argument, la valeur de la variable **Local**, valeur trouvée dans le texte. Ce graphe est récursif pour permettre la reconnaissance des plusieurs lieux cités successivement. Dans l'échantillon de 100 dépêches, 14656 séquences sont reconnues par ce graphe.

Séquence Reconnues par INTEX	Résultat en mode de remplacement
hospital in Hohhot, China ;	<u>Place(hospital in Place(Hohhot))</u> <u>Conjonction(\$Conjo).Localisation2(China);</u>
Centers for Disease Control and Prevention in Atlanta;	<u>Place(the Centers for Disease Control and Prevention in Localisation2(Atlanta));</u>
BEIJING - The World Health Organization;	<u>Place(BEIJING) Conjonction(\$Conjo).Org(The World Health Organization);</u>
from Guangdong province to Hong Kong.	<u>Place(Guangdong province)</u> <u>Conjonction(\$Conjo).Localisation2(Hong Kong);</u>

Table 3. Exemples de séquences reconnues de résultats en mode de remplacement pour la localisation spatiale.

3.4 Les graphes d'évènements épidémiques

Le problème est nettement plus complexe. De nombreux évènements attendus par STEEL, sont évoqués par des verbes d'action comme « admettre », « hospitaliser » (*Admit, hospitalize*). D'autres peuvent être identifiés par des noms communs issus du langage professionnel comme « épidémie » ou d'autres plus généraux mais appliqués dans un contexte linguistique spécifique comme le terme « lien » dans l'exemple suivant : *So far, no link has been found between these cases and the outbreak in Hanoi*. C'est bien « *link* » qui caractérise le fait de « la non-association entre les cas observés et l'épidémie en cours à Hanoi ». Nous avons abordé en priorité les verbes d'action à partir de la liste de tous les verbes trouvés dans le corpus (2854 formes verbales, flexions comprises). Ainsi, autour de ces verbes d'action, on trouve, outre

des notions de temps et de lieu qui sont traités à part, des descriptions de pathologies, de patients, de nombre de cas. Actuellement nous avons traité seulement 94 formes verbales différentes en 40 graphes.

A titre d'exemple, le graphe de la figure 4 représente le verbe d'action « *admit* » et décrit les différentes associations d'informations possibles autour du verbe. Ce graphe peut identifier des séquences comme : *admit new patients*, *admitted as suspect SARS cases*, *admitted for observation etc*. Il sélectionne 332 séquences dans notre échantillon de 100 dépêches.

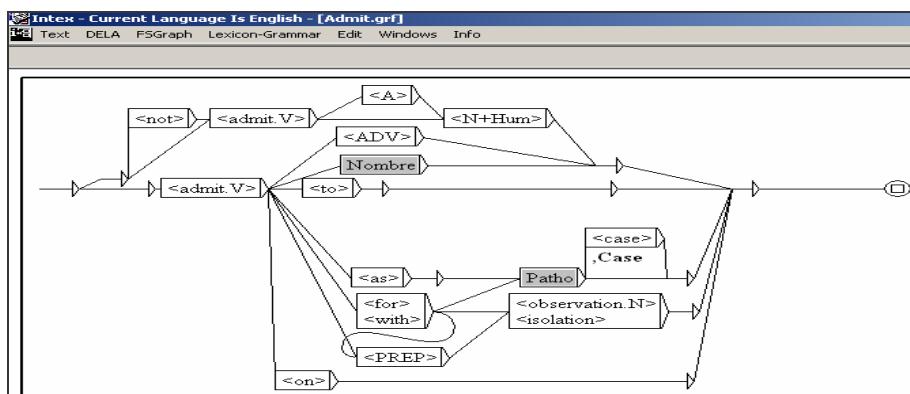


Fig.4 – Extraction des événements engendrés par l'événement « Admission » associé au verbe *Admit*.

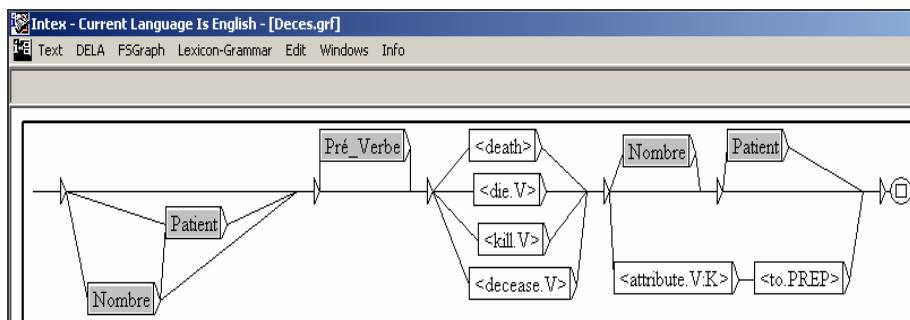


Fig.5 - Extraction des événements engendrés par l'événement « décès » associé aux expressions *death/kill/die...*

Un autre événement identifié dans les textes est le fait « décès » qui prend plusieurs formes linguistiques possibles (*death*, *kill*, *decease..*), elles-mêmes associées à d'autres informations comme « patients », « nombre d'individus » le type de personnes affectées, etc.. (*patient*, *34 people*, *46-years-old female*). Le graphe représentant les différentes formes de « Décès » a sélectionné dans le corpus 894 séquences. Voici quelques séquences sélectionnées : *a 46-year-old female died*, *an American businessman died*, *deaths Health care workers*, *more than 50 deaths*, *have been deaths in individuals*, *has killed at least 34 people*, *at least 9 deaths*.

4 Conclusion et Discussion

Les précédents travaux similaires portant sur le traitement des dépêches épidémiologiques (Damiano et col., 2002, Grishman et col., 2002) reposaient sur des systèmes d'extraction d'information demandant la définition à priori de grilles d'information utilisées par des automates d'identification de formes régulières. Cette approche limite à ces seules grilles l'information qui est extraite, ce qui restreint les services qui peuvent être rendus. Nous avons préféré introduire un certain relâchement dans les contraintes d'extraction d'information en faisant reposer la représentation non plus sur une grille mais sur une théorie, et nous avons répercuté les contraintes de cette théorie sur le dispositif d'analyse. Nous avons pu extraire un ensemble d'expressions temporelles et spatiales présentes dans le corpus, qui sont la base des descriptions de macro-événements du langage STEEL. L'avantage non négligeable d'INTEX est qu'il se fonde sur l'unité lexicale et applique ses transducteurs phrase à phrase. Cependant, de nombreux problèmes restent à résoudre. Nos premiers résultats montrent aussi la difficulté de composer des événements complexes, spécifiques au domaine d'application concerné, à partir d'éléments syntaxiques comme les verbes ou les noms communs même si ceux-ci se rapportent à de concepts majeurs dans l'ontologie du domaine. En effet, non seulement le système doit être capable d'identifier des événements, mais il doit aussi organiser les liens entre ces événements et la composition entre certains d'entre eux, dans un sens correct vis à vis du domaine. D'autant que ces événements ne sont pas systématiquement présents dans une seule phrase, ni même un seul paragraphe ou encore un seul texte mais peuvent se retrouver dans plusieurs textes. En effet, la description de l'évolution d'une épidémie et de ses caractéristiques cliniques, de diffusion dans une population, ne se fait pas en une seule dépêche mais au fil de l'ensemble des dépêches la concernant. Nous devons donc dans l'avenir élaborer une procédure spécifique qui transformera les résultats fournis par les transducteurs d'INTEX en langage STEEL. Ce dernier sera à même de synthétiser les informations issues de plusieurs dépêches.

Une solution serait, afin d'améliorer l'identification de ces différentes expressions, de passer à une forme générale, qui est obtenue grâce à la fonction de génération des transducteurs. On substituerait à toutes séquences d'origine, des séquences générées à partir de marqueurs définis (figés) dans les graphes et de mots ou marqueurs sémantiques récupérés par INTEX dans des variables, à partir du texte d'origine. Cette forme serait alors, dans une prochaine, l'étape traduite dans le langage STEEL. Une autre solution serait aussi de revenir vers une optique plus classique en TALN qui utiliserait des grammaires de type LinkParser permettant de faire plus facilement le lien entre les différentes structures syntaxiques support aux informations recherchées.

Références

- ALLEN, J.F. (1983). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26, 11, p. 832-843.

- CHAUDET H. (2004). Une extension du Calcul des Evènements pour la représentation de récits épidémiologiques. *15 ème journées francophones d'Ingénierie des Connaissances IC'2004*, p.285-296.
- CHAUDET H., PELLEGRIN L., RECH M. (2000). Polymorphisme des besoins d'informations dans le cadre d'une consultation assistée par hypertexte, in M. FIESCHI, O. BOUHADDOU, R. BEUSCART, R. BAUD, Eds; *L'informatique au service du patient*, Comptes rendus des Huitièmes Journées Francophones d'Informatique Médicale Marseille, 30-31 Mai 2000, Springer Verlag Editions. p.157-166. France.
- CHINCHOR N., BROWN, E., FERRO L., ROBINSON P. (1999). Named Entity Recognition Task Definition, *MITRE*, 1999.
- DAMIANOS L., DAY D. et col. (2002). Real users, real data, real problems : the MiTAP system for monitoring bio events. *Proceedings of BTR2002*. The University of Mexico, March 2002.
- DOAN-NGUYEN H. & KOSSEIM L. (2004). Amélioration de la précision dans un système de question-réponse de domaine fermé. *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT-2004)*. p. 325-333. Louvain-la-Neuve.
- FERRO, L., MANI, I., SUNDHEIM, B., WILSON, G. (2001). TIDES Temporal Annotation Guidelines Draft - Version 1.02. *MITRE Technical Report MTR MTR 01W000004. McLean, Virginia: The MITRE Corporation*.
- FILATOVA E. & HOVY E.H. (2001). Assigning Time-Stamp to Event-Clauses. In *Proceedings of the Workshop on Temporal and Spatial Reasoning at the Conference of the ACL*. Toulouse, France.
- GRISHMAN R., HUTTUNEN S., YANGARBER R. (2002), Information extraction for enhanced access to disease outbreak reports *Journal of Biomedical Infoirmatics*. 35. p.236-46.
- JACQUEMART P. & ZWEIGENBAUM P. (2003). Towards a medical question-answering system: a feasibility study. In R. BAUD, M. FIESCHI, P. LE BEUX, P. RUCH, Eds. Actes du colloque *The New Navigators: from Professionals to Patient Medical Informatics Europe*, Studies in Health Technology and Informatics, IOS Press, vol.95, p. 463-468, Amsterdam.
- KATZ G. & AROZIO, F. (2001). The Annotation of Temporal Information in Natural Language Sentences. *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001)*, Toulouse, p. 104--111. France.
- KOWALSKI R. & SERGOT M. (1986). A Logic-based Calculus of Event, *New Generation Computing*, 4, p.67-95
- LAIVENUS K. & LAPALME G. (2002) Evaluation des systèmes de question réponse. Aspects méthodologiques. *Traitement automatique des langues*, 43(3), p. 181-208.
- SCHILDER F. & HABEL C. (2001). From Temporal Expressions To Temporal Information: Semantic Tagging Of News Messages. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, ACL-2001. Toulouse, France, 6-11 July. pp. 65-72.
- SETZER A. & GAIZAUSKAS G. (2000). Annotating Events and Temporal Information in Newswire Texts. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, p. 1287-1294.
- SILBERSTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson: Paris.
- ZWEIGENBAUM, P. (2003) Question answering in biomedicine. In M. DE RIJKE & B. WEBBER, Eds. *Actes du Workshop Natural Language Processing for Question Answering*, EACL- 2003, pages 1-4, Budapest.

Aligner des ontologies lourdes : une méthode basée sur les axiomes

Frédéric Fürst, Francky Trichet

Laboratoire d'Informatique de Nantes Atlantique (CNRS-FRE 2729)
2 rue de la Houssinière - BP 92208 - 44322 Nantes CEDEX 03
{furst,trichet}@univ-nantes.fr

Résumé : L'alignement d'ontologies de domaine est une question centrale dans de nombreuses applications, en particulier le Web sémantique, nécessitant la prise en compte d'une interopérabilité sémantique entre ontologies existantes et couvrant des domaines connexes. Dans cet article, nous introduisons une nouvelle approche de l'alignement d'ontologies qui est essentiellement basée sur l'utilisation des descriptions intentionnelles des concepts et des relations, descriptions appelées *axiomes*. De ce fait, cette approche s'avère principalement dédiée à l'alignement d'ontologies dites *lourdes*. Cependant, elle peut également être appliquée aux ontologies dites *légères* et est ainsi complémentaire des approches plus courantes fondées sur l'analyse des expressions en langage naturel, des extensions des concepts et/ou des structures taxinomiques. Cette nouvelle approche est définie dans le cadre du modèle des Graphes Conceptuels.

Mots-clés : Alignement d'Ontologies, Graphes Conceptuels

1 Introduction

L'alignement d'ontologies de domaine est au cœur de la problématique de la gestion conjointe de plusieurs ontologies, devenue récemment nécessaire dans de nombreuses applications requérant une interopérabilité sémantique, en particulier le Web sémantique. Les stratégies utilisées pour aligner deux ontologies sont diverses. A titre d'exemple, citons la classification hiérarchique (Doan *et al.*, 2004), l'analyse formelle de concepts - Formal Concept Analysis - (Stumme & Maedche, 2001)), l'analyse terminologique des noms de concepts et/ou des définitions en langage naturel (Noy & Musen, 2003) ou l'analyse structurelle des taxonomies de concepts (Bach *et al.*, 2004). Néanmoins, la plupart des travaux actuels porte exclusivement sur des ontologies qualifiées de légères (« *lightweight ontologies* »), *i.e.* des ontologies composées uniquement de taxonomies de concepts et de relations (Gomez-Perez *et al.*, 2003). Aucun outil existant ne propose de fonctionnalités basées sur d'autres composants, en particulier les axiomes qui sont

pourtant essentiels à la spécification de la sémantique des concepts et relations d'une ontologie (Staab & Maedche, 2000).

Le travail présenté dans cet article vise à définir une nouvelle approche de l'alignement basée sur l'usage de tous les composants des ontologies lourdes, *i.e.* (1) structure hiérarchique des concepts, (2) structure hiérarchique des relations et (3) axiomes. Cette approche suppose que les axiomes soient explicitement représentés au niveau conceptuel, et non au niveau opérationnel comme cela est généralement le cas dans les travaux d'ingénierie ontologique. Par exemple, dans Protégé (Noy & Musen, 2003), les axiomes sont directement représentés sous une forme opérationnelle (*i.e.* une règle ou une contrainte dotée d'une sémantique opérationnelle prédéfinie) à l'aide du langage PAL. Du fait que les axiomes sous forme opérationnelle ne peuvent facilement être comparés, les méthodes d'alignement existantes utilisent uniquement les hiérarchies de concepts (parfois les hiérarchies de relations), et ne peuvent exploiter la richesse sémantique apportée par les axiomes, éléments caractéristiques des ontologies lourdes. Notre travail a pour ambition de contribuer à améliorer les techniques actuelles en proposant une approche complémentaire fondée sur (1) la représentation des axiomes au niveau conceptuel et (2) l'utilisation de ces axiomes pour la découverte d'appariements sémantiques entre concepts et relations.

Pour représenter les ontologies lourdes au niveau conceptuel, nous utilisons OCGL (*Ontology Conceptual Graphs Language*) (Fürst *et al.*, 2004), un langage de modélisation fondé sur le formalisme des Graphes Conceptuels (CGs) (Sowa, 1984), et permettant la représentation d'ontologies lourdes. La représentation des axiomes dans un modèle doté de mécanismes de raisonnement basés sur l'homomorphisme de graphes facilite leur comparaison topologique. Cet aspect est au cœur de la méthode que nous proposons : comparaison structurelle d'ontologies par représentation et raisonnement à l'aide de graphes.

La section 2 présente succinctement le langage OCGL. La section 4 introduit les éléments de base de notre méthode d'alignement. La section 4 expose les principes de l'algorithme qui l'implémente. La section 5 commente les résultats d'une expérimentation menée dans le contexte d'une ontologie des relations familiales. La section 6 compare notre approche avec les travaux existants.

2 Contexte du travail

2.1 Le langage de modélisation OCGL

Représenter une ontologie en OCGL (*Ontology Conceptual Graphs Language*) consiste principalement à (1) spécifier le vocabulaire conceptuel du domaine via des concepts, des relations et des instances ontologiques¹ et (2) spécifier la sémantique de ce vocabulaire à travers des schémas d'axiom et des axiomes de domaine (Fürst *et al.*, 2004).

1. Par exemple, en mathématiques, π est une instance ontologique du concept *Nombre*, car l'expression d'axiomes du domaine requiert cette instance. Mais 3.54, par exemple, n'est pas une instance ontologique.

Les **Schémas d’Axiome** proposés par OCGL sont : (1) les *liens ISA* entre deux concepts ou deux relations (propriété de subsomption), utilisés pour structurer les taxinomies de concepts et de relations (en arbres ou treillis), (2) l’*abstraction* d’un concept (appelée *Exhaustive-Decomposition* dans (Gomez-Perez *et al.*, 2003)), (3) la *disjonction* entre deux concepts, (4) la *signature* d’une relation, (5) les *propriétés algébriques* d’une relation (symétrie, réflexivité, transitivité, irréflexivité, antisymétrie), (6) l’*exclusivité* et l’*incompatibilité* entre deux relations (l’incompatibilité entre R_1 et R_2 est formalisée par $\neg(R_1 \wedge R_2)$, l’exclusivité par $\neg R_1 \Rightarrow R_2$) et (7) les *cardinalités* maximale et minimale d’une relation.

OCGL a été implémenté dans un outil, appelé TooCoM (*a Tool to Operationalize an Ontology with the Conceptual Graph Model*), dédié à l’édition et l’opérationnalisation d’ontologies de domaine (Fürst *et al.*, 2004) (TooCoM est disponible sous licence GNU GPL à <http://sourceforge.net/projects/toocom/>). La figure 1 présente un extrait de la hiérarchie des relations d’une ontologie des relations familiales.

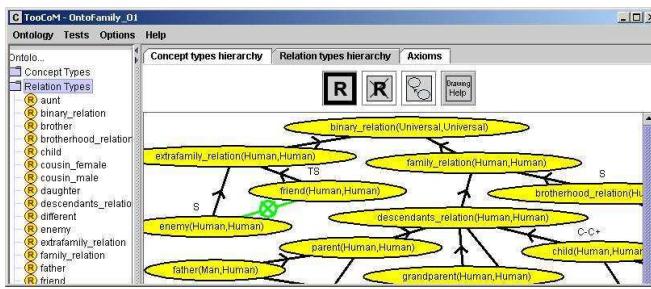


FIG. 1 – Représentation d’une hiérarchie de relations dans TooCoM. Une flèche symbolise un lien de subsomption, un cercle barré une exclusivité, une lettre une propriété algébrique (S pour la symétrie, T pour la transitivité, etc.).

Les **Axiomes de Domaine** diffèrent des schémas d’axiome dans le sens où ils sont totalement spécifiques au domaine considéré alors que les schémas d’axiome caractérisent des propriétés classiques des concepts et relations. La syntaxe graphique d’OCGL utilisée pour exprimer de tels axiomes est basée sur le modèle des Graphes Conceptuels. Ainsi, un axiome est composé d’une partie Antécédent et d’une partie Conséquent, la sémantique formelle d’une telle construction pouvant s’exprimer intuitivement comme suit : *si la partie Antécédent est vraie, alors la partie Conséquent est vraie*. La figure 2 présente le graphe représentant l’axiome « *les ennemis de mes amis sont mes ennemis* ».

2.2 Le domaine d’expérimentation OntoFamily

Afin d’illustrer nos idées, nous considérons dans cet article le domaine simple (mais intuitif) des relations familiales. Ce domaine volontairement limité inclut les notions suivantes : *père, mère, grand-père, grand-mère, fils, fille, cousin, cousine, neveu, nièce, oncle, tante, sœur, frère, épouse, mari, ami, ennemi*. En plus

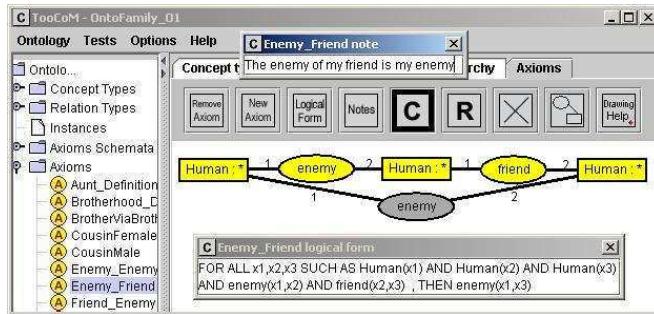


FIG. 2 – Représentation d'un axiome dans *TooCoM*. Les nœuds clairs composent la partie antécédent, les nœuds foncés la partie conséquent. Chaque partie contient des nœuds concepts (rectangles) et des nœuds relations (ellipses).

d'être facilement compréhensible, ce domaine est particulièrement intéressant du fait qu'il requiert la définition d'axiomes de domaine (par exemple, « *les ennemis de mes ennemis sont mes amis* »), non représentables par des schémas d'axiome. Dans le cadre d'un cours d'Ingénierie Ontologique, deux groupes d'étudiants de niveau Master ont travaillé séparément à la définition d'une ontologie de ce domaine. Cette expérience a conduit à la construction de deux ontologies : OntoFamily O_1 et OntoFamily O_2 ². O_1 est composée de 3 types de concepts qui définissent une partition (*Human*, un concept abstrait, et ses 2 sous-concepts *Man* et *Woman* qui sont disjoints), 31 relations binaires structurées en un treillis de profondeur 3, 11 schémas d'axiomes (1 abstraction, 1 disjonction, 1 exclusivité, 3 symétries, 4 cardinalités et 1 transitivité), et 18 axiomes. O_2 est composée de 3 types de concepts (*Human*, qui n'est pas abstrait, et ses 2 sous-concepts *Male* et *Female* qui sont disjoints), 23 relations binaires structurées en un arbre de profondeur 2, 10 schémas d'axiomes et 27 axiomes.

3 Alignement basé sur les axiomes : les principes

L'objectif du processus d'alignement d'ontologies est de découvrir et d'évaluer des liens d'identité entre primitives conceptuelles (concepts et relations) de deux ontologies données, supposées être construites à partir de domaines connexes. La méthode que nous proposons repose sur l'utilisation du niveau axiomatique des ontologies pour découvrir des analogies sémantiques entre primitives, de façon à révéler des identités entre elles et à calculer leur coefficient de similarité, *i.e.* le coefficient qui indique la proximité sémantique de deux concepts ou relations. Deux principes régissent notre approche : (1) l'utilisation de la *stabilité de modélisation* et de la *rareté* d'une propriété conceptuelle pour fixer le poids de

2. Ces ontologies sont disponibles en CGXML sur <http://sourceforge.net/projects/toocom/>. CGXML est le format XML de stockage d'OCGL.

cette dernière dans l'évaluation des appariements et (2) l'utilisation des *méta-représentations* des axiomes de domaine pour comparer leurs structures.

3.1 Postulats concernant la stabilité et la rareté

Les propriétés conceptuelles d'un domaine, exprimées en OCGL par des schémas d'axiome et des axiomes de domaine, peuvent être modélisées de différentes façons. Par exemple, deux hiérarchies de concepts peuvent être différentes si des concepts additionnels sont ajoutés ou non pour les structurer (par exemple le concept *Human* aurait pu être omis dans OntoFamily O₁). De façon similaire, la signature d'une relation peut être différente d'une modélisation à l'autre (par exemple, la notion de *cousin* peut être représentée par seulement une relation *cousin(Human, Human)*, ou par deux relations en introduisant le genre *cousinF(Woman, Human)* et *cousinM(Man, Human)*). La stabilité de modélisation d'une propriété (schéma d'axiome ou axiome de domaine) correspond à son degré de stabilité d'une ontologie à l'autre. La recherche de correspondances entre ontologies doit favoriser les propriétés possédant les meilleures stabilités de modélisation, les analogies entre ces propriétés étant les plus pertinentes.

De plus, au sein d'une même ontologie, une propriété peut être très commune ou, au contraire, extrêmement rare. Par exemple, la propriété de symétrie est très commune. Mais un axiome de domaine est, par définition, très particulier et seulement quelques axiomes de même sémantique formelle peuvent exister dans des ontologies portant sur un domaine donné. La rareté d'une propriété dans les ontologies considérées accroît le poids des appariements découverts grâce à cette propriété. Donc, au début du processus d'alignement, ces raretés doivent être évaluées de façon à adapter le poids de chaque propriété aux caractéristiques des ontologies à aligner.

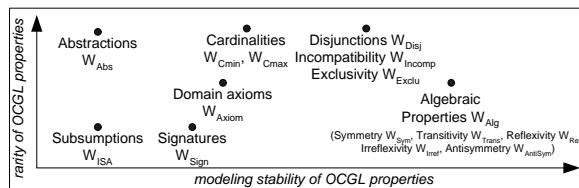


FIG. 3 – Stabilités de modélisation et raretés des propriétés d'OCGL. Les *subsumptions* et les *signatures* sont très communes et peu stables. Les *axiomes de domaine* et les propriétés algébriques sont moins répandues, et la stabilité des *axiomes de domaine* est plus faible que celle des propriétés algébriques du fait que les représentations des propriétés algébriques sont fixées. Les *cardinalités* sont moins stables que les propriétés qui lient deux concepts ou relations (*disjonction, incompatibilité et exclusivité*), car à la fois leurs valeurs et les relations sur lesquelles elles portent peuvent varier.

La figure 3 présente les valeurs relatives de stabilité et de rareté des propriétés

d'OGCL. Ces évaluations des stabilités et raretés ne sont que des postulats permettant de fixer les poids par défaut pour chaque propriété. Ces poids peuvent être modifiés par l'utilisateur dans l'outil pour améliorer les résultats du processus d'alignement. Par défaut, les valeurs des poids ont été ordonnées comme suit, à partir de considérations expérimentales : $W_{Alg}(W_{Sym}, W_{Trans}, W_{Refl}, W_{Irref}, W_{AntiSym}) > W_{Disj} = W_{Incomp} = W_{Exclu} > W_{Cmin} = W_{Cmax} > W_{Axiom} > W_{Sign} > W_{Abst} > W_{ISA}$.

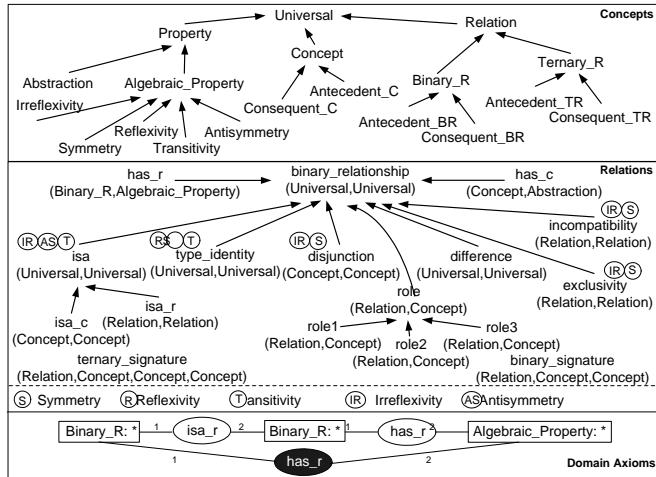


FIG. 4 – Les hiérarchies de concepts et relations de MetaOCGL. L’axiome présenté (en bas de la figure) exprime l’héritage d’une propriété algébrique.

3.2 MetaOCGL : une ontologie de représentation

Afin de détecter des analogies entre axiomes représentés par des graphes, et ainsi de détecter des analogies entre les primitives correspondant aux nœuds des graphes, les axiomes de domaine sont transcrits sous une forme plus abstraite, qui préserve les structures topologiques des graphes initiaux. Ces représentations abstraites sont basées sur une ontologie du langage OCGL, exprimée en OCGL, et appelée **MetaOCGL**. Comme indiqué en figure 4, MetaOCGL inclut tous les concepts d'OCGL et ses relations (relation *isa*, exclusivité/incompatibilité entre relations, disjonction de concepts, liens entre relations et concepts dans un graphe qui exprime un axiome). La figure 5 présente les graphes OCGL dédiés à la représentation des deux axiomes « *les ennemis de mes ennemis sont mes amis* » et « *les ennemis de mes amis sont mes ennemis* », ainsi que les métagraphes correspondant en MetaOCGL.

Les comparaisons entre axiomes représentés en MetaOCGL sont réalisées grâce à l'opérateur de *projection* du modèle des Graphes Conceptuels, opération correspondant à un homomorphisme de graphes : étant donné deux graphes G_1 et

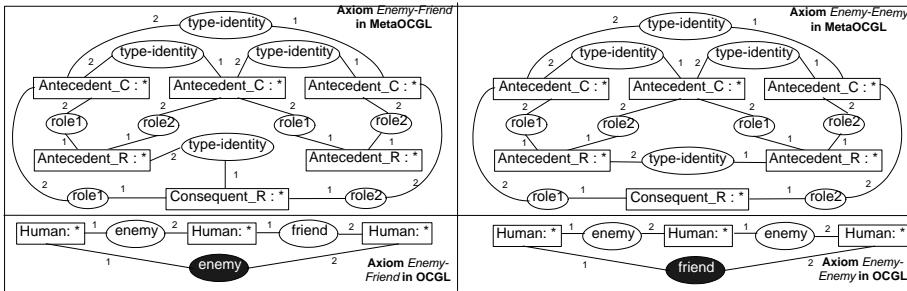


FIG. 5 – Deux axiomes d’OntoFamily O_1 représentés en MetaOCGL. Le lien type_identity indique que les nœuds ont le même type dans l’axiome.

G_2 , qui représentent en MetaOCGL deux axiomes A_1 et A_2 , si deux projections existent de G_1 dans G_2 et de G_2 dans G_1 , alors A_1 et A_2 ont la même structure ; dans ce cas, les axiomes A_1 et A_2 expriment le même type de propriété, et l’analogie entre les deux axiomes peut être étendue aux primitives qui apparaissent dans les axiomes.

4 Alignement basé sur les axiomes : l’algorithme

Notre algorithme prend en entrée deux ontologies O_1 et O_2 , représentées en OCGL, et produit en sortie des similarités potentielles entre 2 concepts ou 2 relations. Le résultat est un ensemble d’appariements (p_i, p'_j, c) , où p_i et p'_j sont respectivement des primitives conceptuelles (concepts ou relations) de O_1 et O_2 , et c le coefficient de similarité entre p_i et p'_j . Bien entendu, étant donné une primitive p_i de O_1 , plusieurs appariements peuvent exister avec des primitives de O_2 (ou aucun), et vice versa. Les schémas d’axiome et les axiomes de domaine sont utilisés pour évaluer ou découvrir des appariements de primitives. Le poids de chaque propriété d’OCGL est utilisé pour moduler son influence sur l’évaluation des appariements.

4.1 Utiliser les schémas d’axiome pour évaluer les appariements

Les schémas d’axiome qui impliquent seulement une primitive (*i.e.* propriétés algébriques et abstractions) sont comparées de O_1 à O_2 , afin de découvrir des appariements de primitives. Si les primitives de l’appariement (p_1, p_2, c) portent toutes deux un même schéma d’axiome, le coefficient c est augmenté du poids du schéma ; si l’appariement n’existe pas, il est créé avec ce poids ; si seule une des primitives porte le schéma, c est diminué du poids du schéma (ou l’appariement est créé avec le poids négatif s’il n’existe pas). Les schémas d’axiome qui impliquent deux primitives (*i.e.* disjonctions, incompatibilités et exclusivités) sont utilisés de même, mais concernent les 4 appariements possibles des 2 primitives.

	Relation r_1 in Ontology 1	Relation r_2 in Ontology 2	Action
Min Card	0 (resp. $c_{min} \geq 1$)	$c_{min} \geq 1$ (resp. 0)	$-2 * W_{cmin}$
	$c_{min} \neq 0$	$c_{min} \neq 0$	$+2 * W_{cmin}$
	$c_{1min} \neq 0$	$c_{2min} \neq 0$ and $\neq c_{1min}$	$-W_{cmin}$
Max Card	∞ (resp. $c_{max} \geq 1$)	$c_{max} \geq 1$ (resp. ∞)	$-W_{cmax}$
	$c_{max} \neq \infty$	$c_{max} \neq \infty$	$+2 * W_{cmax}$
	$c_{1max} \neq \infty$	$c_{2max} \neq \infty$ and $\neq c_{1max}$	$-2 * W_{cmax}$

TAB. 1 – *Modifications du coefficient de l'appariement (r_1, r_2, c) en fonction des cardinalités des relations. c_{min} et c_{max} sont les valeurs des cardinalités.*

Le tableau 1 présente les différentes opérations induites par la prise en compte des analogies ou différences de cardinalités. Si l'appariement entre les relations considérées n'existe pas alors qu'une analogie entre cardinalités est découverte, l'appariement est créé avec le coefficient correspondant.

4.2 Utiliser les axiomes de domaine pour évaluer les appariements

Pour chaque couple d'axiomes (a_1, a_2) , où $a_1 \in O_1$ et $a_2 \in O_2$, les représentations de a_1 et a_2 en MetaOCGL, $meta(a_1)$ et $meta(a_2)$, sont construites. Ces représentations peuvent être enrichies par l'ajout de relations d'identité de type entre les noeuds (cf. figure 5). Deux types d'équivalence topologique sont considérés : l'**équivalence**, qui est attestée lorsque deux projections existent de $meta(a_1)$ dans $meta(a_2)$ et de $meta(a_2)$ dans $meta(a_1)$, sans considérer les relations *type_identity*, et l'**équivalence typée** attestée quand deux projections existent en tenant compte des relations *type_identity*. Bien entendu, le poids d'une équivalence typée est plus élevé que celui d'une équivalence. Une équivalence typée (resp. équivalence) entre deux axiomes augmente le coefficient des noeuds liés par projection, du poids de l'équivalence typée (resp. équivalence). Si aucune projection n'existe (ou seulement une), aucune modification n'intervient. Par exemple, les deux axiomes de la figure 5 sont équivalents car 2 projections existent entre leurs métagraphes sans considérer les relations *type-identity*. En considérant les relations *type-identity*, il n'existe pas de projection, les axiomes ne sont donc pas équivalents typés.

4.3 Résoudre les conflits d'appariements

Du fait qu'elles lient une relation et un ensemble de concepts, les signatures sont uniquement utilisées pour augmenter ou diminuer les coefficients des appariements, et non pour en créer de nouveaux. En outre, deux usages possibles sont distingués : (1) la modification des coefficients d'appariements de relations à partir des appariements entre concepts de leurs signatures et (2) la modification des coefficients d'appariements de concepts à partir des appariements entre relations possédant ces concepts dans leurs signatures. Par exemple, si les relations *mother1(Woman1,Human1)* de O_1 et *parent2(Human2,Human2)* de O_2 sont appariées, les coefficients des appariements de concepts (*Woman1,Human2*)

et (*Human1,Human2*), s'ils existent, sont augmentés de W_{Sign} . Mais si les appariements (*Woman1,Human2*) et (*Human1,Human2*) existent, le coefficient de l'appariement de relations (*mother1,parent2*) peut aussi être augmenté de W_{Sign} . Le choix entre ces deux possibilités est basé sur l'utilisation d'un **Matching Ratio** appelé MR_c (resp. MR_r) et défini entre le nombre d'appariements de concepts (resp. relations) et le nombre moyen de concepts (resp. relations) dans O_1 et O_2 : si MR_c est supérieur à MR_r (*i.e.* la proportion d'appariements de concepts découverts est supérieure à celle des appariements de relations), les signatures sont utilisées pour réévaluer les coefficients des appariements de relations existant à l'aide des appariements de concepts existant, et vice versa. Ainsi, avant d'utiliser les signatures, les appariements de concepts (ou de relations) doivent être résolus, et ensuite les signatures sont utilisées pour affiner les coefficients des appariements de relations (ou de concepts).

La résolution des appariements de concepts ou de relations repose sur les valeurs de leurs coefficients : quand plusieurs appariements existent pour une primitive donnée, l'appariement avec le coefficient le plus élevé est retenu. Bien entendu, des contradictions peuvent exister entre 2 ou plusieurs appariements, que les valeurs des coefficients ne permettent pas toujours de résoudre. Dans ce cas, l'intervention de l'utilisateur est nécessaire pour trancher.

5 Résultats expérimentaux

L'application de notre algorithme sur OntoFamily O_1 et OntoFamily O_2 (avant la phase de résolution des appariements) produit 201 appariements, dont 9 couples de concepts (parmi 9 possibles), et 192 couples de relations (parmi 713 possibles)³. Beaucoup d'appariements n'apparaissent qu'à l'occasion de moins de 4 comparaisons d'axiomes, et jamais lors de la comparaison de schémas d'axiomes. Ils sont donc rejettés du fait de leur faible pertinence et seuls 9 couples de concepts et 69 couples de relations sont ainsi considérés pour la phase de résolution. Les valeurs des rapports MR_c et MR_r sont respectivement de 3 et ≈ 2.55 , les couples de concepts sont donc résolus avant les couples de relations. Le couple (*Human,Human*) possède le coefficient le plus élevé des appariements liant les concepts *Human*. Ensuite, le couple de plus fort coefficient est (*Woman,Female*) et le troisième couple de plus fort coefficient est (*Man,Male*). Tous les liens entre concepts des 2 ontologies sont donc retrouvés par l'algorithme.

Les signatures sont ensuite utilisées pour affiner les valeurs des coefficients des couples de relations. La résolution des couples de relations ainsi constitués conduit à 9 liens entre relations de O_1 et O_2 (parmi 17 liens réellement possibles) : (*aunt,aunt*), (*daughter, daughter*), (*enemy,enemy*), (*father,father*), (*friend,friend*), (*husband,husband*), (*son,son*), (*uncle,uncle*) et (*wife,wife*). Ces résultats, bien que perfectibles, s'avèrent encourageants étant donné le faible niveau de complexité des 2 ontologies considérées. En effet, notre approche pourrait

3. Un rapport détaillé de cette expérimentation est disponible sur <http://sourceforge.net/projects/toocom/>.

s'avérer encore plus pertinente dans un contexte d'ontologies plus hétérogènes et plus riches à la fois en termes de concepts, relations et axiomes. Cette expérience a également montré que l'usage des subsomptions peut améliorer l'efficacité de l'algorithme. Par exemple, les couples de relations (*parent,mother*) et (*parent,father*) ont de forts coefficients, ce qui pourrait permettre de déduire l'appariement (*parent,parent*).

6 Discussion

Actuellement, de nombreux outils sont proposés pour découvrir des correspondances entre ontologies (Kalfoglou & Schorlemmer, 2003). La première façon de classifier ces outils est de considérer l'objectif visé: (1) fusionner deux ontologies pour en créer une nouvelle (*e.g.* iPROMPT (Noy & Musen, 2003)), (2) définir une fonction de transformation d'une ontologie à l'autre (*e.g.* OntoMorph (Chalupsky, 2000)) ou (3) déterminer des correspondances entre concepts (*e.g.* ANCHORPROMPT (Noy & Musen, 2003), GLUE (Doan *et al.*, 2004), S-MATCH (Giunchiglia *et al.*, 2004), FCA-Merge (Stumme & Maedche, 2001) ou ASCO (Bach *et al.*, 2004)). Notre travail s'inscrit dans le cadre de ce dernier objectif tout en y ajoutant la détermination de correspondances entre relations. Une autre façon de caractériser les outils est de considérer le type de représentations utilisées: (1) les noms de classes et les définitions en langage naturel, (2) les hiérarchies de classes et leurs propriétés (ANCHORPROMPT, GLUE), (3) les instances de classes (FCA-Merge) ou (4) les descriptions de classes (S-MATCH). Certains outils s'appuient sur plusieurs représentations, par exemple l'algorithme ASCO utilise à la fois (1), (2) et (4) (Bach *et al.*, 2004). Notre approche repose à la fois sur (2) et (4) et ajoute un nouveau type d'entrée: les axiomes de domaine.

Ensuite, dans (Ehrig & Sure, 2004), les différentes mesures utilisées pour l'alignement d'ontologies sont organisées au sein d'une échelle de similarité à 5 niveaux, intégrant des représentations de plus en plus élaborées dans les mesures : *Entities*, *Semantic Nets*, *Description Logics*, *Restrictions* et *Rules*. Au premier niveau, deux concepts sont égaux s'ils ont même(s) label(s); au second niveau, deux concepts sont égaux s'ils ont mêmes attributs; au troisième niveau, deux concepts sont égaux s'ils ont les mêmes concepts parents. Cependant, pour les niveaux *Restrictions* et *Rules*, aucune mesure n'est proposée. Notre travail doit être considéré comme une extension de cette classification dans le sens où il offre une mesure de similarité basée sur les axiomes de domaine, qui englobent à la fois les niveaux *Restrictions* et *Rules*, puisque nous estimons qu'il n'est pas possible de considérer les règles et les contraintes au niveau ontologique, et que, de notre point de vue, les deux niveaux *Restrictions* et *Rules* doivent être réunis dans un unique niveau baptisé *Axioms* (Fürst *et al.*, 2004).

Enfin, afin de comparer les différentes méthodes de calcul de distances entre entités, (Euzenat & Valtchev, 2004) proposent la taxinomie suivante :

- *Terminological (T)*: comparaison des labels des entités, décomposée en

(TS), approche purement syntaxique et (TL) utilisant en plus un lexique de synonymes et d'hyponymes ;

- *Internal structure comparison (I)*: comparaison des structures internes des entités (par exemple la valeur d'une cardinalité) ;
- *External structure comparison (S)*: comparaison des relations entre entités, décomposée en (ST), comparaison des entités au sein de leurs taxinomies et (SC) comparaison des structures externes en tenant compte des cycles ;
- *Extensional comparison (E)*: comparaison des extensions des entités ;
- *Semantic comparison (M)*: comparaison des interprétations (plus exactement des modèles) des entités.

L'originalité de notre travail est de s'attaquer à la *Semantic Comparaison (M)*, via l'utilisation des axiomes de domaine. Comme rappelé dans (Euzenat & Valtchev, 2004), la seule approche travaillant sur les interprétations des entités est l'algorithme de (Giunchiglia *et al.*, 2004), qui utilise un moteur de raisonnement pour déterminer des subsomptions ou équivalences de classes à partir des équivalences initiales de quelques classes et de l'analyse des relations taxonomiques. Notre travail se différencie de cette approche par la prise en compte non seulement de la relation *sorte-de*, mais également de toutes les représentations sémantiques des entités incluant les schémas d'axiome et les axiomes de domaine. Il est à noter que nous considérons également I et ST, mais pas E. Ce travail étend certains travaux visant à rapprocher deux ontologies exprimées dans le modèle des GCs, mais ne considérant que les hiérarchies de concepts et de relations (et les instances de concepts), et basés sur la comparaison des labels des primitives et leur contexte (place dans les hiérarchies) (Dieng & Hug, 1998).

7 Conclusion

Dans cet article, nous avons introduit une nouvelle approche de l'alignement d'ontologies, principalement basée sur l'utilisation des axiomes. Cette approche a été définie dans un contexte particulier où les composants des ontologies sont représentés à l'aide de graphes et manipulés via les mécanismes de raisonnement du modèle des Graphes Conceptuels. Contrairement aux méthodes d'alignement proposées jusqu'alors, qui pour la plupart considèrent uniquement un sous-ensemble des composants des ontologies légères, notre méthode a pour avantage de prendre en compte tous les composants des ontologies lourdes, vers lesquelles tendent de nombreux travaux, en particulier dans le cadre du Web sémantique.

Ce travail préliminaire se poursuit actuellement sur l'étude de la contribution possible des liens de subsomption et des instances à la méthode d'alignement proposée. Nous étudions également l'apport que pourrait constituer la prise en compte du processus inverse de celui présenté dans ce papier, c'est-à-dire non plus utiliser les axiomes pour découvrir et évaluer des identités entre

concepts/relations, mais utiliser les identités précédemment découvertes entre concepts/relations pour découvrir et évaluer des identités entre axiomes.

Références

- BACH T., DIENG-KUNTZ R. & GANDON F. (2004). On Ontology Matching Problems - for Building a Corporate Semantic Web in a Multi-Communities Organization. In *Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS'2004)*, volume 4, p. 236–243.
- CHALUPSKY H. (2000). OntoMorph: A Translation System for Symbolic Knowledge. In *Principles of Knowledge Representation and Reasoning*, p. 471–482.
- DIENG R. & HUG S. (1998). Comparison of « Personal Ontologies » Represented through Conceptual Graphs. In H. PRADE, Ed., *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'1998)*, p. 341–345: John Wiley and Sons.
- DOAN A., MADHAVAN J., DOMINGOS P. & HALEVY A. (2004). Ontology Matching: A Machine Learning Approach. In *Handbook on Ontologies in Information Systems*, p. 397–416.
- EHRIG M. & SURE Y. (2004). Ontology Mapping - an integrated approach. In *Proceedings of the First European Semantic Web Symposium*, p. 76–91: Springer-Verlag (LNCS 3053).
- EUZENAT J. & VALTCHEV P. (2004). Similarity-based ontology alignment in OWL-Lite. In R. L. DE MANTARAS & L. SAITTA, Eds., *European Conference on Artificial Intelligence (ECAI'2004)*, p. 333–337: IOS Press.
- FÜRST F., LECLÈRE M. & TRICHET F. (2004). Operationalizing domain ontologies: a method and a tool. In R. L. DE MANTARAS & L. SAITTA, Eds., *European Conference on Artificial Intelligence (ECAI'2004)*, p. 318–322: IOS Press.
- GIUNCHIGLIA F., SHVAIKO P. & YATSKEVICH M. (2004). S-Match: an Algorithm and an Implementation of Semantic Matching. In *Proceedings of the First European Semantic Web Symposium*, p. 61–65: Springer-Verlag (LNCS 3053).
- GOMEZ-PEREZ A., FERNANDEZ-LOPEZ M. & CORCHO O. (2003). *Ontological Engineering*. Springer, Advanced Information and Knowledge Processing.
- KALFOGLOU Y. & SCHORLEMMER M. (2003). Ontology mapping: the state of the art. *The Knowledge Engineering Review*, **18**(1), 1–31.
- NOY N. F. & MUSEN M. (2003). The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, **59**(6), 983–1024.
- SOWA J. (1984). *Conceptual Structures : information processing in mind and machine*. Addison-Wesley.
- STAAB S. & MAEDCHE A. (2000). *Axioms are objects too: Ontology Engineering beyond the modeling of concepts and relations*. Research report 399, Institute AIFB, Karlsruhe.
- STUMME G. & MAEDCHE A. (2001). FCA-MERGE: Bottom-up merging of ontologies. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'2001)*, p. 225–234.

Comment ne pas perdre de vue les usage(r)s dans la construction d'une application à base d'ontologies ? Retour d'expérience sur le projet KmP

Alain Giboin¹, Fabien Gandon¹, Nicolas Gronnier¹, Cécile Guigard^{2 et 1}, et
Olivier Corby¹

¹ Projet Acacia, INRIA, Sophia Antipolis
`{giboin|gandon|gronnier|corby}@sophia.inria.fr`
² Telecom Valley, Sophia Antipolis
`{guigard}@sophia.inria.fr`

Résumé : Le constructeur d'une application à base d'ontologies qui veut réaliser une application *utile* et *utilisable* devrait se donner comme ligne directrice non pas simplement d'avoir en vue les usage(r)s de son application, mais surtout de *ne pas les perdre de vue* au cours du processus de construction. Comment ce constructeur peut-il faire pour ne pas perdre de vue les usage(r)s ? On fournit ici quelques éléments de réponse à cette question, qui reposent sur notre expérience du projet pluridisciplinaire RNRT KmP de serveur Web sémantique de compétences inter firmes.

Mots-clés : Analyses d'usage, Applications à base d'ontologies, Gestion des connaissances, Gestion des compétences, Ontologies dans leur contexte d'usage.

1 Introduction

Pour ne pas perdre de vue l'idée maîtresse des essais qu'il composait, l'écrivain français Julien Benda (1867-1956) avait l'habitude d'utiliser une technique astucieuse, qu'on pourrait appeler la *technique du carton sur le chevalet*. Benda¹ décrit ainsi cette technique : « Une fois que je l'ai [l'idée maîtresse], je l'écris sur un carton et la pose sur un petit chevalet de façon à l'avoir toujours sous les yeux. Dès lors je n'écrirai pas un alinéa sans le confronter avec elle et voir s'il s'y relie bien. »

Le constructeur d'une application à base d'ontologies qui veut réaliser une application *utile* et *utilisable* devrait se donner une contrainte du même genre que notre écrivain : celle non pas simplement de prendre en considération les usage(r)s de son application, ou de les avoir en vue, mais surtout de *ne pas les perdre de vue*, de garder le contact avec eux. L'utilisateur, sa tâche et son environnement devraient en effet rester le fil conducteur – qu'on appellera par la suite l'*utilisateur* ou l'*usager*.

¹ Benda, J. (1938). *Un régulier dans le siècle*, Paris, Gallimard ; cité par Etiemble et J. Etiemble (1970). *L'Art d'écrire*, Paris, Seghers, p. 415.

directeur – tout au long du processus de construction de l’application (cf. Giboin, Gandon, Corby, & Dieng, 2002). Pour le constructeur d’applications à base d’ontologies, la situation est cependant plus complexe : *a)* une application est un objet plus compliqué à construire qu’un essai ; *b)* l’écrivain est seul devant son essai ; le développeur n’est pas seul devant l’application : d’autres acteurs participent à sa construction ; en outre, plusieurs de ces co-constructeurs peuvent avoir en charge de ne pas perdre de vue l’utilisateur et donc de se référer régulièrement à l’utilisateur directeur et de rappeler cet utilisateur directeur aux autres participants ; *c)* pour l’écrivain, le carton est l’unique représentation du fil directeur ; pour les co-constructeurs, plusieurs représentations de l’utilisateur directeur peuvent co-exister (scénarios, requêtes, représentations usuelles, etc.) ; les constructeurs doivent expliciter ces représentations, les communiquer, les discuter ; de plus, il n’existe pas qu’un seul utilisateur à ne pas perdre de vue. Conséquence de cette complexité : si les occasions sont plus grandes de prendre contact avec l’utilisateur, elles sont aussi plus grandes de perdre ce contact, de perdre l’utilisateur de vue : parce qu’une représentation n’est pas communiquée, parce qu’elle est mal comprise ou parce qu’elle n’est pas acceptée.

La question devient alors cruciale : comment, dans une situation complexe comme la construction d’une application à base d’ontologies, ne pas perdre de vue les usagers et leurs usages ? Pour reprendre notre exemple du départ : quel serait, pour le constructeur d’application, l’équivalent de la technique du carton sur le chevalet pour l’écrivain ? quel serait le carton ? le chevalet ? qui serait l’écrivain ? quel serait l’essai à rédiger ?

Cet article fournit quelques éléments de réponse à ces questions. Ces éléments reposent sur notre expérience du projet pluridisciplinaire RNRT KmP de serveur Web sémantique de compétences inter firmes. L’article est organisé de la manière suivante : on présente d’abord l’application KmP ; on décrit ensuite le cadre et la procédure que nous avons utilisés pour analyser quand, comment et pourquoi, dans le projet KmP, nous avons pris contact avec les usage(r)s, perdu ce contact et, le cas échéant, repris ce contact ; puis on présente quelques cas analysés ; nous terminons sur la suite que nous comptons donner à ce travail d’analyse².

2 L’application KmP

KmP³ (*Knowledge Management Platform*) est un serveur Web sémantique permettant à des entreprises et des organismes de recherche de la Telecom Valley

² On notera que cet article a été rédigé en écho à l’article, à notre avis fondateur, de N. Aussenac-Gilles, A. Condamines, et S. Sulzman (2002), sur la prise en considération des applications dans la constitution des produits terminologiques (ex. : les ontologies), article où les auteurs mettent en évidence certains facteurs « dev[ant] être pris en compte pour espérer construire des produits utilisables et utilisés».

³ On décrit ici KmP-CORESE, le prototype KmP basé sur le moteur de recherche sémantique CORESE (Corby, Dieng, Faron-Zucker, 2004), qui utilise les graphes conceptuels. On ne rend pas compte de KmP-SCARCE, basé sur le moteur de composition sémantique SCARCE (Garlatti, Iksal, & Tanguy, 2004)

(Sophia Antipolis) de cartographier et d'échanger leurs compétences, soutenus en cela par des institutionnels locaux.

KmP est le résultat du projet exploratoire RNRT KmP (janvier 2003 – mars 2005), un projet pluridisciplinaire ayant impliqué plusieurs équipes de recherche spécialisées en gestion, économie, psychologie ergonomique et informatique : Laboratoire Rodige (UNSA-CNRS), Laboratoire Latapses (UNSA-CNRS), Projet Acacia (INRIA Sophia Antipolis), GET (Telecom Paris et ENST Bretagne), Association Telecom Valley (Sophia Antipolis). Ce projet était soutenu par le Laboratoire des usages de Sophia Antipolis. Faute de place, nous présenterons à grands traits KmP⁴.

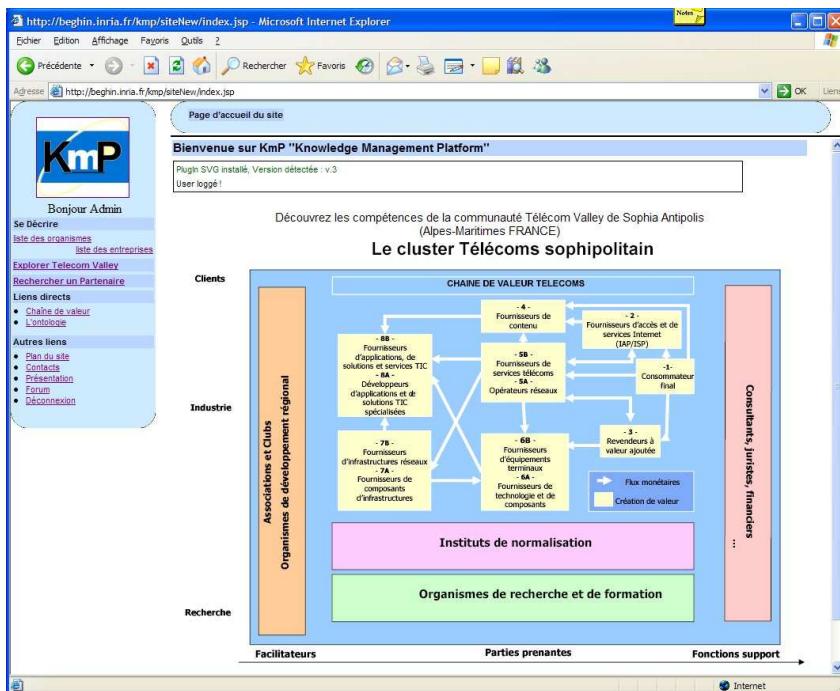


Fig. 1 – Page d'accueil de l'application KmP montrant la chaîne de valeur

Scénarios et requêtes.— KmP a été construit autour de trois grandes catégories de scénarios d'usage, dont on indique ici les buts : *Scénarios 1* : avoir et donner une visibilité générale de Telecom Valley ; *Scénarios 2* : favoriser les coopérations entre entreprises ; *Scénarios 3* : favoriser les coopérations entre recherche publique et recherche privée. KmP permet d'obtenir des réponses à des requêtes en rapport avec ces scénarios, par exemple : *Quelles sont les entreprises de la Telecom Valley dont la compétence est la conception de logiciels embarqués ? Quelles sont les nouvelles compétences apparues dans la Telecom Valley depuis ces trois derniers mois ?*

⁴ Pour obtenir plus d'informations sur KmP et accéder au serveur KmP, on peut consulter le site du projet : <http://www-sop.inria.fr/acacia/soft/kmp.html>.

Quels sont les laboratoires de recherche de la Telecom Valley spécialisés dans les systèmes informatiques pour le Knowledge Management et qui ont déjà monté des partenariats de R&D avec des entreprises ?

Ontologies et modèles.— Dans KmP, la description et la recherche de compétences sont facilitées par des ontologies (ex. : ontologie des ressources technologiques) et des modèles (ex. : modèle de la chaîne de valeur). Parmi les concepts principaux que recouvrent ces ontologies et modèles, citons : la *compétence* – qui se définit par une *action*, un *délivrable*, un *système d'offre* et un *bénéficiaire* –, la *ressource* utilisée pour mettre en œuvre la compétence (cette ressource peut être technologique, scientifique ou managériale), la *chaîne de valeur*, qui décrit les *acteurs* (entreprises) ou « segments de valeur » participant à la création d'une valeur économique, la *relation de partenariat*, le *cluster* et le *pôle* – le premier regroupant des acteurs aux compétences *complémentaires* (qui relèvent du même système d'offre) et le second des acteurs aux compétences *similaires* (qui réalisent le même type d'action).

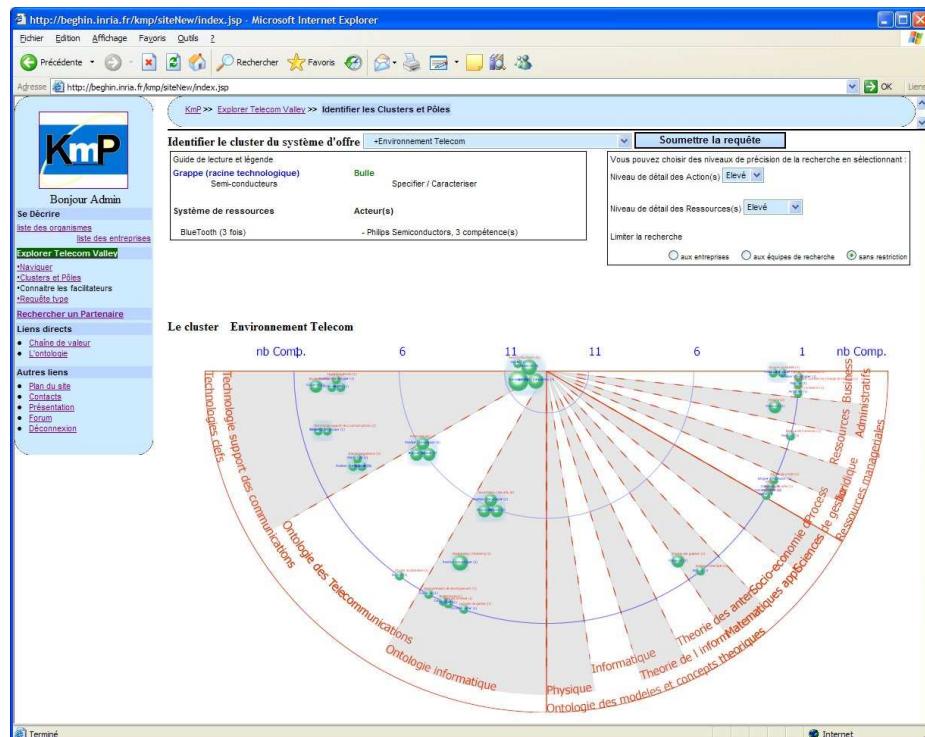


Fig. 2 – Cluster (dynamiquement composé) des compétences relevant du système d'offres des Télécommunications tel qu'on le trouve dans la Telecom Valley.

Fonctionnalités et interfaces.— KmP est construit autour de deux grands types de fonctionnalités découlant des scénarios et reposant sur les ontologies et les modèles : 1) fonctionnalités d'édition des compétences d'entreprises ou

d'organismes de recherche et 2) fonctionnalités de recherche et de visualisation des compétences (ex. : naviguer dans la chaîne de valeur ; rechercher un partenaire). Les interfaces proposées dans KmP sont des interfaces d'application à base d'ontologies et non des interfaces de gestion d'ontologies. Autrement dit, ces interfaces ont été autant que possible construites en rapport avec la tâche et les représentations de l'usager directeur plutôt qu'en rapport avec la tâche et les représentations de l'ontologue (voir par exemple les figures 1 et 2).

Technologies sous-jacentes.— KmP combine les technologies du Web sémantique (RDF, RDFS), du Web structuré (XML, XSLT) et du Web classique (HTML, CSS, SVG) pour intégrer les données issues de sources différentes, répondre à des requêtes formulées de différents points de vue, adapter le contenu aux utilisateurs, analyser, regrouper et fournir des indicateurs sur la Telecom Valley. KmP repose sur l'intégration de multiples composants : bases de données pour la persistance en *back-end*, serveurs Webs en JSP et servlets pour les *front-ends*, et serveur Web sémantique CORESE pour les capacités de traitement sémantique.

3 Cadre d'analyse des contacts et ruptures avec les usage(r)s

Pour analyser et caractériser la manière dont nous avons (re)pris et perdu contact avec les usager(r)s durant la conception du prototype KmP, nous avons ébauché un cadre d'analyse tournant autour de la notion de *représentation*.

Processus de gestion des représentations d'usage(r)s.— La (re)prise de contact ou la perte de contact surviennent à différents moments d'un processus de gestion des représentations inclus dans l'activité de construction de l'application à base d'ontologies. Nous envisageons la construction d'une application à base d'ontologies comme une activité de « conception collective centrée utilisateurs » (cf. Visser, Darses et Détienne, 2004) et, plus précisément, de « conception participative » (cf. Caelen & Jambon, 2004), ce que l'on a appelé « co-conception » (cf. Thomas, Giboin, Garlatti, et l'équipe KmP, 2003 ; Rouby & Thomas, 2004). On admet que cette activité est sous-tendue par un processus d'argumentation dans lequel les constructeurs doivent convaincre les usagers d'utiliser l'outil qu'ils construisent, et les usagers doivent convaincre les concepteurs d'intégrer leur point de vue dans l'outil à construire. Ce processus d'argumentation met en œuvre des *représentations d'usage(r)s* – c'est-à-dire des représentations externes reflétant une certaine vue des usage(r)s –, dont certaines deviendront des représentations communes de référence⁵. Lors de ce processus les interlocuteurs cherchent à tour de rôle à : *a*) donner accès ou avoir accès aux représentations d'usager(s), *b*) les faire comprendre ou les comprendre, *c*) les faire accepter ou les accepter, et *d*) en faciliter l'usage ou les

⁵ Il existe une vaste littérature sur ces représentations communes, que l'on dénomme de différentes façons, par exemple : référentiel commun, représentation intermédiaire, mécanisme de coordination, objet frontière, genre communicationnel, etc. (pour un état de l'art sur les référentiels communs et les notions connexes étudiées du point de vue de la psychologie ergonomique, voir Giboin, 2004). On notera qu'une distinction est faite dans cette littérature entre représentations communes externes (ou physiques) et représentations communes internes (ou mentales ou ressources cognitives). On ne considère ici que les représentations externes.

utiliser. Les représentations acquièrent alors le statut de représentations accessibles, comprises, acceptées ou utilisées⁶. Lors des différentes étapes de la gestion des représentations, les interlocuteurs mettent en œuvre différentes opérations, telles que : expliciter la représentation, la vérifier, la discuter, l'adapter, surveiller son utilisation, etc. Ces différentes opérations vont déterminer en partie les *rôles représentationnels* que vont jouer les personnes participant à la conception de l'application.

Rôles représentationnels.— La (re)prise ou la perte de contact ont lieu entre *rôles représentationnels*. On distingue deux grands types de rôles représentationnels : 1) des rôles de représentants – les personnes qui représentent les usage(r)s, représentants *proches* (spécialistes des usages) et représentants *éloignés* (informaticiens) – et 2) des rôles de représentés – les usagers représentés. Le principal rôle de représenté est l'*Usager directeur*, c'est-à-dire l'utilisateur à ne pas perdre de vue (à la limite, chaque utilisateur singulier ne devrait pas être perdu de vue). L'usager directeur est dans notre cas un utilisateur pilote de KmP. Plusieurs rôles de représentants peuvent être identifiés, parmi lesquels ceux de : – *Preneur de vue* (celui qui va prendre la vue de l'utilisateur, autrement dit, l'auteur de la représentation ; une catégorie particulière d'auteur est le *Contributeur de vue*) ; – *Porteur de vue* (celui qui prend en charge ou défend la vue ou les représentations) ; – *Traducteur de vue* (celui qui va expliquer la vue, ou la représentation, la traduire de manière compréhensible par un interlocuteur donné) ; – *Vérificateur de vue* (celui qui vérifie que la vue prise de l'utilisateur est pertinente) ; – *Veilleur de vue* (celui qui surveille que la vue est bien prise en considération et qui suit les évolutions des usage(r)s et répercute celles-ci sur la vue et les représentations).

Représentations d'usage(r)s.— La (re)prise et la perte de contact portent sur des *représentations d'usage(r)s*, qui peuvent être envisagées en fonction de leurs dimensions⁷ : – *Type* (le type de représentation exprimant la vue sur l'utilisateur : artefacts à construire [application, ontologie, interface...], objets intermédiaires [scénario, *story-board*, requête...]) ; – *Objectif** (l'objectif pour lequel la représentation est utilisée, ou le produit attendu qui reflètera la vue : ontologie, application à base d'ontologies, etc.) – *Contenu** (l'information ou la connaissance exprimée dans la représentation : compétence, chaîne de valeur, etc.) – *Forme** (la forme sous laquelle est exprimée la représentation : abstraite ou concrète ; formelle ou informelle ; générique ou spécifique ; etc.) ; – *Cycle de vie** (l'étape du processus de construction où intervient la représentation : analyse des besoins, spécifications fonctionnelles, etc.) ; – *Cycle d'argumentation* (l'étape atteinte dans le processus d'argumentation : représentation accessible, représentation comprise, représentation acceptée, représentation utilisée) ; – *Valeur* (l'importance que chaque rôle accorde à la vue) ; – *Provenance* (les sources d'où provient la représentation : un document, une description orale, etc.) ; – *Acteurs intéressés* (les acteurs ayant un intérêt dans la représentation ; ces acteurs déterminent en partie la valeur de la représentation).

⁶ Cette description reprend, en les adaptant, des éléments du modèle d'argumentation du logicien J.-B. Grize (1990).

⁷ Les dimensions marquées d'un astérisque (*) sont inspirées des dimensions des scénarios définies dans la méthode CREWS de conception à base de scénarios (cf. Rolland et al., 1998)

Application du cadre aux documents de conception de KmP.— Les composants que l'on vient de décrire ont été utilisés de manière combinée pour caractériser les points de contact ou de perte de contact avec les usage(r)s dans le projet KmP. Une perte de contact pourra ainsi s'expliquer comme le rejet d'une représentation ayant une grande valeur pour un usager directeur. Une reprise de contact comme la traduction d'une représentation formelle en une représentation non formelle.

Le matériel sur lequel a porté l'analyse sont divers documents de conception de l'application KmP, composés ou récupérés par différents membres de l'équipe de conception du projet KmP : transcriptions d'entretiens avec les usagers, comptes rendus de réunion, documents de spécification, copies écran des versions successives de l'application KmP, rapports d'évaluation de l'application KmP, documents d'entreprise, etc. L'expérience KmP étant très riche d'enseignements, nous nous limiterons à quelques cas de (re)prises et de pertes de contact avec les usage(r)s.

4 Quelques cas de (re)prises et pertes de contact

Prises de contact par l'implication d'utilisateurs (représentés).— Un moyen bien connu de prendre contact avec les usagers est d'impliquer ces derniers directement dans la conception. C'est ce que nous avons fait en adoptant une méthode de conception participative, ou co-conception, impliquant des utilisateurs (représentés), appelés « utilisateurs pilotes de KmP » : membres d'entreprises et d'organismes de recherche de la Telecom Valley, institutionnels locaux. Ces utilisateurs étaient regroupés dans un comité de pilotage. Certains de ces utilisateurs ont participé directement à la construction des ontologies des ressources technologiques. Ils ont même sollicité cette participation directe dans le but de s'approprier ces objets au départ mystérieux pour eux qu'étaient les ontologies.

Prises de contact par l'implication de représentants proches de l'utilisateur.— Un moyen complémentaire, et connu lui aussi, de prendre contact avec les usagers est d'impliquer des spécialistes des usages, qui sont des représentants proches des utilisateurs. Des spécialistes de sciences de la gestion, de sciences économiques et de psychologie/ergonomie ont participé à la conception du prototype KmP. La participation de gestionnaires et d'économistes a d'ailleurs permis de suivre la recommandation de Caelen & Jambon (2004), d'« étendre [la conception participative] aux dimensions socio-économiques de l'usage – en intégrant les sociologues, anthropologues et économistes dans le processus [de conception] ». Les gestionnaires et les économistes étaient les représentants les plus nombreux : 13 au total, dont 7 sont intervenus tout le long du projet et 6 autres ponctuellement. Par leur maîtrise des concepts organisationnels reflétant des usage(r)s (chaîne de valeur, compétences d'entreprise, etc.), ces spécialistes ont joué un rôle prépondérant dans la construction de l'application, comme *preneurs de vue*, et comme *porteurs de vue*.

Les psychologues/ergonomes, moins nombreux (1+2)⁸, ont surtout joué le rôle de traducteurs de vue et de vérificateurs de vue.

Ces représentants proches ont permis aux représentants éloignés – les informaticiens – de réaliser le prototype en ayant l’assurance de ne pas être totalement déconnectés des usager(s). La connexion était facilitée par le fait que l’un des informaticiens avait une formation initiale en psychologie cognitive, et l’un des psychologues/ergonomes avait une formation initiale en informatique. Les informaticiens étaient au nombre de 12 (5 + 7).

Prise et maintien de contact par les objets intermédiaires.— Une manière significative de prendre contact avec les usage(r)s et de garder ce contact est de communiquer par scénarios d’usage. Dans KmP, nous avons repris la méthode des scénarios que nous avions utilisée dans le projet européen IST CoMMA (cf. Gandon, 2002 ; Giboin *et al.*, 2002). Cette méthode a été adaptée à la nouvelle application et à l’entrée de nouveaux analystes des usages : gestionnaires et économistes (cf. Pascal & Rouby, 2004). L’ouverture à ces analystes a permis d’obtenir une vue plus complète des usage(r)s, une vue qui reflète davantage les aspects organisationnels, fondamentaux dans KmP. Les scénarios de KmP ayant été élaborés à partir d’entretiens avec des utilisateurs réels décrivant des besoins réels, cela a conforté la prise de contact, mais aussi son maintien. Ces représentations avaient en effet une valeur forte pour les usagers directeurs.

La prise de contact est passée aussi par les requêtes et réponses liées aux scénarios d’usage. Requêtes et réponses contiennent en effet les termes et les relations entre ces termes qui doivent être retenus pour construire une ontologie utile et utilisable. On est d’autant plus proche des usage(r)s que l’on recueille les requêtes et les réponses réelles des usagers directeurs. Nous avons essayé le plus possible de récupérer des documents d’usagers où trouver ces questions et ces réponses.

La prise de contact s’est faite également par les maquettes d’interfaces ou *story-boards* élaborés dans un premier temps par les *représentants proches* à partir des scénarios d’usage. Pour établir un contact plus étroit, il a fallu dans un deuxième temps demander aux usagers eux-mêmes de participer directement à l’élaboration des maquettes.

Perte de contact avec les destinataires des représentations.— Une perte de contact a eu lieu avec certains destinataires des représentations. C’est ainsi que les utilisateurs pilotes membres de SSII ayant participé au groupe de travail sur l’ontologie des ressources informatiques ont demandé que viennent aussi à ce groupe des donneurs d’ordre. Ces derniers leur auraient fourni des exemples de requêtes auxquelles ils devaient répondre en tant que SSII ; les réponses fournies devaient inclure des concepts utilisés dans les requêtes. Pour les membres des SSII, l’ontologie qu’ils étaient en train de construire devait aussi servir aux donneurs d’ordre. Il n’y a pas eu de séances de travail réunissant à la fois des SSII et des donneurs d’ordre. Plus généralement, nous aurions été encore plus en contact avec

⁸ On notera que quelques mois avant le commencement du projet RNRT, trois étudiants d’un DESS d’Ingénierie des Ressources humaines (dont les responsables sont des enseignants-rechercheurs en psychologie sociale) ont contribué à l’élaboration d’une version préliminaire des ontologies de KmP à partir d’entretiens menés auprès d’utilisateurs pilotes.

les usage(r)s si, de manière systématique, nous avions placé des usagers en situation réelle de conseil (via le courriel par exemple) et recueilli les requêtes et les réponses qui ont été échangées dans cette situation. Nous n'avons recueilli que quelques exemples d'échanges de ce type.

Perte et reprise de contact sur l'acceptabilité des représentations.— Ce cas a été observé par exemple à propos de la chaîne de valeur, qui n'était pas familière aux usagers. Avant de montrer cette représentation aux usagers, et pour identifier leurs représentations équivalentes du moment, on a proposé une procédure de vérification de la pertinence de la chaîne de valeur, qui consistait en trois étapes : 1) sans se reporter à la chaîne de valeur, ni à un quelconque document d'entreprise, décrire en quelques lignes le rôle de son entreprise ; 2) prendre un document d'entreprise de son choix (plaquette de présentation, page web, etc.) et reporter la définition du rôle de son entreprise figurant sur ce document ; 3) en se reportant à la chaîne de valeur, indiquer le segment auquel son entreprise peut être identifiée ; préciser si les intitulés des segments conviennent ou non ; dans la négative, indiquer les modifications que l'on souhaiterait apporter aux intitulés. Grâce à cette procédure, on a pu constater que, pour accepter la représentation de la chaîne de valeur, certains utilisateurs souhaitaient qu'apparaisse dans les intitulés de certains segments un terme crucial pour eux, par exemple le terme « Solutions » pour le représentant d'une entreprise donneur d'ordre.

Perte de contact avec la valeur d'une représentation.— Les pertes de contact apparaissent aussi lorsqu'on oublie que le produit à construire ayant le plus de *valeur* pour l'usager directeur, c'est l'application à base d'ontologies plus que l'ontologie elle-même : l'ontologie n'est qu'un moyen de rendre l'application plus pertinente. Or ce type de perte de contact est apparu à plusieurs reprises dans le projet. Il a semblé se produire en effet une sorte de fixation sur l'ontologie, ou d'*obnubilation ontologique*, qui a conduit les constructeurs, y compris les représentants proches, à prendre parfois l'ontologie comme la finalité de leur tâche. De sorte que l'on en est venu à représenter telle quelle dans l'interface la représentation interne (à la machine et/ou à l'ontologue) de l'ontologie. On voit d'ailleurs encore sur l'interface de KmP des traces de cette fixation ontologique. Par exemple, sur le radar de l'interface des clusters (voir figure 2), le terme « ontologie » apparaît alors qu'il ne devrait pas, car l'utilisateur ne manipule pas des ontologies dans sa tâche, mais des ressources, etc. Il faudrait donc penser un peu moins à l'*ontologie*, et un peu plus à la *téléologie*, c'est-à-dire aux objectifs et aux tâches de l'usager.

Perte et reprise de contact avec les formes compréhensibles des représentations.— Le décrochage avec les usage(r)s intervient lorsqu'on privilégie l'ontologie dans son état formel par rapport à l'ontologie dans son état non formel – l'ontologie en langue naturelle par exemple, ou représentée sous forme d'images ou de photographies, comme dans SemTalk (Fillies & Sure, 2002) – ou lorsqu'on privilégie l'ontologie interne par rapport à l'ontologie telle qu'elle s'incarne dans l'interface. On reprend contact lorsqu'on ancre l'ontologie dans les formes extérieures où elle intervient : requêtes, documents, graphiques, etc. Toutes ces formes sont celles manipulées par les utilisateurs. Si l'on accepte ce point de vue, on devrait alors accepter l'idée de compléter les différentes formes d'engagement

(sémantique, ontologique, computationnel) décrites par Bachimont (2000) par un *engagement d'usage* ou *engagement d'interface*, qui porteraient sur la représentation externe ou d'usage des ontologies.

Cas similaire : une interface avait été élaborée, sans interaction avec les usagers ou les représentants proches, pour indiquer le niveau de détail dans la description des clusters. Ces niveaux de détail étaient exprimés en termes de niveau de profondeur dans l'arbre ontologique. Cette représentation était illisible pour les usagers, car trop proche des mécanismes internes de calcul. Le contact a été repris quand il a été demandé aux représentants proches de définir ce qu'ils entendaient par niveau de détail, et comment le représenter dans l'interface. D'où une représentation plus lisible du niveau de détail (voir figure 2, zone supérieure droite).

Perte et reprise de contact avec les sources pertinentes des représentations.—

On a constaté une perte de contact lorsqu'on a cherché à construire l'ontologie des technologies à partir d'ontologies ou de taxinomies existantes : les usagers ne se retrouvaient pas complètement dans ces ontologies. Ils ont pu reprendre contact lorsqu'ils ont pu rapporter le contenu de ces ontologies à leur métier. Plus généralement on pourrait se demander si l'introduction de « modèles étrangers », de normes, d'ontologies existantes, etc., au nom de la réutilisabilité n'est pas un risque de décrochage supplémentaire. Les ressources réutilisables ne seraient ainsi utilisables que si les utilisateurs se les sont auparavant appropriés. Ce qui impliquerait que ce soit les usagers eux-mêmes qui sélectionnent les ressources à utiliser. C'est ce qu'ont fait par exemple les utilisateurs pilotes membres de SSII, qui ont participé au groupe ayant construit l'ontologie des technologies informatiques.

Perte de contact par confusion sur la valeur des représentations. Reprise de contact par adaptation des représentations .— Il est arrivé que les représentants proches s'éloignent des utilisateurs en confondant la valeur qu'ils attribuaient aux représentations qu'ils proposaient aux usagers (par exemple la chaîne de valeur) à la valeur que les usagers eux-mêmes leur attribuaient. Pour garder le contact avec les usagers, les représentants doivent ainsi accepter de voir modifiées ou remplacées les représentations qu'ils proposent.

Pertes volontaires de contact et prévision de reprises de contact.— La construction d'une application à base d'ontologies étant une tâche complexe, on ne peut envisager de pouvoir garder le contact sur tout. Il arrive donc que l'on décide volontairement de perdre le contact, mais en prévoyant de reprendre ce contact à un moment plus propice. Ainsi avons-nous anticipé quelques cas de modifications des représentations proposées par les représentants directs.⁹ Premier cas : le besoin de restructurer la chaîne de valeur. On a alors proposé une fonctionnalité permettant à l'utilisateur de déplacer les segments de valeur de façon à rapprocher ceux qu'il souhaiterait voir ensemble. Deuxième cas : la création possible d'autres chaînes que la chaîne de valeur. Lors d'un entretien avec le représentant d'un donneur d'ordre qui nous parlait des projets auxquels son entreprise pouvait participer, l'idée nous est

⁹ Il s'agit aussi d'accepter et de gérer le découplage entre les représentations internes et les représentations externes, c'est-à-dire des représentations externes directement calquées sur les représentations internes sont utiles au développement et la maintenance, mais il faut accepter la divergence lorsqu'il s'agit de représentations d'usage(r)s.

venue de construire avec cet utilisateur pilote une représentation que nous avons appelée la *chaîne de projets*. Concrètement, nous avons proposé à l'utilisateur pilote de répartir ces projets sur un axe recherche/application, et nous lui avons demandé de situer des entreprises ou des organismes de recherche qu'il connaissait le long de cet axe. Ces différentes représentations (et d'autres du même genre dont on n'a pas parlé ici) n'ont pas été utilisées dans KmP, mais ont été gardées en réserve afin de préparer des évolutions probables d'usage du prototype KmP.

5 Conclusion

Nous venons de décrire quelques cas de (re)prises et pertes de contact avec les usager(s) au cours du projet KmP. Ces cas fournissent des indications sur la technique à mettre au point pour ne pas perdre de vue l'utilisateur au cours d'un projet de conception d'application à base d'ontologies. On a pu voir que cette technique devrait être beaucoup plus complexe que la technique du carton sur le chevalet utilisée par l'écrivain solitaire, en particulier parce qu'elle implique plusieurs acteurs qui doivent coordonner différentes représentations d'usage(r)s.

Un prochain objectif est de spécifier cette technique, et donc d'aller plus avant dans l'élaboration de notre cadre d'analyse et de nos analyses des situations de maintien du contact avec les usage(r)s. L'analyse que nous avons réalisée a porté sur certains documents de conception de KmP (la mémoire externe du projet). Nous comptons approfondir cette analyse sur d'autres documents. Nous comptons aussi mener une analyse fondée sur des entretiens post-projet avec les protagonistes du projet KmP (la mémoire interne du projet).

Si, dans le retour d'expérience que nous venons de rapporter, nous avons voulu faire la part des aspects négatifs (pertes de contact) aussi bien que des aspects positifs (prises et reprises de contact avec les usage(r)s du projet KmP), il nous faut souligner que le bilan final du projet est avec certitude positif. Dans un rapport daté de mars 2004, le Conseil scientifique du Laboratoire des usages de Sophia Antipolis soulignait déjà que KmP était le plus innovant, le plus complet et le plus réussi des dix projets se réclamant de ce Laboratoire des usages. Le projet KmP a beaucoup évolué depuis, et l'intérêt porté par ses utilisateurs potentiels s'est accru. Plusieurs suites sont d'ores et déjà prévues au projet KmP, en particulier un projet KmP-2, dont l'objectif est d'étendre la cartographie des compétences inter firmes du site de Sophia Antipolis à la région Provence-Alpes-Côte d'Azur.

Références

- AUSSENAC-GILLES, N., CONDAMINES, A., & SZULMAN, S. (2002). Prise en compte de l'application dans la constitution de produits terminologiques. In J. Le Maître (Ed.), *Information, Interaction, Intelligence : Actes des 2e Assises Nationales du GDR I3*, Cépaduès Editions, pp. 289-303, décembre 2002.

- BACHIMONT, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances. In J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (Eds.), *Ingénierie des connaissances, évolutions récentes et nouveaux défis*. Paris: Eyrolles.
- CAELEN, J. JAMBON, F. (2004). Conception participative par « moments ». *Actes de la 16^{ème} conférence francophone sur l'interaction homme-machine (IHM'04)*, ACM International Conference Proceedings Series ISBN : 1-58113-926-8, p. 29-36.
- CORBY, O., DIENG-KUNTZ, R., FARON-ZUCKER, C. (2004). Querying the Semantic Web with the CORESE search engine. In R. Lopez de Mantaras and L. Saitta (Eds), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI2004), subconference PAIS'2004, Valencia, 22-27 August 2004, IOS Press, p. 705-709.
- FILLIES, C., SURE, Y. (2002). On Visualizing the Semantic Web in MS Office. In *Proceedings of the 6th International Conference on Information Visualisation (IV02)*, 10-12 July 2002, London, England, pp. 441-446.
- GANDON, F. (2002). *Ontology Engineering : a survey and a return on experience..* Rapport de Recherche RR-4396, INRIA, Sophia Antipolis, Mars, 2002.
- GARLATTI, S., IKSAL, S., TANGUY, P. (2004). SCARCE: an Adaptive Hypermedia Environment Based on Virtual Documents and Semantic Web. In S.Y. Chen and G.D. Magoulas (Eds), *Adaptable and Adaptive Hypermedia Systems*, Idea Group Inc., pp. 206-224.
- GIBOIN, A. (2004). La construction de référentiels communs dans le travail coopératif. In J.M. Hoc et F. Darses (Eds.). *Psychologie ergonomique: tendances actuelles* (pp.119-139). Paris, PUF.
- GIBOIN, A., GANDON, F., CORBY, O., DIENG, R. (2002). *User Assessment of Ontology-based Tools: A Step Towards Systemizing the Scenario Approach.* in « Proceedings of EON'2002 : Evaluation of Ontology-basedTools, OntoWeb-SIG3 Workshop at the 13th International Conference on Knowledge Engineering and Knowledge Management EKAW 2002 », pages 66-73, Sigüenza, Spain, September 30, 2002, 2002.
- GRIZE, J-B., 1990: *Logique et langage*. Ophrys, Paris.
- PASCAL A., ROUBY, E. (2004). Une méthode de construction des scénarios d'usage à la croisée des théories de la cognition distribuée et de la structuration. *Journée de recherche sur « Les approches cognitives en sciences de gestion : transversalité des objets et méthodes innovantes »*, Evry, 30 septembre.
- ROLLAND, C., BEN ACHOUR, C., CAUVET, C., RALYTÉ, J., SUTCLIFFE,A., MAIDEN, N.A.M., JARKE, M.. HAUMER, P., POHL, K., DUBOIS, E., HEYMANS, P. (1998). A Proposal for a Scenario Classification Framework. *Requirements Engineering Journal*, Vol. 3, No. 1, Springer Verlag, pp.23-47.
- ROUBY, E., THOMAS, C. (2004). La codification des compétences organisationnelles : l'épreuve des faits. *Revue française de Gestion*, 149, pp.51-68.
- VISSEUR, W., DARSE, F., & DÉTIENNE, F. (2004) Approches théoriques des activités de conception en psychologie ergonomique. In J.M. Hoc et F. Darses (Eds.) *Psychologie Ergonomique : tendances actuelles* (pp. 97-118). Paris : PUF.

Modèles Sémantiques de Composants pour l'Ingénierie des Systèmes d'Information

Gwladys Guzélian, Corine Cauvet

LSIS

Université Paul Cézanne, Aix-Marseille III
Avenue Escadrille Normandie Niemen
13397 Marseille cedex 20
gwaldys.guzelian@lsis.org
corine.cauvet@univ.u-3mrs.fr

Résumé. L'ingénierie à base de composants s'impose peu à peu dans le développement des systèmes d'information (S.I.). Pourtant, elle n'a pas encore atteint son niveau de maturité. Le terme même de composant est souvent défini de façon imprécise et parfois contradictoire, les modèles de composants proposés peuvent avoir des finalités et des contextes d'utilisation très différents. Dans ce papier, nous nous intéressons aux modèles sémantiques de composants, c'est-à-dire des modèles visant à associer aux composants une connaissance qui guide leur usage. Nous étudions trois types de modèle : les approches « Web Services Sémantiques », les approches « patrons » et les approches de modélisation de domaine. L'étude comparative de ces approches permet de dégager des différences importantes en terme de finalité, de spécification et d'aide à l'utilisation des composants.

Mots-clés : Composants sémantiques, systèmes d'information, Web Services, Patrons, Modélisation de domaine.

1 Introduction

L'ingénierie à base de composants s'impose peu à peu dans le développement des systèmes d'information (S.I.). Le rôle croissant et diversifié que jouent le Web, l'Internet et l'Intranet dans la conception et la mise en ligne d'applications va amplifier ce phénomène.

Pourtant, l'ingénierie des systèmes d'information à base de composants n'a pas encore atteint son niveau de maturité. Sur le plan méthodologique, il n'existe pas d'approche de développement complètement orientée composant. Le terme même de composant est souvent défini de façon imprécise et parfois contradictoire, les modèles de composants proposés peuvent avoir des finalités et des contextes d'utilisation très différents (Fellner & Turowski, 2000), (Heinemann & Councill, 2001). Malgré cette diversité, la tendance qui s'impose est de considérer un composant comme une brique logicielle permettant d'organiser et de réutiliser du logiciel. En conséquence, les

composants sont des artefacts logiciels, leur spécification est relativement technique et leur usage reste limité à la phase de conception d'architecture logicielle et de production logicielle.

La possibilité d'utiliser une approche orientée « composant sémantique » présente pourtant un réel intérêt. La description du problème auquel le composant répond, du contexte dans lequel il peut être utilisé et des lois de composition qui permettent de l'assembler à d'autres composants est essentielle pour mettre en œuvre une approche systématique de développement à base de composants.

Dans ce papier, nous étudions trois types de modèles de composants qui relèvent de l'approche « composant sémantique ».

- L'approche « **Web Services Sémantiques** »,
- L'approche « **Patrons** »,
- L'approche « **Modèles de domaine** ».

Il existe plusieurs études comparatives des approches orientées composants (Fellner & Turowski, 2000), (Heineman & Councill, 2001). Ces études ont essentiellement porté sur les composants technologiques. L'émergence de l'approche « **Web Services Sémantiques** » conduit à une nouvelle forme de composant qui vise à intégrer dans chaque composant une description sémantique pour en faciliter l'usage (la recherche, l'adaptation et la composition). Nous pensons que cette approche présente des ressemblances avec les approches « patrons » en génie logiciel et les approches « modèles de domaine » en ingénierie des connaissances. Dans ce papier, nous comparons ces différentes approches sur un même ensemble de critères. Les critères ont été choisis pour permettre une comparaison des modèles conceptuels sous-jacents aux différentes approches.

Le papier est organisé en 3 sections. Dans la **partie 2**, nous présentons les principes de spécification d'un modèle sémantique de composant. Dans la **partie 3**, nous analysons chacune de ces approches en fonction des principes présentés. Dans la **partie 4**, nous effectuons une analyse comparative de ces approches.

2 Principes de spécification des composants sémantiques

Il existe de nombreuses définitions de la notion de composant. Toutes ces définitions font apparaître une grande diversité des modèles proposés pour décrire les composants. Nous rappelons qu'un modèle sémantique de composants vise à associer aux composants une connaissance qui guide leur usage. Nous caractérisons un modèle sémantique au moyen de six principes : l'abstraction, la variabilité, la contextualisation, la composition, l'interopérabilité et la localisation. Ces six principes ont été choisis pour, d'une part, caractériser les approches à un niveau conceptuel et d'autre part mettre en évidence à la fois leurs points communs et leurs différences.

2.1 Principe d'abstraction

Par analogie avec l'approche objet, un composant comporte une «interface» visible par les clients susceptibles de l'utiliser et une «implémentation» qui contient la connaissance effectivement réutilisable fournie par le composant. Appliqué aux composants sémantiques, le principe d'abstraction vise à mettre en avant le besoin auquel un composant répond plutôt que les détails de la solution qu'il fournit.

Nous pensons qu'un composant doit posséder une interface centrée sur un problème métier, la partie implantation devant fournir une solution technique. Aujourd'hui, un certain nombre d'auteurs proposent d'utiliser des modèles de buts («Business Goals») ou des modèles de tâches pour décrire ces problèmes métier (Eriksson & Penker, 2000), (Motta & Zdrahal, 1998).

2.2 Principe de variabilité

Le principe de variabilité permet d'associer à un même concept plusieurs points de vue. Par exemple, le concept de «*personne*» peut avoir différentes représentations dans le domaine de la gestion, il peut correspondre à un abonné que l'on gère dans une bibliothèque, à un client que l'on gère dans une banque ou encore à un assuré géré dans une compagnie d'assurance.

Dans les composants orientés problème, la variabilité est exprimée à travers les différentes solutions possibles d'un même problème. Au niveau technique le principe de variabilité est mis en œuvre par différents mécanismes (Gurp & Bosch, 2001) comme les paramètres, l'héritage, les générateurs d'applications, la programmation «orientée aspect», la technologie à base de «frames»... Ce principe est la clé de la réutilisation, il permet l'adaptation du composant dans différentes situations et donc sa réutilisation dans différents contextes.

2.3 Principe de contextualisation

Complémentaire au principe de variabilité, le principe de contextualisation permet de discriminer les différentes variantes d'un même concept. Ce principe est essentiel pour guider le choix d'un composant. Dans la spécification d'un composant, la connaissance contextuelle définit la situation dans laquelle le composant est particulièrement adapté. Ce principe prend toute son importance lorsqu'un composant a été spécifié selon une approche orientée problème. En effet un problème de conception peut être résolu de différentes manières, en fonction de besoins particuliers à satisfaire ou de contraintes spécifiques à respecter. Ces différentes solutions doivent être différencierées en explicitant leur contexte d'application pour guider le choix de la meilleure.

2.4 Principe de composition

La mise en œuvre d'une approche à base de composants nécessite l'assemblage des composants pour aboutir à une solution globale. Les règles d'assemblage des composants doivent être intégrées et spécifiées au sein des composants. Les règles d'assemblage peuvent être décrites à différents niveaux d'abstraction. Au niveau

logiciel, les langages de description d'architectures (ADL : « Architecture Description Languages ») utilisent des connecteurs pour modéliser les interactions entre les composants et les règles qui régissent ces interactions. A un niveau conceptuel, les connecteurs peuvent exprimer les lois de fonctionnement d'un domaine telle une relation entre un client et une ressource ou une relation contractuelle entre deux partenaires, un client et un fournisseur par exemple. Dans une approche composant, il est important de gérer les composants et les connecteurs comme des concepts de même niveau. Les composants comme les connecteurs doivent pouvoir être réutilisés et adaptés.

2.5 Principe d'interopérabilité

L'interopérabilité est essentielle pour permettre l'assemblage et la coopération de composants hétérogènes et souvent distribués.

Comme pour la composition et la variabilité, l'interopérabilité peut être traitée à différents niveaux. Au niveau technique, les modèles tels que CORBA, COM / DCOM ou EJB visent cet objectif d'interopérabilité au moyen de protocoles de communication. Au niveau conceptuel, l'interopérabilité est basée sur des modèles de référence ou des ontologies (Vargas Solar & Doucet, 2002) et des règles de correspondance assurant le passage entre le modèle de référence et les modèles spécifiques à chaque environnement.

2.6 Principe de localisation

Les composants sont des artefacts dont l'usage nécessite la mise en œuvre d'un processus de recherche. Le principe de localisation répond à un besoin de recherche de composants dans une ou plusieurs bases. Ce principe peut être mis en œuvre au sein d'une base de composants par un modèle d'organisation de type hiérarchique, réseau... qui guide la recherche. Une autre façon de satisfaire ce principe est de spécifier et d'implanter le processus de réutilisation pour supporter entièrement la recherche en tenant compte du besoin exprimé et de la situation courante. Le principe de localisation doit aujourd'hui être étendu pour prendre en compte des sources de composants distribuées.

Ces six principes permettent de faire la différence entre objet et composant. Ils permettent aussi de distinguer composant sémantique et composant logiciel. La prise en compte de la variabilité, de la connaissance contextuelle et de l'orientation problème dans la spécification des composants permet d'associer dans un composant une connaissance réutilisable et une connaissance qui guide sa réutilisation.

3 Modèles sémantiques de composants

Dans cette section, nous étudions trois types de modèles sémantiques de composants sur la base des principes définis dans la section précédente.

3.1 Les approches WSS (Web Services Sémantiques)

Dans ces approches, les composants sont considérés comme des services externes (disponibles sur le Web) pouvant coopérer pour réaliser une certaine fonctionnalité. Nous présentons d'abord les « Web Services » et ensuite leur extension avec les « Web Services Sémantiques ».

3.1.1 Les Web Services

D'une façon générale, les Web Services sont des fragments d'application de gestion identifiés par une URL. Ils correspondent à des composants logiciels réutilisables qui réalisent des tâches spécifiques en utilisant des mécanismes standards orientés Web (OWL-S, 2004).

Un service est décrit à l'aide du langage WSDL qui le spécifie en termes de messages et d'opérations qu'il propose. La description d'un service avec WSDL comprend deux parties indépendantes (**Fig. 1**) : l'**interface** qui définit les éléments caractérisant les informations communes d'une catégorie de services et la partie **implémentation** qui décrit comment un service d'une interface particulière est implanté.



La figure 1 (Srivastava & Koehler, 2003) illustre avec le méta langage WSDL un service de planification de voyage dans une agence. Chaque opération est associée à un message d'entrée (« input message ») et/ou de sortie (« output message »). De ce fait, l'opération « Replan Itinerary To Flight Service » associée au message de sortie « BookingFailure » signifie que l'exécution de l'opération «re-planification d'une réservation d'un vol d'avion » est effectuée lorsque la réservation est annulée. L'expression des Web Services avec WSDL décrit les messages de façon syntaxique sans préciser leur sémantique.

3.1.2 Les Web Services Sémantiques

Les « Web Services Sémantiques » visent à enrichir la représentation des Web Services avec des connaissances relatives à l'objectif du service, au processus mis en œuvre par le service et à ses contraintes d'utilisation. Ces connaissances peuvent alors être exploitées par des agents pour automatiser la recherche, l'invocation, la composition et l'exécution des Web Services. Les langages de description des « Web Services Sémantiques » assurent le marquage sémantique des Web Services, et rendent explicite la connaissance nécessaire à la sélection et à la composition (McIlraith et al., 2001).

Par exemple, l'approche OWL-S (OWL-S, 2004) est une extension de l'approche DAML-S qui propose une ontologie de services dans laquelle chaque service est décrit selon trois dimensions : le « **Profile** » du service qui précise les fonctionnalités du service, le « **Modèle** » du service qui décrit le processus mis en œuvre par le service, et le « **Grounding** » du service qui spécifie en termes de messages les règles d'interaction avec le service.

Cette ontologie de services enrichit WSDL et UDDI pour permettre un accès et une utilisation du service à partir de son contenu (objectif, messages....) plutôt que par des mots clefs.

3.1.3 Analyse des approches WSS

Tout d'abord, l'approche OWL-S vérifie le principe de localisation et de contextualisation. Par exemple, si l'on cherche un service permettant de commander un appareil photo numérique, on peut formuler des requêtes complexes comme « commander un appareil photo numérique de marque Sony et d'un poids inférieur à 500 gr ». On peut également affiner le contexte d'utilisation du service en précisant les notions de coût et de qualité de service tels les délais de livraison.

Ensuite, les principes de composition et d'interopérabilité sont étendus afin d'engendrer des services complexes combinant différents Web Services. Par exemple, la composition peut être utilisée pour faciliter la planification des vacances d'une personne qui souhaite ne pas dépenser plus de 1000 euros entre la réservation d'hôtel, le service de location de voiture, et les billets d'avion. Le plus souvent un service complexe est modélisé comme un processus (processus d'orchestration) faisant coopérer plusieurs activités. De plus, le principe d'interopérabilité est implicitement satisfait puisqu'il est possible de composer des services différents. Même s'ils sont implantés différemment, il faut par exemple, pouvoir comparer deux dates fournies par des services différents. L'utilisation d'une ontologie de services dans OWL-S permet ainsi de définir un langage commun à tous les services pour le partage et la réutilisation des données sur le Web.

D'autres approches de ce type, autres que OWL-S, possèdent des aspects sémantiques légèrement différents. Par exemple, dans l'approche IRS (Cabral et al., 2004), l'orientation problème est suggérée au moyen de modèles de tâches pour décrire sémantiquement les Web Services et les objectifs auxquels ils répondent (Motta & Zdrahal, 1998). En ce qui concerne l'approche WSMF, elle exploite la notion de but pour décrire les problèmes que devraient résoudre les Web Services (Cabral et al., 2004).

3.2 Les approches « Patrons »

L'approche « patrons » répond au besoin de formaliser et capitaliser des solutions à des problèmes récurrents de conception. Dans cette approche, un composant (appelé « patron ») est spécifié sous la forme d'un triplet <Problème, Solution, Contexte> (Gamma et al., 1995), (OMG, 2003), (Fowler, 1997).

3.2.1 Description des approches « Patrons »

Dans le triplet <Problème, Solution, Contexte>, le problème est un objectif de conception que souhaite atteindre le concepteur en réalisant un système d'information. La solution est un produit de conception représenté sous forme de modèles conceptuels, d'architectures... Le contexte définit la situation pour laquelle la solution décrite dans le composant est adaptée. La **figure 2** illustre partiellement un patron.



Fig. 2 – Un exemple de patron : le patron « Composite » (Gamma et al., 1995).

Ce patron fournit une représentation (sous forme d'un diagramme de classes) pour décrire des objets complexes. La structure du patron comprend l'intention, la motivation et la solution. Par analogie au triplet <Problème, Solution, Contexte>, l'intention correspond au problème à résoudre, la solution est décrite avec le langage UML et la motivation représente le contexte sous forme d'exemples d'application.

3.2.2 Analyse des approches « Patrons »

Dans ces approches, la plupart des composants proposés fournissent des solutions réutilisables qui aident à résoudre des problèmes-type de conception. Les principes mis en oeuvre par ce type d'approche concernent essentiellement les principes d'abstraction et de contextualisation. Le principe d'abstraction est vérifié par l'orientation problème. En effet dans cette approche, la description d'un composant contient de manière explicite une partie problème et une partie solution.

Les différentes approches telles que les « patrons de conception » (Gamma et al., 1995), les « patrons d'analyse » (Fowler, 1997) et les « Assets » de l'OMG (OMG, 2003) proposent différentes rubriques pour exprimer la connaissance contextuelle. Ces rubriques précisent les situations et les choix de conception qui ont conduit à la solution.

3.3 Les approches de modélisation de domaine

Une autre approche pour la conception de composants sémantiques est basée sur la modélisation de domaine. Ces approches sont issues soit du domaine de

l'Intelligence Artificielle (IA), par exemple l'approche LISA soit du domaine du Génie Logiciel (GL), telle que la méthode FODA. Ces approches visent à produire des modèles génériques ou modèles de domaine. Elles se distinguent dans la manière de considérer un domaine. Dans l'approche IA, un domaine est vu comme un ensemble de problèmes à résoudre et est conceptualisé par un ensemble de buts qui peuvent être opérationnalisés. Dans l'approche GL, un domaine est vu comme un ensemble de caractéristiques communes aux applications de ce domaine. Un domaine est aussi appelé une « ligne de produit ».

3.3.1 Les approches modèles de domaine orientées « IA » : La méthode LISA

L'approche LISA place l'utilisateur dans le contexte de résolution de problèmes. Dans le langage LISA, les buts caractérisent l'ensemble des problèmes et sous-problèmes. A chaque but sont associées des "méthodes" qui sont déclarées à priori comme les mieux adaptées pour le résoudre. Dans cette approche, le modèle de buts est relativement riche : un but est défini à l'aide d'un état-but, d'un contexte, de critères de satisfaction, de méthodes associées et de préférences. Dans la **figure 3**, le but « définition du réseau à étudier » est associé à trois méthodes possibles pour le réaliser. La première et la seconde méthode sont décomposées en deux sous buts et la troisième méthode est terminale (Jacob-Delouis & Krivine, 1995).

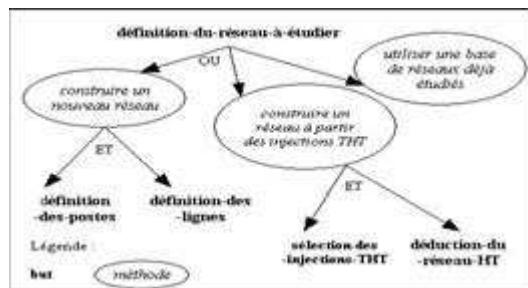


Fig. 3 – Le but « définition du réseau à étudier » et ses méthodes associées.

3.3.2 Les approches orientées « ligne de produit » : La méthode FODA

Dans l'approche FODA (Kang et al., 2003), le modèle le plus pertinent du point de vue de la réutilisation est le modèle de caractéristiques appelé « *features model* ». Ce modèle de caractéristiques permet de mettre en évidence, sous la forme d'un ensemble de hiérarchies, les caractéristiques communes et discriminantes des systèmes d'un même domaine. Les caractéristiques peuvent exprimer des fonctionnalités, des besoins en performance ou en ergonomie, mais aussi des besoins d'implantation (type de système d'exploitation...).

Le modèle précise les caractéristiques **obligatoires** pour un système du domaine ainsi que celles qui sont **optionnelles**, il fournit également les **alternatives** qui engendreront des choix au moment de la construction d'un système particulier. Le modèle définit également des **règles de composition** entre des caractéristiques. Par

exemple, l'utilisation d'une caractéristique peut imposer l'utilisation d'une autre. Enfin, les modèles de caractéristiques fournissent des **arguments** qui aident dans le choix d'options ou d'alternatives.

3.3.3 Analyse des approches de modélisation de domaine

D'un côté, LISA satisfait les principes d'abstraction en utilisant une orientation problème et de contextualisation grâce à l'introduction du concept de contexte dans la définition des buts.

Le principe de variabilité est également satisfait puisque chaque but est associé à plusieurs méthodes de résolution. Les concepts de but, méthode et contexte capturent des connaissances qui guideront l'utilisation des composants.

Le principe de localisation est aussi respecté. Le processus de recherche est un processus de satisfaction de buts induit par la hiérarchie de buts. Ce processus exploite la notion de contexte pour guider le choix de la méthode la plus adaptée au problème que souhaite résoudre l'utilisateur.

D'un autre côté, l'approche FODA respecte le principe de variabilité puisque le modèle des caractéristiques exprime tous les systèmes possibles d'un même domaine. De plus, le principe de contextualisation est satisfait au moyen, d'une part, des caractéristiques qui sont des propriétés permettant de discriminer les différents systèmes d'un même domaine et d'autre part, des arguments qui guident le choix des caractéristiques (Gurp & Bosch, 2001).

4 Comparaison des modèles sémantiques de composants

Les six principes proposés pour caractériser un modèle sémantique de composant ont permis d'étudier les différents modèles (**Table 1**).

Table 1. Evaluation des approches en fonction des principes de spécification

Principes	Approches	WSS	Patrons	Modèles de domaine	
				IA	Ligne de produit
Abstraction	-		orientation problème	orientation problème	-
Variabilité	oui (mais peu existante)		-	oui (pertinente)	
Contextualisation	formel : métalangage		informel : Phrase en langage naturel	informel : Phrase en langage naturel	
Composition	oui		-	-	
Interopérabilité	SOAP et XML		-	-	
Localisation	UDDI			Hiérarchies	

Dans cette partie, nous exploitons ces principes pour dégager des types d'approches. Nous utilisons trois points de vue conduisant chacun à une classification des modèles :

- le point de vue «**finalité**» étudie les modèles en fonction de l'usage des composants,
- le point de vue «**approche de spécification**» précise l'orientation choisie dans la nature des connaissances spécifiées au sein des composants,
- le point de vue «**support du processus d'utilisation**» caractérise le niveau d'aide fourni à l'utilisateur durant le processus de manipulation des composants.

4.1 Le point de vue finalité

Les propositions de modèles de composants sont nombreuses et leur utilisation dans le développement de systèmes d'information peut viser différents objectifs. Certaines approches sont orientées « réutilisation » et d'autres « intégration ».

-La **réutilisation** : il s'agit de considérer les composants comme des briques de développement existantes, pouvant être réutilisées dans plusieurs situations. Les principes d'abstraction et de variabilité caractérisent l'approche orientée « réutilisation ». Les logiciels de type «ligne de produit», les bibliothèques de composants relatifs à un domaine et les patrons visent essentiellement un objectif de réutilisation. Ces composants possèdent des propriétés de générnicité leur permettant d'être réutilisable dans différents contextes.

-L'**intégration** : il s'agit de considérer les composants comme des services externes (disponibles le plus souvent sur le Web). L'intégration implique d'une part, l'accès transparent et uniforme aux services et l'échange de données éventuellement hétérogènes et distribuées. D'autre part, les composants peuvent s'exécuter à distance et coopérer pour réaliser une certaine fonctionnalité. Les principes d'interopérabilité, de composition et de localisation sont caractéristiques de l'approche orientée « intégration » et sont employés essentiellement par les approches WSS.

Bien que les objectifs de réutilisation et d'intégration soient complémentaires, les approches orientées composant existantes visent de façon généralement exclusive l'un ou l'autre de ces objectifs, conduisant ainsi à des propositions de modèles de composant très différents.

4.2 Le point de vue approche de spécification

Les modèles de composants étudiés font apparaître deux orientations dans la nature des connaissances spécifiées dans les composants.

Dans l'**orientation problème** la spécification d'un composant est centrée sur l'expression d'un problème. Dans cette approche le principe d'abstraction est mis en avant en distinguant explicitement les parties problème (interface du composant) et solution (implantation). Par exemple, l'orientation problème est sous-jacente aux nombreuses approches de type patrons qui définissent un composant comme une solution réutilisable dans un certain contexte pour répondre à un problème. Il en est de même dans l'approche LISA qui utilise la notion de but pour exprimer les problèmes et les sous problèmes. Par ailleurs, l'ajout d'une couche sémantique dans les approches

WSS permet de décrire des éléments tels que le besoin auquel le service répond. Des recherches doivent cependant encore être menées pour formaliser les buts et définir de véritables modèles de buts.

A l'inverse, **l'orientation solution** spécifie les composants à un niveau logiciel. Dans ces approches, le principe d'abstraction est peu respecté. Par exemple dans FODA le concept de « caractéristique » est utilisé pour définir des propriétés d'un système (donc d'une solution). On peut noter ici que les Web Services ont été étendus pour faire émerger les Web Services Sémantiques, cette extension ayant contribué à passer d'une approche solution à une approche plus orientée problème des Web Services.

4.3 Le point de vue support du processus d'utilisation des composants

Les différents modèles de composants étudiés permettent de spécifier une connaissance relative à leur utilisation. L'intérêt de cette connaissance est de fournir une aide au moment de l'exploitation des composants. On peut envisager trois niveaux d'aide :

- **Guidage** : Le processus est entièrement exécuté par l'utilisateur. Dans les modèles de type patrons, la connaissance contextuelle n'est pas formalisée et le processus de recherche et d'assemblage reste manuel. Cependant cette connaissance exprimée en langage naturel est facilement compréhensible par les utilisateurs.

- **Semi-automatique** : Le processus est automatisé mais nécessite des choix humains. Par exemple, LISA permet d'interpréter les connaissances spécifiées au sein des buts et méthodes telles que les préférences, les contextes favorables, et les paramètres, ce qui facilite l'automatisation et l'opérationnalisation du modèle LISA. Cependant, l'utilisateur peut intervenir dans le choix des méthodes à appliquer.

- **Automatique** : Le processus est réalisé par une application logicielle et l'utilisateur n'intervient pas dans le déroulement du processus. Ce type de support est un objectif des recherches actuelles sur les WSS. Les systèmes à base d'agents sont souvent expérimentés avec l'approche WSS pour rechercher les services et les orchestrer.

5 Conclusion

Dans cet article, nous avons présenté, d'une part, les principes que doit satisfaire un modèle sémantique de composants et d'autre part, trois types d'approche orientées modèle sémantique de composants. Nous définissons un composant sémantique comme un composant intégrant dans sa spécification une connaissance relevant de son usage (pour quel problème, dans quel contexte, avec quel autre composant l'assembler...). Les approches Web Services Sémantiques (WSS), les approches « Patrons » et les approches modèles de domaine sont représentatives de cette nouvelle forme de composant. Toutes intègrent dans des formes différentes une connaissance contextuelle qui guide plus ou moins l'usage des composants.

L'étude de ces trois types d'approche a permis de dégager des différences en termes de finalité, d'orientation de spécification et de support dans le processus de manipulation.

L'évaluation de ces approches montre une évolution importante des modèles de composants. Ils deviennent de plus en plus riches pour modéliser la connaissance relative à leur usage. Il s'agit d'une évolution indispensable si l'on veut mettre en œuvre de manière systématique et instrumentée une approche orientée composant.

Références

- CABRAL L., DOMINGUE J., MOTTA E., PAYNE T., HAKIMPOUR F. (2004), Approaches to Semantic Web Services: An Overview and Comparisons. In Bussler, C. and Davies, J. and Fensel, D. and Studer, R., Eds. Proceedings First European Semantic Web Symposium (ESWS2004) The SemanticWeb: Research and Applications, Lecture Notes in Computer Science 3053(LNCS3053), pages pp. 225-239, Heraklion, Crete, Greece.
- ERIKSSON H.E, M. PENKER M. (2000), Business Modeling with UML – Business Patterns at Work, OMG Press, Wiley, ISBN 0-471-29551-5.
- FELLNER K.J., TUROWSKI K. (2000), Classification Framework for business Components, Proc. of the 3rd Hawai International Conference on System Sciences.
- FOWLER M. (1997), Analysis Patterns – Reusable Object Models, Addison-Wesley, 1997.
- GAMMA E., HELM R., JOHNSON R.E., J. VLISSIDES J. (1995), Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley.
- GURP J.VAN, BOSCH J. (2001), On the notion of Variability in Software Product Lines, Proc. of WICSA.
- HEINEMAN G., W. COUNCILL (2001), Component-Based Software Engineering: Putting the Pieces Together, Addison-Wesley.
- JACOB-DELOUIS I., KRIVINE J.P. (1995), LISA : un langage réflexif pour opérationnaliser les modèles d'expertise, *Revue d'intelligence artificielle*, Vol. 9, n°1, 1995, p. 53 à 88.
- KANG KYO KIM C., LEE K., LEEJ. , KIM S. (2003), Feature Oriented Product Line Software Engineering : Principles and Guidelines, chapter 2 in Domain Oriented Systems Development: Practices and Perspectives, Taylor & Francis, 2003, p. 29-46.
- McILRAITH SHEILA A., TRAN CAO S., AND HONGLI Z., Semantic Web Services, *IEEE Intelligent Systems*, Special Issue on the Semantic Web, (Vol. 16, No. 2) p. 46-53, March/April 2001.
- MOTTA E., ZDRAHAL Z. (1998), A Library of Problem-Solving Components Based on the Integration of the Search Paradigm with Task and Method Ontologies, dans *International Journal of Human-Computer Studies*, vol. 49, n°4, octobre 1998.
- OMG (2003), Reusable Asset Specification (RAS), Draft RFC submitted to OMG, version 2.1, août 2003.
- OWL-S (2004), The OWS Services Coalition : OWL-S : Semantic Markup for Web Services, version 1.0 available at <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>.
- SRIVASTAVA B., KOEHLER J. (2003), Web service composition, current solutions and open problems, ICAPS 2003, Workshop on Planning for Web Services, Trento, Italy.
- VARGAS SOLAR G., DOUCET A. (2002), Médiation de données : solutions et problèmes ouverts, In Actes des 2^{èmes} Assises nationales du GdR I3, Décembre 2002.

Indexation de documents AV : ontologies, patrons de conception et d'utilisation*

Antoine Isaac^{1,2}, Bruno Bachimont^{1,3}, Philippe Laublet²

¹ Institut National de l'Audiovisuel, Direction de la Recherche
4, Av. de l'Europe - 94366 Bry-sur-Marne
{aisaac,bbachimont}@ina.fr

² LaLICC, CNRS-UMR 8139, Université Paris IV-Sorbonne
Philippe.Laublet@paris4.sorbonne.fr

³ Université de Technologie de Compiègne

Résumé : Nous rappelons que l'indexation est nécessaire aux systèmes documentaires audiovisuels, et qu'elle exige ce que les langages documentaires existants ne peuvent apporter : une articulation convenable entre relations, contrôle et raisonnement. Dans le cadre de projets expérimentaux comme OPALES, nous avons pu constater qu'une indexation utilisant des ontologies pouvait remédier à cela. Néanmoins, il convient de diminuer la complexité de cette indexation pour l'indexeur ; nous proposons pour cela de généraliser l'usage de patrons d'indexation. Nous nous intéressons également à la manière dont on peut introduire ces patrons d'utilisation dès la conception des ontologies : on peut les articuler avec des patrons de conception de haut niveau, afin d'obtenir des ontologies qui soient légitimées tant sur le plan théorique que sur le plan applicatif.

Mots-clés : Indexation, Documents audiovisuels, Ontologies, Patrons de conception, Patrons d'utilisation

1 Introduction

L'indexation des documents audiovisuels (AV) est une activité ancrée dans le quotidien de l'INA. Le contexte de cette pratique est cependant en train d'évoluer : environnement technique, avec la numérisation et les outils afférents, mais aussi contexte organisationnel, avec l'émergence de petites communautés travaillant de manière précise, exigeante, à partir des documents indexés.

Pour parvenir à accorder les nouvelles techniques avec les usages existants ou pressentis, l'INA mène une réflexion s'appuyant à la fois sur sa connaissance des

*Ce travail a été initié pour le projet RIAM OPALES. Nous tenons à en remercier l'ensemble des participants, et en particulier Véronique Malaisé, avec qui nous avons travaillé pour l'élaboration de certaines des ressources ontologiques présentées ici.

pratiques documentaires, sur des travaux fondamentaux et sur des projets expérimentaux. L'un de ceux-ci, OPALES, a eu pour ambition de tester la faisabilité d'une approche d'indexation utilisant des ontologies, dans un cadre restreint à des communautés ciblées (Isaac A. et al., 2004). Notre participation à ce projet a été l'occasion de prolonger un effort, déjà ancien à l'INA, portant sur la conception d'ontologies pour l'indexation AV (Bachimont B., 2000), en nous concentrant davantage sur le volet formel et inférentiel de ces artefacts.

Nous allons d'abord voir comment les besoins en indexation AV amènent à se tourner vers la représentation de connaissances, et comment la complexité liée à un tel choix peut être réduite par le recours à des structures relationnelles pré-définies, des *patrons d'indexation*. Nous discuterons ensuite de l'introduction de ces patrons d'utilisation dans la conception des ontologies, au regard de propositions récentes centrées sur la notion de *patron de conception*. Nous montrerons en particulier comment l'utilisation de connaissances de raisonnement permet d'articuler ces deux approches en un cadre méthodologique cohérent.

2 Vers l'indexation ontologique des documents AV

2.1 Documents AV et indexation : besoins et problèmes

La recherche et la réutilisation de documents AV se font suivant des critères plus liés à leur contenu qu'aux informations catalographiques sur leur cycle de vie – date de production, auteur, etc. A l'INA, on trouve par exemple des requêtes de documents concernant « Malraux, en support film couleur » ou des « images d'actualités illustrant des tempêtes violentes »... Ce besoin concerne tout autant le public des chercheurs qui utilisent ces objets comme support de leur travail que celui des spécialistes en production audiovisuelle.

Pour utiliser le document AV, on est ainsi obligé d'*interpréter* son contenu. En effet, la forme audiovisuelle, contrairement à la forme textuelle, ne repose pas sur un système fonctionnel d'assignation conventionnelle de sens aux unités de contenu, tel que la langue (Bachimont B., 1998). En fonction d'un point de vue particulier lui permettant de choisir parmi les interprétations potentiellement infinies d'une image, un opérateur doit donc effectuer une analyse du contenu documentaire pour en déterminer les aspects intéressants à fins de recherche ou de réutilisation. Il doit ensuite reformuler le résultat de cette analyse sous une forme plus adaptée à la réinterprétation et à l'exploitation : il *indexe*¹.

Un système documentaire AV se doit donc de consigner une interprétation du contenu des documents, et d'instrumenter cette interprétation de manière efficace. Si l'index sert de point de passage obligé entre l'utilisateur du système et le document, on doit faire particulièrement attention aux conditions de sa production et de sa réception. Se pose en particulier le problème de la *continuité sémantique* : les interprétations de celui qui indexe et de celui qui accède à

¹Dans le cas du texte, on peut se servir des éléments de contenu repérables physiquement, les mots, pour indexer un document. Cette approche, à la base de la recherche de contenu textuel, demeure inapplicable dans le cas de l'audiovisuel.

l'index doivent correspondre, et être respectées lors de l'exploitation des index par un système informatique. Par exemple, pour l'indexation d'images médicales, **opération** doit renvoyer à une action chirurgicale, et non à un calcul.

Pour pallier ces difficultés, on utilise des langages documentaires comme les *thesaurus*, qui contrôlent le vocabulaire des index et replacent les termes employés dans leur contexte d'usage *via* un réseau sémantique de spécialisations et d'équivalences. On utilisera **opération chirurgicale**, notion spécialisant **action-médecine**. Un système utilisant un thesaurus peut également, en renvoyant des documents traitant de notions liées à celles d'une requête, apporter une assistance cohérente par rapport à ce qu'attendent indexeur et chercheur.

Le recours au thesaurus ne peut cependant tout résoudre. Ainsi, la relation hiérarchique structurant les champs sémantiques mélange souvent subsomption *stricto sensu* et méréconomie, et peut causer de nombreuses imprécisions : **don du sang** spécialisant **opération chirurgicale**, par exemple. Ensuite se posent des problèmes d'expressivité : il faudrait utiliser dans les index des relations entre les descripteurs, relations qui dépendraient plus des notions du domaine d'application et de leur actualisation dans les documents que des liens sémantiques génériques trouvés dans un thesaurus.

Certains thesauri proposent bien un mécanisme permettant de préciser le rôle dans la situation décrite des entités auxquelles réfèrent les descripteurs : les *facettes*. A l'intérieur d'un même domaine, on va classer les notions suivant des catégories de haut niveau – *personne*, *action*, etc. – qui indiquent implicitement des relations entre ces notions. Ce mécanisme peut se combiner à celui des *formulaires*, où des champs correspondant à un schéma adapté à l'application structurent les descriptions : par exemple, un champ *participant* apparaîtra dans un formulaire dédié à la description d'une action. Néanmoins, ces liens restent souvent très généraux, et les mécanismes proposés ne sont pas suffisamment souples pour préciser leur type en fonction du cas descriptif rencontré.

2.2 Des index sémantiques relationnels

Il faut donc que le système documentaire puisse utiliser un langage d'indexation capable d'articuler des concepts et des relations de manière explicite et contrôlée. Une solution est de se tourner vers les techniques de représentation des connaissances, et en particulier vers celle s'appuyant sur des *ontologies*.

Expressivité

Les ontologies permettent de spécifier des primitives de connaissance – concepts, relations et règles associées – pouvant être employées pour construire des index structurés adaptés à la richesse du contenu documentaire. Ainsi, dans l'une des applications du projet OPALES, nous pouvons trouver un documentaire qui contient une séquence d'archives clarifiant un discours historique et une séquence d'interview clarifiant un aspect technique. Plutôt que de livrer pêle-mêle les termes correspondants, il sera évidemment préférable de créer l'index relationnel de la figure 1.

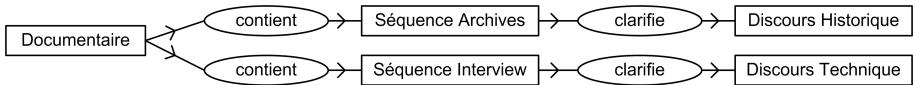


FIG. 1 – Index ontologique

Contrôle de la cohérence sémantique

Les ontologies contiennent des spécifications formelles, obéissant à une sémantique bien établie et exploitables par des systèmes de raisonnement. Ceci permet de contrôler automatiquement le vocabulaire et la structure des index : on ne peut pas utiliser n'importe quelle notion, et on ne doit pas articuler n'importe comment les notions entre elles². Les définitions formelles des notions présentes, plus précises que celles autorisées par la relation hiérarchique d'un thesaurus, peuvent être couplées à des spécifications langagières (Bachimont B., 2000). Il est ainsi possible de replacer correctement chaque concept et relation dans son contexte d'usage, et ce par des moyens immédiatement compréhensibles par des utilisateurs humains. La continuité interprétative est donc assurée d'un bout à l'autre de la chaîne documentaire.

Assistance à la recherche d'index

Plus que ne le permet un thesaurus, une ontologie peut accroître la pertinence d'un système d'information, en facilitant la correspondance entre les requêtes effectuées et les index présents. On peut en effet ajouter, aux côtés des hiérarchies de concepts et de relations, des axiomes logiques, connaissances de raisonnement qui s'appliquent à ces primitives. Par exemple, dans une ontologie dédiée à la représentation des rapports entre les éléments de contenu des documents AV, nous avons introduit une règle de composition entre les relations **contient** et **clarifie**. Muni de cette règle, un système d'inférence déduira de l'index de la figure 1 que le documentaire indexé éclaircit à la fois des questions historiques et techniques (figure 2). L'information peut sembler triviale, mais une prise en compte pertinente, même élémentaire, de ces relations propres aux domaines d'application est précisément reconnue comme un point d'amélioration des systèmes documentaires existants (Tudhope D. *et al.*, 2001).

La recherche est ainsi rendue plus robuste, ce qui permet d'alléger une partie de la charge liée à la complexité de l'indexation. En effet, même dans un contexte applicatif limité, un indexeur ne peut pas tout expliciter : il doit se repérer sur un fonds de connaissances qu'il espère partagé par le chercheur. Or réinterpréter un index conceptuel est toujours plus difficile que comprendre une liste de mots-clefs. Et reformuler une requête en cas de résultat négatif devient plus délicat. Un système de recherche ontologique, de par sa nature formelle et logique, a justement pour but de renvoyer un index s'il implique la requête

²Les domaines de relations, la hiérarchie de subsomption et des axiomes comme l'exclusion entre concepts sont des connaissances qui permettent de contrôler les assertions relationnelles.

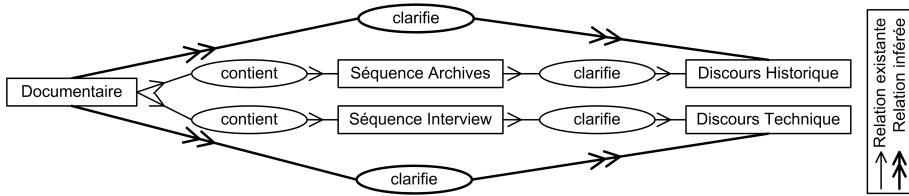


FIG. 2 – Application d'une règle d'inférence à un index

d'après les connaissances consignées dans l'ontologie, et non plus uniquement en cas de correspondance exacte. Utiliser des calculs d'inférence aide grandement à diminuer les risques induits par l'utilisation d'index conceptuels dans les processus de recherche.

2.3 Des patrons d'indexation pour la construction et l'utilisation de descriptions complexes

Le recours à des index ontologiques, s'il permet d'optimiser le fonctionnement d'un système d'information, va donc de pair avec une complexification de ce système : l'accès aux index peut être plus délicat, leur création aussi. Il devient alors nécessaire de guider l'indexeur dans la tâche difficile de sélection et d'articulation des notions rendant compte de son analyse. L'ontologie permet de contrôler la structure des descriptions, mais elle ne prescrit rien : les domaines des relations apportent bien une forme de contrainte locale, mais ils ne donnent pas la structure générale d'un index... L'expérience dans le cadre d'un projet appliqué comme OPALES montre pourtant qu'il existe une grande régularité en ce qui concerne le contenu des annotations, tant du point de vue conceptuel que structurel. Cette observation a favorisé l'utilisation de *graphes patrons* mobilisant les concepts et les relations apportés par l'ontologie.

Un patron d'indexation est une construction relationnelle adaptable, qui présente, dans un contexte d'indexation typique, l'articulation entre les concepts et les relations ontologiques les plus caractéristiques, *structurants*, d'une application. Grâce à lui, un indexeur dispose d'une structure de connaissances accessible, explicite, globalement pertinente pour son point de vue, qu'il peut aisément reconfigurer pour rendre compte de ses interprétations. Le graphe de la figure 3 montre l'un de ces patrons, tiré d'une des applications d'OPALES : l'utilisateur peut copier un tel graphe dans son index, puis le modifier en spécialisant les concepts et relations présentes, en en ajoutant d'autres ou bien en en retirant, suivant les éléments d'information reconnus pertinents. Il est à noter que les patrons d'indexation peuvent également être présentés lors de la construction d'une requête pour la guider – et en optimiser le résultat – en illustrant la manière dont la base d'index est construite. Ils jouent alors un rôle semblable à celui de formulaires d'interrogation.

L'utilisation de guides de description n'a rien de fondamentalement nouveau.

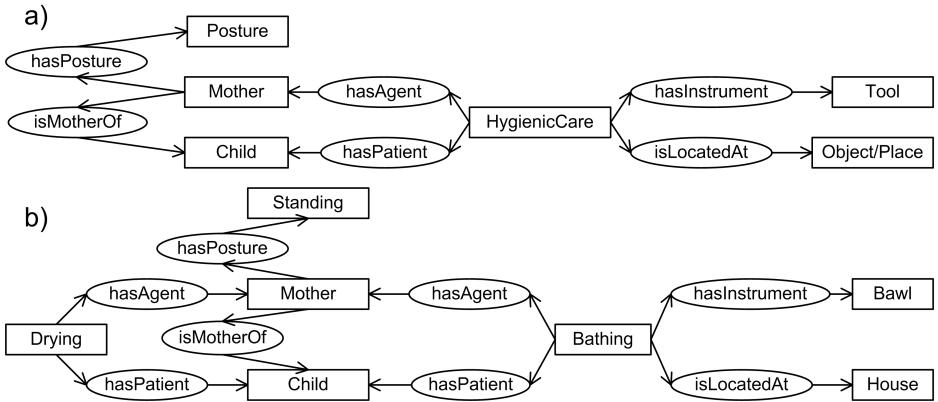


FIG. 3 – (a) Un patron d'indexation pour la description de soins sur des enfants, et (b) une spécialisation possible

Déjà, dans le cadre de l'indexation textuelle, on utilise couramment des références – *grilles* – d'indexation recommandant, pour des thèmes ou des événements précis, des descripteurs pertinents. On peut même, à la manière d'un formulaire, faire apparaître les notions préconisées dans des champs. De telles idées ont fait l'objet d'adaptations plus ou moins explicites dans des expériences de création d'annotations à base de connaissances (Schreiber A. Th. *et al.*, 2001; Hyvönen E. *et al.*, 2004). Mais, curieusement, toutes tendent à figer les descriptions produites en proposant des interfaces de type formulaire – où l'on peut par exemple choisir les concepts joints par une relation sans pouvoir modifier celle-ci – alors qu'il paraît tout à fait possible de concilier cohérence et souplesse lors de l'utilisation des patrons d'indexation. Les observations tendent en effet à exhiber des patrons plutôt simples, tant en termes qualitatifs – le niveau, basique pour l'application, auquel appartiennent concepts et relations – que quantitatifs – un patron contient généralement un petit nombre de notions.

Il est intéressant de constater que le patron a également un rôle à jouer dans la spécification du fonctionnement inférentiel du système à base de connaissances. En effet, sa configuration et les notions qu'il présente sont par définition proches des index produits, ce qui autorise à le considérer comme un structure pivot, à laquelle on peut comparer les requêtes et les descriptions de manière à faciliter leur rapprochement. De fait, dans un système ontologique riche, de nombreuses connaissances de raisonnement – définitions de concepts, axiomes relationnels – accompagnent les notions qu'il mobilise. Autour d'un patron, on peut trouver des règles qui vont soit déduire de nouvelles connaissances à partir de la structure existante (pour répondre à des questions dont les besoins ne seraient pas directement traités par le patron, figure 4a), soit ramener des connaissances « non standard » à un format plus proche de celui qui est généralement utilisé pour les requêtes (figure 4b).

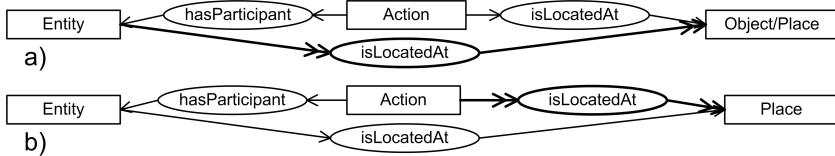


FIG. 4 – Connaissances de raisonnement et patron

2.4 Utilisation des patrons dans le projet OPALES

Au cours du projet OPALES, nous avons mis en œuvre cette approche pour deux points de vues applicatifs : la description de vidéos relatives à la petite enfance, et l’analyse de documents pédagogiques mettant en lumière l’articulation entre procédés AV et présentation d’un discours scientifique.

Ces expérimentations ont nécessité la construction de deux ontologies comprenant chacune quelques centaines de concepts et dizaines de relations, ainsi que quelques dizaines de règles d’inférences. Ici, c’est le formalisme des graphes conceptuels (Sowa J., 1984) qui a été retenu : il permet une visualisation immédiate des ressources, et des raisonnements – reposant sur des mécanismes de projection – satisfaisants au vu de nos besoins³.

Les représentants des domaines concernés (six enseignants et documentalistes pour les deux domaines) ont été impliqués dans la création et l’évaluation de ces ressources. Pour l’évaluation, une dizaine d’heures de vidéo a été finement indexée (une petite centaine d’index) à partir des graphes patrons d’indexation, ce qui a permis de valider l’adéquation de notre approche, mais aussi sa robustesse. La souplesse offerte par la combinaison des graphes patrons et des règles d’inférence a en effet permis de se limiter à un patron par point de vue applicatif, ce qui a facilité l’appropriation de cette démarche par des utilisateurs non-experts en représentation de connaissances.

3 Ingénierie ontologique et patrons d’indexation

Les patrons d’indexation sont, on l’a vu, intimement liés à l’efficacité du système d’information. Dès lors, la question de leur introduction dès l’étape de conception des ontologies se pose. Cette activité cruciale, difficile, doit offrir, au moindre coût, des garanties d’efficience pour le système qui utilise son produit. Nous allons montrer comment des *patrons de conception* peuvent être introduits, de sorte que la conception bénéficie à la fois d’une légitimité théorique liée à l’adhésion à des principes propices à consensus et d’une reconnaissance « métier » apportée par une prise en charge correcte des structures de description typiques.

³D’autres langages pourraient convenir, comme le langage RDF du w3c, pourvu qu’on les utilise avec des outils d’inférence suffisamment élaborés.

3.1 Patrons de conception

De nombreuses réflexions méthodologiques ont été conduites pour proposer des cycles de vie généraux ou des critères de cohérence « nettoyant » les ontologies, comme ceux de (Guarino N. & Welty C., 2002). Cependant, ces propositions sont plus axées sur la qualité du développement de l'ontologie lui-même que sur les compétences du système d'information qui l'utilisera : on sait comment construire, mais on ne sait pas ce qu'il faut spécifier afin que le système soit pertinent pour une application donnée. Pour cela, des travaux essaient d'initier la conception avec des composants issus d'ontologies de haut niveau satisfaisant des critères d'utilisabilité : les connaissances génériques doivent être facilement mobilisables lors de la production des ontologies, et leur contenu doit être adapté à un éventail applicatif considéré à présent comme prioritaire. On compte ainsi rendre plus explicite l'engagement ontologique plaçant la conceptualisation de l'application dans une vision « propre » du monde qui l'environne.

C'est une telle vision qui sous-tend les travaux actuels sur les patrons de conception ontologiques, ainsi nommés en référence aux *design patterns* de l'ingénierie logicielle. Gangemi, dans (Gangemi A. & Mika P., 2003), propose de réutiliser les notions de haut niveau de l'ontologie DOLCE (Gangemi A. *et al.*, 2002), obéissant aux principes de cohérence formelle énoncés par Guarino, au sein d'une structure relationnelle qui présente ces notions dans un contexte d'utilisation – ici, les « Descriptions et Situations ». Le patron qui en résulte (*D&S*, cf. figure 5) présente des *descriptions de séquences d'événements* qui ordonnent des entités temporelles (*perdurants*), des *rôles* que des entités physiques (*endurants*) peuvent jouer dans ces événements, ainsi que des *paramètres* qui sont utilisés pour décrire rôles et événements, et prennent leurs valeurs dans des *régions* plus ou moins abstraites. Cette structure générique cherche à s'abstraire d'un domaine particulier tout en résolvant un problème de description ontologique donné. Elle est aisément adaptable à la conceptualisation d'un domaine particulier, pourvu que l'usage visé par cette nouvelle conceptualisation soit similaire à celui qui a dicté l'élaboration du patron. Il suffit de rattacher aux notions qu'il introduit celles, plus spécifiques, du domaine d'application envisagé. Les objets résultant de cette adaptation forment une modélisation réduite aux notions les plus générales du domaine, dite *noyau*, qui pourra être à son tour être spécialisée pour introduire les éléments de connaissances nécessaires à des besoins plus précis.

Nous avons tenté de reproduire cette approche pour une ontologie dédiée à la description des documents AV (Isaac A. & Troncy R., 2004) : le patron D&S cible en effet un usage qui pouvait convenir à nos besoins. Pour adapter ces notions au domaine AV, nous avons appliqué le patron aux deux activités structurant le cycle de vie du document audiovisuel de télévision, avant son archivage : la production et la diffusion. Du point de vue de la diffusion, la description d'un document AV suppose un enchaînement d'événements de diffusion, tels que l'émission d'un programme sur un *canal de diffusion*. Des rôles de diffusion, comme *diffuseur* ou *récepteur*, sont joués par des entités qui peuvent être des *organisations* ou des *personnes*. Et, dans la description de ces événements, nous trouvons des paramètres comme la *date de diffusion* ou le *taux d'audience*,

valués respectivement par des *dates* et des *taux*. La figure 5 montre comment ces concepts spécialisent ceux du patron de conception pour obtenir un premier patron d'utilisation, que nous appelons de haut niveau.

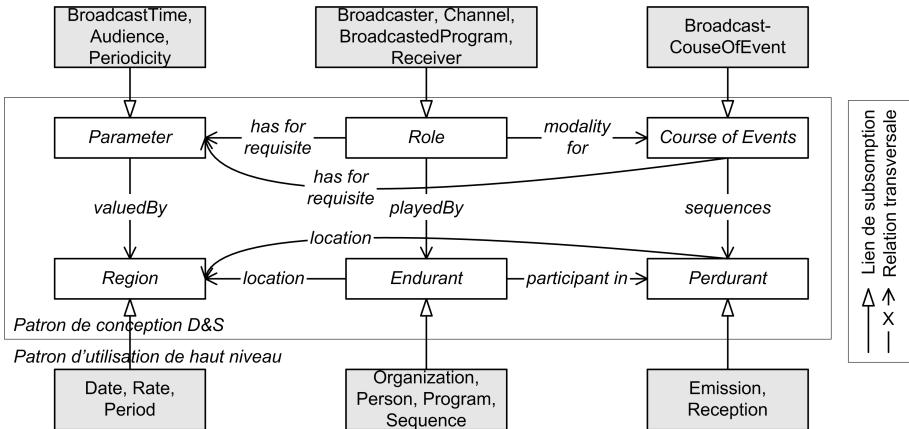


FIG. 5 – Introduction de concepts généraux de l’audiovisuel suivant le patron de conception D&S de (Gangemi A. & Mika P., 2003)

3.2 Vers une solution articulant patrons de conception et patrons d'utilisation

Rattacher une ontologie de domaine à un patron de conception de haut niveau améliore indéniablement sa qualité intrinsèque : on clarifie l’engagement ontologique en adhérant explicitement à une conceptualisation qui a bénéficié d’une effort théorique important, et qui de plus apporte des connaissances de raisonnement se répercutant, *via* la relation de subsomption, aux notions du domaine. On peut envisager que l’ontologie sera plus facilement partageable, et réutilisable par les modélisateurs d’autres domaines. Pour autant, sera-t-elle utilisable par l’indexeur ou l’expert du domaine ? La complexité des notions impliquées par des considérations abstraites peut masquer la vue applicative sur le domaine, et limiter la pertinence de l’ontologie. Spécialiser un patron de conception ne suffit pas toujours à obtenir un véritable patron d’utilisation. Pour cela, on doit faire la part entre les décisions théoriques d’un patron de conception quant à l’organisation du monde à décrire, et le domaine qui peut exiger une modélisation plus pragmatique. Dans un cadre applicatif concret, il faut veiller à se placer au niveau des notions typiques de l’application – qui ne sont pas nécessairement les notions générales du domaine – et chercher à obtenir une structure vraiment proche du besoin descriptif.

Par exemple, une étude des besoins en matière de description documentaire AV (reposant en grande partie sur l’examen des notices produites à l’INA et

concernant à ce titre autant le catalogage que l'indexation) aboutit à l'obtention d'un patron d'utilisation dont la structure relationnelle, illustrée en figure 6, est beaucoup plus simple que celle esquissée en figure 5. Les descriptions dont nous avons réellement besoin sont centrées plus sur les documents que sur les situations de production ou de diffusion : certaines subtilités informationnelles du patron de conception ne seront donc pas nécessairement à expliciter pour notre application. Par exemple, savoir qu'un document donné joue le rôle de programme diffusé dans une séquence d'événements de diffusion est relativement inutile pour nous. On préférera plutôt des propriétés simples permettant de relier directement le programme à sa date de diffusion, sa mesure d'audience, etc.

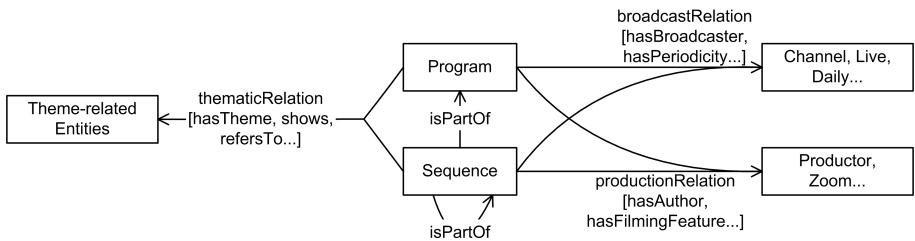


FIG. 6 – Patron de description d'un document AV

Nous proposons, pour faire coexister patron d'utilisation de haut niveau et patron applicatif au sein d'un cadre méthodologique cohérent, d'articuler les deux formes d'expression par des connaissances de raisonnement formelles. L'objectif est d'autoriser un système d'inférence à passer de l'une à l'autre à chaque fois que cela est faisable. En particulier, il faudra présenter les liens « simples » demandés par le patron d'utilisation comme autant de *raccourcis relationnels* de chemins présents dans le patron de haut niveau. Par exemple, on peut considérer qu'une relation *was broadcasted at* entre un programme et une date est utile si l'on ne veut pas dire que le programme joue le rôle de message dans une séquence d'événements qui admet pour paramètre une date de diffusion valuée par ladite date. Il nous faut alors introduire un axiome (cf. figure 7) qui permettra de gérer simultanément les concepts et les relations des deux points de vue⁴.

De telles connaissances de raisonnement, exprimées de manière formelle – et donc exploitables de façon automatique – autorisent un système à gérer simultanément un point de vue adéquat cognitivement et opérationnellement avec les usages du domaine, et un autre se référant plus à des principes de modélisation de haut niveau. Cette situation a deux avantages. D'abord, la connaissance est utilisable de manière satisfaisante pour l'application visée : on peut spécifier de manière naturelle des connaissances de raisonnement pertinentes, et on

⁴On doit remarquer qu'une équivalence parfaite n'est pas forcément atteignable. Dans l'exemple donné, une information exprimée selon le patron de haut niveau contiendra des assertions qu'on ne pourra pas déduire précisément – on utilisera des quantificateurs existentiels – de connaissances issues du patron d'utilisation.

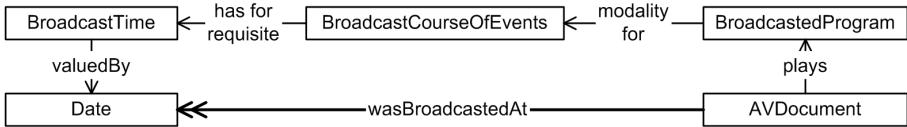


FIG. 7 – Déduction d’une relation à partir d’assertions respectant le patron D&S

prescrit bien une description adaptée aux préoccupations applicatives. Ensuite, cette connaissance acquiert un statut plus consensuel et peut être réutilisable par d’autres applications construites autour du même noyau ontologique, ce qui ne peut qu’améliorer l’intéropérabilité de ces systèmes.

4 Conclusion

L’indexation, nécessaire pour exploiter les documents AV, peut profiter de techniques permettant de créer et de manipuler des index en tant que véritables connaissances structurées. Pour cela, il faut des ontologies riches, capables d’assister efficacement le processus de recherche, mais complexes à concevoir et à utiliser. Nous avons donc utilisé des patrons d’indexation, conçus ici comme des structures relationnelles typiques d’un domaine d’application. Couplée à celle de connaissances de raisonnement, leur utilisation diminue la charge cognitive pesant sur les utilisateurs.

Poursuivant cet effort de légitimation des systèmes ontologiques, nous avons montré que dans certains cas on peut rattacher les patrons d’utilisation ontologiques à des patrons de conception de haut niveau apportant une consolidation théorique bienvenue. Pour ne pas trop s’écartez des besoins de description réels, souvent très simples, nous proposons d’utiliser des connaissances de raisonnement dédiées au passage d’un mode d’expression de la connaissance à un autre. Légitime en termes d’usages, l’ontologie le sera aussi sur le plan théorique.

Des travaux partageant nos préoccupations (Rector A. et al., 1999) notent cependant les problèmes de calculabilité et d’expressivité relatifs à de telles solutions. Dans cette approche, il s’agit bien en effet d’introduire des représentations intermédiaires entre l’utilisateur et la connaissance exprimée formellement, mais celles-ci sont pas gérées par les mécanismes de raisonnement ontologiques standards. Ce choix est dû à la décision fondamentale d’exclure ces « connaissances d’encodage » des ontologies, mais aussi aux capacités parfois restreintes des systèmes de raisonnement concernés.

Si nous-mêmes considérons que les patrons et leur articulation avec des connaissances plus abstraites font bien partie de la spécification d’une conceptualisation *située*, nous ne devons cependant pas négliger le problème de l’opérationnalisation des connaissances de raisonnement. A un travail qui a validé une hypothèse relative à l’utilisation des ontologies – les patrons d’indexation – puis détaillé des éléments de conception et d’exploitation de ces connaissances – le couplage

entre patron de conception et d'utilisation – il conviendrait d'ajouter des tests opérationnels à plus grande échelle. Il importe de valider un choix de formalismes et d'outils dont les possibilités expressives et l'efficacité calculatoire permettront de tirer pleinement parti d'une approche dont les retombées en termes d'usage sont indéniables.

Références

- BACHIMONT B. (1998). Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. *Document numérique*, **2**(3-4).
- BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis*. Eyrolles.
- GANGEMI A., GUARINO N., MASOLO C., OLTRAMARI A. & SCHNEIDER L. (2002). Sweetening ontologies with DOLCE. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2002)*, Siguenza, Spain.
- GANGEMI A. & MIKA P. (2003). Understanding the semantic web through descriptions and situations. In *International Conference on Ontologies, Databases and Applications of Semantics (ODBASE'03)*, Catania, Italy.
- GUARINO N. & WELTY C. (2002). Evaluating ontological decisions with Onto-Clean. *Communications of the ACM*, **2**(45).
- HYVÖNEN E., SAARELA S. & VILJANEN K. (2004). Application of ontology techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, Heraklion, Greece.
- ISAAC A., COUROUNET P., GENEST D., MALAISÉ V., NANARD J. & NANARD M. (2004). Un système d'annotation multiforme et communautaire de documents AV : OPALES. In *Journée sur les Modèles Documentaires de l'Audiovisuel, SDN 2004*, La Rochelle, France. <http://archivesic.ccsd.cnrs.fr/>.
- ISAAC A. & TRONCY R. (2004). Designing an audio-visual description core ontology. In *Workshop on Core Ontologies in Ontology Engineering, 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004)*. CEUR online Proceedings, <http://ceur-ws.org/Vol-118/>.
- RECTOR A., ZANSTRA P.E., SOLOMON W. D., ROGERS J. E., BAUD R. & AL. (1999). Reconciling user's needs and formal requirements : Issues in developing a re-usable ontology for medicine. *IEEE Transactions on Information Technology in BioMedecine*, **4**(2).
- SCHREIBER A. TH., DUBBELDAM B., WIELEMAKER J. & WIELINGA B. (2001). Ontology-based photo annotation. *IEEE Intelligent Systems*, **16**(3).
- SOWA J. (1984). *Conceptual structures : information processing in mind and machine*. Addison-Wesley, Reading (MA US).
- TUDHOPE D., ALANI H. & JONES C. (2001). Augmenting thesaurus relationships : Possibilities for retrieval. *Journal of Digital Information*, **1**(8).

Modélisation de l'activité d'annotation discursive pour la conception d'un collecticiel support à l'herméneutique

Gaëlle Lortal¹, Myriam Lewkowicz¹, Amalia Todirascu-Courtier²

¹ Laboratoire CNRS ISTIT, équipe Tech-CICO, Université de technologie de Troyes,
{lortal, lewkowicz}@utt.fr

² Université Marc Bloch de Strasbourg, ea 1339
amalia.todirascu@umb.u-strasbg.fr

Résumé : Dans un contexte où les échanges médiatisés s'accroissent, le document numérique devient central. Pour soutenir les échanges et la construction d'une interprétation collective autour de ce document, il est nécessaire de développer des collecticiels pertinents, support à l'herméneutique. Dans ce cadre, l'annotation est un fragment de discours à propos d'un texte, un support à l'argumentation. Dans cet article, il s'agit de présenter une démarche de conception d'outil fondée sur la modélisation de l'activité d'annotation discursive. Ce modèle est issu des travaux en rhétorique antique et médiévale sur la production de discours. De ce modèle d'activité d'annotation, nous dégagons des primitives de conception de l'outil support à l'herméneutique.

Mots-clés : Annotation, discours, création collective de sens, conception de collecticiel.

1 Introduction

La place omniprésente des documents dans nos organisations a donné lieu à des travaux de recherche focalisés sur le document, comme par exemple en France, les travaux au sein du Réseau Thématisé Pluridisciplinaire 33 : « Documents et contenu : création, indexation, navigation » (RTP-DOC). Ce réseau distingue trois orientations de recherche associées au document : l'analyse du document comme forme (études de la structure du document pour sa manipulation), comme signe (étude de l'intentionnalité du document), et comme médium (étude du statut du document dans les relations sociales) (Pédauque, R.T., 2003). Le travail de recherche que nous présentons dans cet article s'intègre dans la seconde problématique qui considère le document « porteur de sens et doté d'une intentionnalité [...] indissociable du sujet en contexte qui le construit ou le reconstruit et lui donne sens » (Pédauque, R.T., 2003, p. 3). Dans cette vision du document comme signe, on s'intéresse plus particulièrement à la création d'un document, à l'interprétation d'un document, c'est-à-dire aux signes qui le constituent. Ces questions sont abordées dans cet article sous l'angle de la lecture *critique* des documents, que nous opposons à une lecture qui ne

serait guidée par aucun principe productif, qui ne viserait ni un savoir, ni la production d'un autre texte.

Dans cet article, nous présentons tout d'abord l'annotation comme élément discursif central d'une lecture critique. Nous exposons ensuite les principes méthodologiques pour la conception d'un collecticiel support à l'herméneutique, puis faisons le point sur les travaux existants sur la modélisation des activités liées à l'écrit. Nous proposons (section 5) un modèle de l'activité de production de discours issu de la rhétorique, décliné (section 6) dans le cadre d'une activité instrumentée. Ce modèle sert de base à la conception du collecticiel dont nous décrivons les fonctions (section 7). Nous précisons enfin comment nous proposons d'utiliser des techniques issues du Traitement Automatique de la Langue (T.A.L.) afin d'appareiller le support textuel.

2 L'annotation : une production discursive de la lecture critique

Une lecture critique est productive d'une interprétation qui éclaire non seulement le texte lu, mais également d'autres textes. Elle permet de contextualiser le texte et de le faire entrer ainsi dans un contexte discursif. Cette lecture critique peut être soutenue par un ensemble de fragments textuels reliés au texte et formant son « contexte » (contexte textuel), les annotations. Elle peut donc être productive d'un autre texte, commentaire ou critique, le corps textuel de l'annotation marquant une argumentation autour d'un texte, un point de vue d'un lecteur. Nous nous intéressons plus particulièrement à une lecture critique à plusieurs, permettant la construction d'une interprétation partagée du document initial. Cette élaboration d'une interprétation partagée au sein d'un collectif participe selon nous à la construction collective du sens (Weick, 1979). Ce dernier conçoit en effet la construction collective du sens dans les organisations (*collective sensemaking*) comme un processus de réduction collective de l'ambiguïté perçue d'une situation. C'est en échangeant, en débattant, que les membres de l'organisation vont clarifier puis partager des compréhensions de situations (retranscrites dans des documents), ce qui construira du sens petit à petit. Les travaux de Weick mettent l'accent sur le processus de création du sens, son émergence et son évolution, et non pas sur la représentation collective du sens. Le sens collectif n'est donc pas forcément un sens partagé par un collectif. L'interprétation collective de documents, traces des actions menées dans l'organisation, permet de tirer parti des documents tout en étant capable de sortir éventuellement du cadre dans lequel les documents ont été rédigés. Ce processus est également le support de l'identité individuelle, car, par le biais de ces interactions, chaque acteur met à l'épreuve et fait évoluer son identité.

Nous proposons de soutenir l'herméneutique, l'élaboration de discussions critiques ou explicatives autour de textes, en développant des stratégies d'interactions médiatisées autour des documents numériques, souvent textuels. L'herméneutique consiste à recréer un sens autour d'un texte qui ne possède plus qu'intrinsèquement son contexte. Ce sens laissé au sein du texte même par l'auteur est redécouvert par une succession d'énoncés expliquant une interprétation, se complétant, se répondant,

construisant de nouvelles pistes d'interprétation. L'interprétation des textes est traditionnellement accompagnée de gloses, commentaires et autres annotations ancrées au texte même, ou reliant différents textes ou fragments de texte. Nous proposons donc de soutenir cette collaboration discursive autour des documents par un système permettant l'annotation de ces documents, avec une finalité d'interprétation et d'appropriation, finalité non assistée par les outils informatiques d'annotations textuelles actuels, qui ne permettent de déposer que des annotations isolées, du type commentaire textuel, pauvrement indexées (date, nom d'auteur) et difficilement utilisables comme support aux interactions au sein d'un collectif. Or dans un contexte herméneutique d'interprétation méthodique de textes, le corps textuel des commentaires est placé au rang de discours, son contexte est formé notamment par le rôle de l'auteur, le contenu sémantique, la place de cette annotation dans le fil de discussion. Cette contextualisation est essentielle pour tracer la logique de conception d'une interprétation, voire d'un concept.

Des travaux ont été menés au KMI (Knowledge Media Institute) sur ces fonctions de commentaires discursifs d'un document. Ils ont donné lieu au "Digital Document Discourse Environment" (D3E) (Sumner et al., 2000), outil web dans lequel des échanges de messages « autour » d'un document peuvent avoir lieu, mais la conception de cet outil n'étant pas liée à une étude de l'activité d'analyse documentaire, il n'existe pas de réflexion sur le processus coopératif d'interprétation partagée. De plus, la réflexion sur la visualisation et la réutilisation de ces messages est peu approfondie ; en effet, les messages sont présentés de manière arborescente, et indexés selon des attributs standards (date, auteur, titre). Tout se passe comme si on avait associé un forum au document. Or, de nombreux travaux soulignent que les discussions en ligne sont souvent désorganisées et confuses, à cause du développement fréquent de multiples fils de discussion et de conversations parallèles. On peut citer par exemple (Marcoccia, 2004), qui évoque le phénomène de digression thématique à l'intérieur d'un forum, qui se fait progressivement, en parcourant une chaîne de messages introduisant chacun un développement thématique par rapport au message précédent. Le résultat peut être une véritable « décomposition thématique » (*topic decay*, Herring, 1999).

3 Principes méthodologiques pour la conception d'un collecticiel support à l'herméneutique

La situation de recherche dans laquelle nous évoluons nous amène à définir de nouvelles pratiques à assister (l'interprétation collective de documents numériques). De ce fait, un processus de conception classique en informatique basé sur une analyse des besoins, ou sur une analyse de l'activité existante, pour en déduire des primitives de conception, n'est pas adapté. Comme le dit Tchounikine : « la recherche en informatique a donc ici un rôle fondamental, celui d'inventer de nouveaux possibles » (Tchounikine, 2002-b, p.207).

La démarche que nous proposons est inspirée du positionnement méthodologique adopté dans le champ de la conception des EIAH (Environnements Informatiques pour l'Apprentissage Humain) par Baker (Baker, 2000), repris par Tchounikine

(Tchounikine, 2002-a). Ces auteurs distinguent les modèles comme outil scientifique des modèles pour la conception de systèmes. Les premiers permettent d'utiliser une théorie pour comprendre ou prédire une situation ou une activité ; les seconds traduisent les premiers en un modèle permettant la conception et l'implémentation de systèmes supports à la situation ou à l'activité.

Or les théories issues des sciences humaines habituellement mobilisées lors de la conception de collecticiels (théorie de l'activité, théorie de l'apprentissage, théorie de l'agir communicationnel...) sont difficiles à exploiter telles quelles pour en déduire des primitives de conception, ou il est difficile de transposer leurs définitions dans un cadre médiatisé par un système informatique. Le travail de conception consiste donc à définir de nouveaux modèles, avec de nouveaux concepts, en accord avec la théorie, pour décrire l'artefact assistant et traçant les interactions. La théorie nous permettra ensuite d'analyser les traces ainsi mémorisées.

Nous proposons donc la démarche suivante, illustrée en figure 1 : dans le cadre d'une théorie en sciences humaines ou sociales adaptée aux phénomènes que l'on souhaite assister/observer, nous proposons un modèle de description de ces phénomènes. Ce modèle opérationnalise la théorie et est une base de réflexion à la définition des situations dans lesquelles ces phénomènes seraient médiatisés à l'aide d'un système informatique. Cette réflexion conduit à un modèle de l'activité instrumentée, mettant en jeu à la fois les chercheurs en sciences sociales garant du modèle de description, et les chercheurs en informatique, comprenant et maîtrisant les propriétés propres des outils informatiques. Ce modèle de l'activité instrumentée est ensuite matérialisé dans un modèle de conception, spécification du collecticiel. Ce collecticiel supportant les interactions est également un moyen privilégié de recueil de corpus. Ce corpus, analysé à l'aide de la théorie mobilisée, nous permettra de faire évoluer notre compréhension des phénomènes à l'étude.

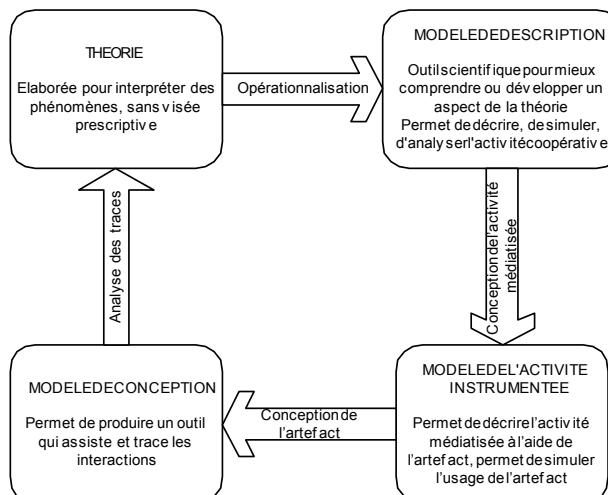


Fig. 1 - Démarche de conception d'un collecticiel

Il nous semble que bien que la phase de conception de l'activité instrumentée soit toujours présente lors de la conception d'un artefact informatique, les activités de cette phase ne sont pas explicitées. Tout se passe comme s'il était possible de définir directement des primitives de conception d'un artefact assistant une activité à partir de la modélisation du déroulement de cette activité en face à face, que cette modélisation soit réalisée par des chercheurs en sciences sociales ou non. Or nul ne nierait que cette instrumentation a un impact sur l'activité. C'est au cours de cette phase de conception de l'activité médiatisée que les échanges entre chercheurs en sciences humaines (SHS) et chercheurs en technologies de l'information et de la communication (STIC) vont pouvoir s'exprimer afin de construire un modèle commun reflétant à la fois les principes directeurs de l'activité et les possibilités en terme d'assistance. Cette phase permet le passage de relais vers des questions de conception qui seront explicitées dans un modèle de conception décrivant les fonctions de l'outil.

Dans la section suivante, afin de définir le modèle de description le plus adapté à notre problématique d'herméneutique, nous évoquons les travaux ayant analysé les activités centrées sur les documents.

4 Quelle théorie pour l'analyse de l'activité de production de discours ?

Dans le domaine de la psychologie cognitive de nombreux chercheurs ont étudié les activités mentales liées à l'écrit, en distinguant les activités de compréhension de texte et de production de texte.

En ce qui concerne les modèles de compréhension, l'accent a été mis sur la mise en mémoire de fragments de texte, fragments nécessairement résumés. Un des modèles les plus cités dans ce domaine est celui du processus de construction-intégration de (Kintsch, 1988) ; La compréhension de texte y est décrite comme un cycle alternant des phases de construction d'une représentation mentale cohérente d'un texte en cours de lecture et des phases de sélection ou non des fragments de texte pour la mise en mémoire (intégration). Des recherches ont été menées pour utiliser cette théorie descriptive à des fins constructives, par exemple pour la définition de principes de conception de documents hypermédia facilement appropriables par le lecteur (Iksal, Garlatti, 2000). Ces auteurs proposent un guide de « bonnes pratiques » pour la conception de documents, afin d'assurer notamment la cohérence du texte. Ces documents sont ensuite présentés de manière à ce que le lecteur soit assisté dans la construction de son modèle mental, l'objectif étant la minimisation du coût cognitif lors de la lecture des documents.

En ce qui concerne les modèles de production, l'accent est mis sur les processus rédactionnels de planification, de mise en texte et de révision, et le modèle de contrôle permettant d'appliquer ces processus. Les auteurs fréquemment cités dans ce domaines sont notamment (Hayes, Flower, 1980) qui ont proposé des modèles de stratégies rédactionnelles. Là encore, cette théorie descriptive a été utilisée dans des travaux qui ont donné lieu à des logiciels support aux processus rédactionnels. On

peut citer par exemple les travaux de (Piolat et al., 1989) qui utilisent la combinaison de trois logiciels (scripsis, scripap, scriprev), chacun étant focalisé sur un processus (planification, mise en texte, révision). L'objectif de ces travaux n'est toutefois pas de proposer des outils de production de texte au sein d'une organisation, mais de fournir un cadre à l'étude expérimentale de la production de texte.

Comme nous l'avons présenté en section 3, notre approche consiste à concevoir un collecticiel sur la base d'une analyse de l'activité collective que ce collecticiel entend assister. Les modèles descriptifs de compréhension ou de production proposés par la psychologie cognitive que nous avons cités ci-dessus, ne nous paraissent pas appropriés pour la conception d'un outil support à l'herméneutique car ils séparent des phases de mise en mémoire et de mise en texte. En effet, l'herméneutique mêle l'activité d'écriture durant la lecture – annotations – à celle de lecture pour produire du sens. Les phases de lecture/mémorisation et de rédaction/intégration sont donc associées. Dans des conceptions liées à la didactique de l'écrit, la lecture et la rédaction sont également considérées comme des phases d'une activité générique liée au support écrit (Barré de Miniac, 2000). Nous proposons donc de mobiliser un modèle de la production de discours issu de la rhétorique antique et médiévale représentant la mémorisation et la production discursive en un cycle complet.

5 Modèle de production de discours

L'écrit est le lieu d'interactions complexes et évolutives entre des facteurs affectifs, cognitifs et linguistiques (Barré de Miniac, 2000). Nous nous intéresserons plus particulièrement aux facteurs cognitifs en tant que facteurs organisateurs des concepts en mémoire et en texte et aux facteurs linguistiques en tant que marques à la fois d'un type de discours spécifique et de la sémantique du document en « contexte ».

Nous retrouvons ces deux types de facteurs dans la rhétorique. Des théories rhétoriques d'Aristote à celles d'Hugues de St Victor en passant par Cicéron ou Quintilien, la production de discours est enseignée suivant un processus défini. La rhétorique aristotélicienne se focalise sur une production finale de discours oral sans nier pour autant une phase mémorielle nécessaire à toute production. Cette phase de mémorisation est mieux représentée par la rhétorique que nous appellerons mémorielle portée par des penseurs cités par (Carruthers, 1990), tels que Quintilien (*L'institution oratoire*), Cicéron (*De oratore*, *De inventione*) ou Tullius (*Ad Herennium*) dans l'Antiquité, puis Hugues de St Victor (*Didascalicon*), Fortunatianus (*Artis rhetoricae libri tres*) ou Julius Victor (*Ars rhetorica*) au Moyen-Âge. Dans cette approche de la rhétorique, un continuum entre la partie mémorielle plus « logique » ou « dialectique » et la partie stylistique, rédactionnelle, est observable. La rhétorique est considérée comme une alliance entre structuration et éloquence.

Le processus de production de discours tel qu'il est préconisé dans ce contexte comporte deux phases : « Divisio » et « Compositio ». La Divisio se fait au cours de la lecture et représente l'étape de division d'un texte en unités intelligibles, en segments brefs mémorisables. La Compositio, elle, est l'assemblage ordonné, l'agencement convenable des « res » (objets conceptuels comme physiques) des segments mémorisés (fig.2). Ces phases de mémorisation, la Divisio, et de création, la Compositio, sont elles-mêmes divisées en étapes soutenues par l'utilisation d'annotations.

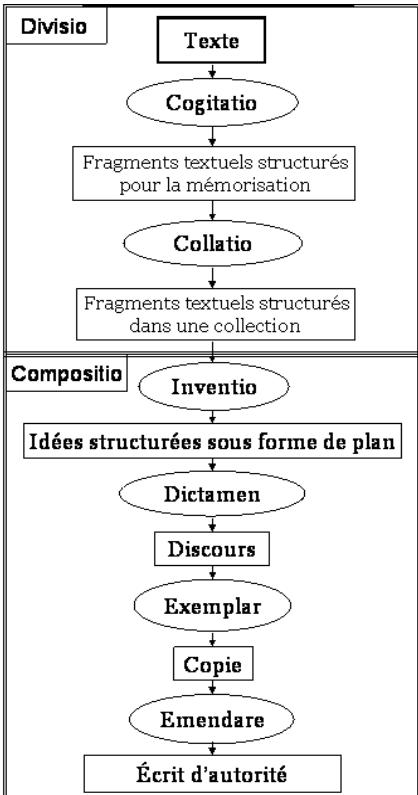


Fig. 2 : modèle de production de discours

est proche de celui de Collatio dans la mesure où il s'agit de créer des liens sémantiques entre divers éléments mémorisés, au niveau de la « res » (objets conceptuels, idée) pas au niveau du mot. Un plan est formé, c'est-à-dire un ensemble hiérarchisé d'idées, une structure argumentative par exemple.

La phase suivante sera celle de mise en mot de ce plan conceptuel, une phase classique de rédaction, le « Dictamen ». Nous voyons à cette étape la création physique du discours, classiquement sur support encore modulable (brouillon), où seul le style, le choix des termes, donc la forme textuelle du discours peut être modifiée.

La phase d'Exemplar n'est que la mise en support pérenne d'un discours strictement identique à ce que l'on trouve en sortie du processus du Dictamen.

La dernière phase et non la moindre dans cette succession de processus est celle de l'Emendare où la copie finale du discours diffusée est commentée publiquement, par l'ajout des commentaires, « notae » ou arguments d'un auteur au texte original, faisant ainsi du texte une référence, un écrit faisant autorité.

Ce modèle représente un mode de production de discours fortement soutenu par la mémoire. Dans un contexte de travail collaboratif médiatisé par ordinateur, la création discursive doit être soutenue par un outil adéquat permettant de stocker, de créer et partager les informations. Afin de concevoir cet outil, nous souhaitons donc tout d'abord modéliser cette activité de production de discours instrumentée, activité que nous n'analysons ici que dans un cadre de documents numériques textuels.

6 Modèle de production de discours instrumentée

La déclinaison du modèle de production de discours dans un cadre instrumenté nous permet de définir les étapes suivantes à préconiser (figure 3) : tout d'abord, le texte du document est *segmenté* pour être mis en mémoire sous la forme de fragments mémorisables. Ces segments sont ensuite *indexés* pour éviter la perte de la structuration du document comme unité. Il est important d'indexer les segments chronologiquement pour marquer la hiérarchie des différents paragraphes dans un texte, des différents mots dans un paragraphe,... Ce type d'indexation concerne toutes les métadonnées associables automatiquement à un élément déposé (localisation, auteur, date, ...). L'indexation doit également servir à lier les nouveaux fragments déposés à l'ensemble conceptuel existant. Nous obtenons alors un ensemble de segments textuels liés sémantiquement à d'autres segments textuels. La phase de *structuration* représente un processus de hiérarchisation, d'organisation des idées, selon un plan chronologique. Un plan détaillé est défini, contenant toutes les idées nécessaires à la mise en mot du discours. C'est la phase où les « res » (concept) contenus dans les fragments textuels indexés sont réutilisés et réorganisés en un nouveau document. La

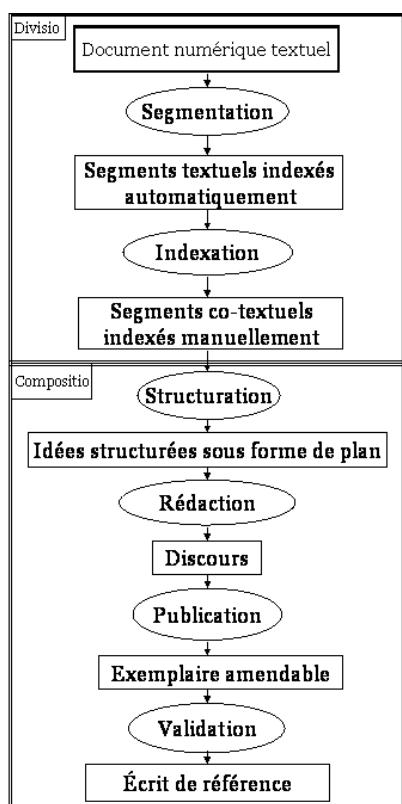


Fig. 3 – Modèle de production de discours instrumentée

phase de *rédaction* est une mise en texte du plan donnant en résultat le discours. Ce discours n'est pas l'objectif final de cette activité dans cette vision de la rhétorique, puisqu'il est ensuite *publié* pour en faire un objet amendable par d'autres lecteurs/auteurs de la communauté. Cette phase de commentaire par d'autres membres de la communauté est primordiale pour permettre la *validation* de l'exemplaire, voire son amélioration, et constituer un écrit d'autorité, un discours faisant référence dans la communauté. L'annotation d'un document dans un cadre herméneutique consiste donc selon nous à suivre un processus de mise en discours d'idées organisées. Il s'agit en effet, suite à la lecture d'un document, d'engager un processus qui permet d'ajouter une idée, une opinion, structurée sous forme textuelle.

Par exemple, dans un contexte de travail collaboratif, on peut considérer la mise en partage d'un document afin qu'il soit commenté. Après une phase de visualisation du texte, d'une lecture, un segment sera mis en valeur de façon à signaler l'ancre d'un élément discursif en rapport avec ce segment. Cette mise en valeur pourra se faire par des techniques classiques de surlignage, soulignement, encadrement, colorisation de segments de tailles variables (du mot, ou d'une partie d'un mot, au paragraphe, ou d'un ensemble disjoint d'éléments). Suite à cette segmentation et ce choix d'élément(s) à annoter, une phase d'indexation pourra intervenir, consistant en une mise en relation des segments. Il sera alors question de tirer des liens sémantiques entre des éléments pour les structurer entre eux et former un ensemble de segments textuels organisés selon leur sens, donné par l'utilisateur. L'annotation consistera en effet en un ancrage, une relation géographique, mais aussi en un corps, un discours qui fait sens, qui prend place dans un « co-texte », l'ensemble des segments textuels en mémoire indexés par des mots-clés compréhensibles et structurants pour un utilisateur humain. Lors de la rédaction de cette annotation, une phase d'organisation du discours à écrire sera nécessaire, c'est la structuration des « res », des concepts en mémoire, qui donnera naissance à un plan constitué d'arguments structurés hiérarchiquement. La phase de rédaction permettra de constituer le corps de l'annotation qui sera lisible par un membre de la discussion herméneutique après sa publication et donc sa diffusion.

Tout comme un texte de référence, l'annotation peut être entérinée grâce à un nouveau lien amené à celle-ci. Une réponse à un commentaire permet alors de participer au fil de discussion initié par la première annotation. Le passage du commentaire individuel à l'annotation discursive marquant la naissance d'une argumentation autour d'un document s'articule grâce à un modèle de coopération sous-jacent représentant les interactions créées par l'utilisation d'annotation. Ce modèle est en cours d'élaboration et n'est donc pas présenté ici.

Le modèle de production de discours instrumentée présenté dans cette section est issu d'un modèle d'activité de production de discours venant de la rhétorique. Il nous permet dans un premier temps de représenter la rédaction individuelle d'annotation et dans un second temps de décrire les spécifications d'un collecticiel assistant ce type de production discursive par le biais d'annotation.

7 Pistes de réflexion pour la conception d'un collecticiel support à l'herméneutique

Conformément au modèle que nous préconisons, le collecticiel doit permettre aux utilisateurs de visualiser un document, de le segmenter, de créer des associations de divers types (indexation, assemblage) entre les différents fragments, de rédiger le discours du corps de l'annotation, ou de la publier. La phase de validation, optimisant la collaboration au travers des réponses à un discours, nécessite la mise en place d'une fonction d'association spécifique de type « réponse à » une annotation. Le modèle discursif devant permettre un aller-retour constant entre lecture et écriture, la fonctionnalité de visualisation est également prépondérante. Entre autres pour cet objectif, l'utilisation d'un plug-in dans un navigateur semble pertinente. En effet, n'étant qu'un ajout à un navigateur naturellement utilisé et donnant accès en lecture à de nombreux documents du Web, ce plug-in permet de visualiser à la fois le document et le corps de l'annotation en cours de rédaction, ou en cours d'indexation. Cette annotation est saisie dans une fenêtre « pop-up », puis indexée pour autoriser sa récupération après publication et la création d'un ensemble de documents structurés.

Nous proposons d'adopter une architecture client-serveur respectant le standard Annotea du W3C (Kahan et al., 2001), pour la gestion des annotations et le développement d'outils d'annotation, basé sur une description des annotations en RDF améliorant la collaboration au travers de métadonnées partagées. Le serveur d'annotations Zannot respectant ce standard (Zannot, 2003) conserve les annotations dans une base de données RDF et les utilisateurs peuvent interagir avec le serveur par le client Annozilla (Annozilla, 2004) afin de rechercher, créer ou supprimer une annotation. Une annotation suit une notation en RDF qu'il est possible de personnaliser (ajout de valeurs d'attributs au schéma d'annotation Annotea), ce qui permet d'ajuster le modèle en fonction de notre besoin. Nous avons choisi de réutiliser le plug-in de Mozilla nommé Annozilla qui fournit l'interface de gestion de l'annotation. Nous proposons de l'augmenter avec des fonctionnalités plus précises d'indexation de l'annotation et de visualisation des annotations en fonction des critères d'indexation.

L'indexation de l'annotation peut se faire automatiquement par l'outil (date, auteur, codifications des annotations répondues, fil de discussion chronologique automatique) ou manuellement par l'utilisateur, selon trois dimensions : domaine, argumentation, organisation (Zacklad et al., 2003). Dans le second cas (choix d'une valeur représentant le contenu de l'annotation pour chacune des dimensions), l'indexation peut être fastidieuse et nous souhaitons donc soutenir l'utilisateur dans cette tâche en lui proposant des termes (mots-clés) automatiquement grâce à des outils de Traitement Automatique de la Langue (T.A.L.). Pour cette opération, nous souhaitons utiliser un algorithme de mise en correspondance entre le corps de l'annotation et trois ontologies dimensionnelles semi-formelles. Le système d'annotation proposera alors à l'utilisateur des mots-clés pour chacune des dimensions. L'utilisateur décidera ensuite si l'indexation proposée est pertinente et s'il faut alors la conserver comme métadonnée de son annotation. Une fois la validation effectuée par l'utilisateur, l'annotation est stockée avec ses métadonnées sur le serveur d'annotations.

8 Conclusion, perspectives

Dans un environnement où le document numérique est omniprésent, nous nous sommes penchés sur une problématique spécifique de création de sens, l'interprétation collective de document, appelée herméneutique dans la tradition des textes, et soutenue par l'activité d'annotation. L'annotation est un medium privilégié d'interprétation collective engageant une communication dans un collectif, un fil de discussion. Afin de soutenir cette activité dans un cadre médiatisé, nous proposons de concevoir un collecticiel support au discours via les annotations. Les fonctionnalités de celui-ci se fondent sur un modèle d'activité de production de discours représentant les différentes étapes de production d'une annotation discursive. Nous proposons pour cet outil une architecture respectant le standard Annotea du W3C et des fonctionnalités soutenues par des techniques de T.A.L.

Dans le cadre d'une conception itérative du collecticiel, une maquette de l'outil a montré la faisabilité de l'approche. Finalisée, elle permettra une évaluation de nos hypothèses sur le modèle de production de discours, mais aussi sur le statut et les objectifs de l'annotation.

En parallèle, une expérience sur le processus d'annotation dans un groupe de conception est en cours et les données recueillies permettront d'avancer sur la définition des dimensions argumentative et organisationnelle indexant les annotations.

Les annotations seront observées non seulement pour affiner une typologie des annotations dans le cadre d'activités de conception collaborative, mais aussi pour proposer un modèle du processus d'interaction au travers des annotations déterminant les fonctionnalités coopératives de l'outil.

Références

- ANNOZILLA (2004), <http://annozilla.mozdev.org/>
- BAKER M. (2000). The roles of models in Artificial Intelligence and Education Research: a prospective view. *International Journal of Artificial Intelligence in Education Research*. Vol 11(2), p. 122-143.
- BARRE DE MINIAC C. (2000) *Le rapport à l'écriture : Aspects théoriques et didactiques* coll. Savoirs mieux Ed. Septentrion Presses Universitaires, Ch. Barré de Miniac.
- BUITELAAR P., OLENIK D., HUTANU M., SCHUTZ A., DECLERCK T., ET SINTEK, M. (2004), *Towards Ontology Engineering Based on Linguistic Analysis*, in Proceedings of LREC'2004, Lisbon, may 2004, ISBN 2-9517408-1-6, pp.7-11
- CARRUTHERS M. (1990) *The Book of Memory: A Study of Memory in Medieval Culture*. New York: Cambridge University Press.
- CIMIANO, P. HOTHO, A., ET STAAB S. (2004), *Clustering Concept Hierarchies from Text*, in Proceedings of LREC'2004, Lisbon, may 2004, ISBN 2-9517408-1-6, pp. 1721-1724.
- FAYOL M. (1997). *Des idées au texte : psychologie cognitive de la production verbale, orale et écrite*. Paris: PUF.
- GARLATTI S., IKSAL S. (2000). Méthodologie de conception de documents électroniques adaptifs sur le Web. IN . GAIO, M., TRUPINCIDE, E., *Document Electronique Dynamique, Actes du troisième colloque international sur le document électronique : CIDÉ'2000*.

- HAYES J. R. & FLOWER, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, NJ: Lawrence Erlbaum.
- HERRING, S.C. (1999). Interactional Coherence in CMC. *Journal of Computer-Mediated Communication* 4(4) : www.ascusc.org/jcmc/vol4/issue4/
- JACQUEMIN C. ET BOURIGAULT D. (2003), Term Extraction and Automatic Indexing, in Mitkov R. (ed), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2003, pp. 599-615
- JACQUEMIN, C., ET TZOUKERMAN, E. (1999), NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25-74, Kluwer, Boston, MA, 1999
- KAHAN J., KOIVUNEN M.-R., PRUD'HOMMEAUX E., ET SWICK R.R. (2001) *Annotea : an open RDF Infrastructure for Shared Web Annotations*, Proceedings of WWW10, May 1-5 2001, Hong-Kong, pp. 623-632.
- KINTSCH W. (1988). The role of knowledge in discourse comprehension: A Construction-Integration model. *Psychological Review*, 95, 163-182.
- MARCOCCIA M., (2004) *On-line polylogues: conversation structure and participation framework in internet newsgroups*, Journals of Pragmatics, 36 (2004) 115-145.
- PEDAUQUE, R.T. (2003). Document : forme, signe et médium, les re-formulations du numérique, *working paper*, version 3- 8 juillet 2003, <http://rtp-doc.enssib.fr>
- PIOLAT A., FAROLI F., & ROUSSEY J.-Y. (1989). La production de texte assistée par ordinateur. In G. Monteil, & M. Fayol (Eds.), *La psychologie scientifique et ses applications* (pp. 177-184). Grenoble : Presses Universitaires de Grenoble
- ROUSSELOT, F., FRATH, P., ET OUESLATI, R. (1996), *Extracting concepts and relations from Corpora*. In Proceedings of the Workshop on Corpus-oriented Semantic Analysis, European Conference on Artificial Intelligence, ECAI 96, Budapest, 12 August 1996.
- SUMNER T., BUCKINGHAM SHUM S., WRIGHT M., BONNARDEL N. , PIOLAT A. & CHEVALIER A. (2000). Redesigning the peer review process : A developmental theory-in-action. In R. DIENG, A. GIBOIN, G. DE MICHELIS & L. KARSENTY (EDS.), *Designing cooperative systems: The use of theories and models* (pp. 19-34). Amsterdam : I.O.S. Press
- TCHOUNKINE P. (2002-a). Pour une ingénierie des Environnements Informatiques pour l'Apprentissage Humain. *Revue I3 Information-Interaction-Intelligence*. Vol. 2, n°1, Cepadues Editions.
- TCHOUNKINE P., (2002-b), Conception des environnements informatiques d'apprentissage : mieux articuler informatique et sciences humaines et sociales, in Baron G.L., Bruillard E. (eds.), *Les technologies en éducation : Perspectives de recherche et questions vives*, p. 203-210, Paris : INRP - MSH - IUFM de Basse Normandie.
- WEICK K.E., (1979) *The Social Psychology of organizing*, New York, Random House.
- ZACKLAD M., LEWKOWICZ M., BOUJUT J-F., DARSEES F., ET DETIENNE F. (2003), *Formes et gestion des annotations numériques collectives en ingénierie collaborative*, actes des journées Ingénierie des Connaissances 2003, Laval.
- ZANNOT (2003), <http://www.zope.org/Members/Crouton/ZAnnot/>

Un traitement sémantique par ontologie pour l'indexation de documents dans un référentiel métier

Wilfried Njomgue Sado^{1,2}, Dominique Fontaine¹

¹ UMR CNRS 6599 Heudiasyc, Université Technologie de Compiègne,
BP 20529, F-60205 Compiègne

{wilfried.njomgue-sado, dominique.fontaine}@hds.utc.fr

² Suez Environnement CIRSEE Pôle Informatique Métier
38, rue du Président Wilson, F-78230 Le Pecq

Résumé : Cet article présente et évalue une approche sémantique qui a été greffée sur une approche originellement linguistico-statistique pour l'indexation de document. Elle combine en amont l'annotation sémantique du document à indexer via l'utilisation d'une ontologie de domaine, l'analyse linguistique du document et enfin l'analyse statistique par la décomposition en valeurs singulières des mots composant le document. Le système d'indexation qui a été développé sur cette base a pour tâche d'affecter tout nouveau document aux activités d'un référentiel métier préexistant. Nous en présentons les résultats, obtenus lors des diverses expérimentations menées sur un corpus de documents propre à la société Suez-Environnement.

Mots-clés : indexation, ontologie, représentation, gestion des connaissances

1 Introduction

Les spécialistes de la documentation assurent le stockage et la diffusion de l'information initialement fixée sur différents supports. Or, ces opérations exigent au préalable un traitement intellectuel des documents à savoir l'indexation. L'indexation en recherche d'information ne couvre pas seulement les aspects d'accès aux données mais aussi la représentation sous forme réduite d'un document par rapport à sa structure et à son contenu sémantique. On parle alors d'indexation par le contenu.

Cette problématique est celle du projet de gestion des connaissances, qui est initié par la Direction Technique et de Recherche de Suez-Environnement, et qui concerne tous ceux qui gèrent le patrimoine textuel du groupe. L'objectif général est d'accroître la valeur d'usage de l'Intranet du groupe qui est un outil clé de stockage, de partage et de diffusion d'information au sein de l'entreprise. Il est un réservoir de nombreuses connaissances (expérience et savoir-faire), généralement proposées sous forme de rapports et de notes techniques. Il favorise les échanges d'expériences entre les exploitants et des utilisateurs répartis sur l'ensemble des continents, en mettant à leur disposition des connaissances utiles à la réalisation de leurs activités.

Ce projet vise en particulier à faciliter l'accès aux documents qui supportent ces connaissances. Pour ce faire, un référentiel métier, une importante taxonomie qui décrit l'ensemble des activités ou métiers de l'entreprise, a été élaboré par des experts de l'entreprise. Il permet à l'utilisateur de sélectionner les documents qui l'intéressent. Il faut donc au préalable que ces documents aient été affectés judicieusement. Ce travail incombait jusqu'alors à des intervenants humains qui indexaient les documents de façon manuelle.

La masse de documents à introduire étant considérable, il a été entrepris d'automatiser ou plutôt de semi-automatiser le processus d'indexation. Nous avons alors conçu un système fondé sur une approche à la fois linguistique et statistique, qui assure l'affectation des nouveaux documents, après validation de l'utilisateur (Njomgue et Fontaine, 2004b). Nous avons ensuite fait l'hypothèse qu'il serait possible d'améliorer les résultats en adoptant au préalable une approche à caractère sémantique, objet principal de cet article.

Après une présentation du problème tel qu'il nous a été posé à l'origine, nous rappelons quelques caractéristiques de l'approche linguistico-statistique qui a d'abord soutenu notre système d'indexation. Ensuite, nous présentons les principes de l'approche à caractère sémantique qui, dans la version remaniée du système d'indexation, va finalement précéder le traitement linguistico-statistique. Celle-ci repose principalement sur la construction et l'utilisation d'une ontologie du domaine qu'il a fallu adapter à nos besoins spécifiques d'indexation. Nous faisons alors une comparaison des résultats obtenus avec ou sans traitement sémantique, sur une même collection de documents. Enfin, nous concluons sur quelques perspectives.

2 Problématique et processus

L'élément déterminant dont la présence influence et conditionne la totalité de notre démarche est le référentiel métier dont la connaissance nous est donnée *a priori*. Ce référentiel est actuellement une arborescence dont les feuilles sont les activités élémentaires du groupe dans le domaine de l'eau. L'auteur _ nom donné à celui qui introduit un nouveau document numérisé _ s'efforce de le classer en parcourant cette arborescence et en identifiant les activités qui selon lui le caractérisent au plus près. Cette tâche s'avère fort fastidieuse : en effet, l'auteur est d'abord censé connaître la plupart des activités de l'entreprise, hypothèse fort risquée, puis doit élaborer sa propre représentation du document, et enfin choisir certains métiers du référentiel parmi la multitude des possibilités. Il est donc essentiel de l'aider dans cette tâche en l'automatisant partiellement ou totalement, et si possible de réduire le temps nécessaire à son accomplissement.

Les particularités et les contraintes de la problématique sont alors les suivantes :

- l'indexation du document est faite relativement aux activités de l'entreprise, et non relativement aux mots du document. Le référentiel est donc à la fois une contrainte dans la mesure où il nous faut se conformer à un usage, à des pratiques, et aussi un support d'informations susceptibles d'orienter et d'aider l'indexation.

- nous ne sommes pas responsables de l'intégrité et de la pertinence du référentiel qui comporte des relations entre concepts dont la sémantique est pour le moins variable voire parfois indiscernable. En outre, il ne nous est pas permis de modifier cette structure parce qu'elle résulte d'un grand effort fait par la compagnie pour clarifier ses activités.
- le système doit être intégré dans un environnement Notes : cette intégration n'a pas été sans incidence sur certains choix techniques.
- nous sommes en mesure d'évaluer systématiquement les résultats produits par le système. En effet, nous les comparons à ceux fournis manuellement par l'auteur du document. Nous faisons l'hypothèse que les propositions faites par l'auteur sont pertinentes et donc qu'elles ne sont pas à remettre en cause. Cette contrainte est extrêmement forte car la diversité des auteurs fait qu'ils n'ont pas toujours la même compréhension du référentiel. Cet élément accroît la nécessité de concevoir un système semi-automatique, la décision revenant finalement à l'auteur.

On sait que l'indexation doit détecter les concepts caractérisant le mieux le document. Cette tâche est difficilement automatisable dans son intégralité, la plupart des systèmes n'indexant pas de façon totalement autonome les textes numérisés. L'indexation devient alors semi-automatique, l'intervention humaine étant garante de la qualité des résultats. C'est également le cas du système présenté, qui après examen du nouveau document propose une liste ordonnée d'affectations possibles à des feuilles du référentiel métier. L'auteur doit alors opter pour certaines d'entre elles.

Les systèmes d'indexation conçus ces dernières années empruntent des approches diverses : linguistique, statistique, plus rarement sémantique. En particulier, les méthodes mixtes apparaissent complémentaires et connaissent un succès notable, au vu des résultats positifs qu'elles fournissent : la meilleure d'entre elles semble être l'approche linguistico-statistique (LS) que nous avons aussi retenue pour notre processus d'indexation. Une première version de notre système a alors été élaborée par application d'une méthode LS.

Au fil des nombreuses expérimentations que nous avons menées, les résultats obtenus, répondant majoritairement à nos attentes (Njomgue et Fontaine, 2004a), ont certes permis de vérifier le bien-fondé de cette approche, mais ont aussi parfois révélé, ponctuellement, des insuffisances fâcheuses en termes de classification. Après analyse des cas considérés comme non conformes, nous avons estimé que la cause de ces défaillances résidait essentiellement dans l'absence de traitement ou de pré-traitement sémantique (Njomgue et Fontaine, 2004a), et même que, compte-tenu de l'existence du référentiel, la prise en compte d'informations à caractère sémantique s'imposait.

On remarquera à cet égard que la démarche sémantique sur la plupart des systèmes d'indexation se résume à un regroupement morphologique et/ou synonymique des termes clés extraits lors de la phase linguistique. Nous avons voulu aller plus loin dans l'analyse et le traitement sémantique des documents à indexer. Nous avons alors émis l'hypothèse que l'utilisation adéquate d'une ontologie du domaine, spécifiquement conçue pour l'indexation, pourrait être la clef de l'amélioration du système existant.

En définitive, le processus d'indexation de la version modifiée, considéré dans sa globalité, comporte plusieurs phases où s'enchaînent des traitements linguistiques, statistiques (Njomgue et Fontaine, 2004b) et donc plus récemment sémantiques. Le schéma descriptif de ce processus, commenté dans les *sections 3 et 4*, est le suivant :

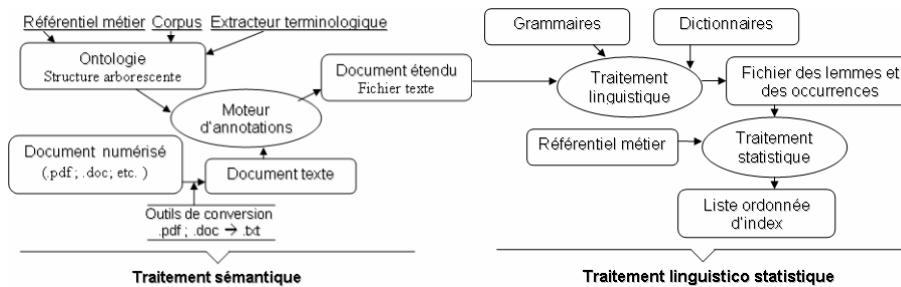


Fig. 1 – Schéma descriptif du processus d'indexation

On remarquera que, parmi les différentes options envisageables, notre choix fut de commencer par un traitement sémantique, dans le but de modifier le document afin d'en obtenir une représentation plus significative des concepts qui le caractérisent. La version résultante est alors soumise au traitement linguistico-statistique, qui pour ne pas biaiser la comparaison des méthodes a été strictement laissé dans la configuration de la version précédente du système.

Il va de soi que nous aurions pu faire un choix différent, par exemple celui qui aurait consisté à maintenir en premier le traitement linguistique et statistique pour seulement ensuite appliquer un traitement sémantique, cette fois-ci non plus dans le but de corriger le document, mais plutôt directement les résultats de l'indexation.

Le processus d'indexation linguistico-statistique ayant déjà été présenté par ailleurs (Njomgue et Fontaine, 2004b), nous en faisons une description simplifiée, puis nous décrivons les principaux aspects de la méthode de traitement sémantique, et en particulier ceux de l'ontologie sur laquelle elle s'appuie. On notera qu'à chaque phase de ce processus nous avons adopté une combinaison de différentes techniques, associées en parallèle ou en séquence. Les choix ont alors été fondé soit, *a priori*, sur des caractéristiques connues de ces techniques, soit, *a posteriori*, sur des résultats d'expériences.

3 Le traitement linguistico-statistique (LS)

Le traitement LS vise *in fine* à établir une correspondance entre le document d'origine et les activités du référentiel métier : il s'agit donc doublement de pointer les concepts représentatifs du document, (i.e.) d'accomplir une tâche d'indexation, et d'associer à cette forme indexée du document des feuilles du référentiel métier, (i.e.) d'effectuer une tâche de classification.

Le *document texte*, dénué d'images, vidéo, etc. est constitué uniquement des composantes textuelles du *document numérisé* initial. Le *document étendu* est obtenu par enrichissement du *document texte*. Le traitement L.S. débute par une analyse linguistique du *document étendu*, qui consiste à extraire les termes composant le document, puis s'achève par une analyse statistique qui vise à mettre en valeur les termes importants.

L'analyse linguistique comprend séquentiellement des analyses morphologique et syntaxique, qui extraient et éventuellement regroupent les mots en fonction respectivement de leurs formes et de la syntaxe des phrases. Elle comprend également une étape dont la sémantique n'est pas totalement exclue, à savoir un regroupement synonymique autour termes clés du référentiel. Dans cette phase linguistique, sont notamment mises à contribution des grammaires locales, des dictionnaires spécialisés et des listes de synonymes relatifs aux termes du référentiel métier. Des techniques de lemmatisation (réduction automatique des mots à une forme de surface canonique), de stemming (réduction des formes de surfaces similaires à un seul concept), de stop-list (élimination de mots non pertinents pour l'indexation), en rapport avec le référentiel métier, sont mises en oeuvre.

L'analyse statistique qui est menée presuppose qu'il existe d'une part une relation entre la fréquence d'un terme dans un document et son importance pour ledit document, et d'autre part un lien entre l'importance d'un terme et le nombre de documents du corpus qui le contiennent. Elle discrimine les mots extraits à l'issue de la phase linguistique selon leurs occurrences et leurs co-occurrences, puis estime la proximité entre le document et les thèmes du référentiel.

A cet effet, de nombreuses techniques sont mises à contribution, parmi lesquelles, le latent semantic indexing ou LSI, la pondération des termes, en fonction de leur *contexte local* _par rapport au document _, de leur *contexte global* _par rapport à la base de données_ et de leur *contexte positionnel* _par rapport aux autres termes (leurs co-occurrences). Le processus statistique mis en oeuvre ici est une combinaison de diverses méthodes statistiques contribuant à l'indexation des documents, combinaison valable sous l'hypothèse qu'il est possible et pertinent d'associer les avantages des unes et des autres.

4 Traitement sémantique

4.1 Approche par enrichissement du document

Le système d'indexation doté d'une approche IS donne des résultats souvent satisfaisants (Njomgue et Fontaine, 2004b) comme nous le rappelons dans la section 3, mais fournit aussi des résultats parfois inadéquats, lorsqu'on les compare à ceux proposés par les auteurs des documents sur lesquels les expérimentations ont été conduites. Ces insuffisances s'expriment en termes de *silence*, lorsque les propositions d'affectation d'un document, émises par le système, ignorent les propositions privilégiées par l'auteur dudit document, ou de *bruit*, lorsque les

propositions d'affectations d'un document accordent une place prioritaire à des affectations que l'expert a lui-même jugées impropre.

L'analyse des cas problématiques nous a permis de discerner essentiellement trois situations qui en sont à l'origine : la présence de mots ou d'expressions ambigus, l'absence ou la sous représentation de certains mots ou de certaines expressions pourtant jugés importants, et *a contrario* la surreprésentation et donc la survalorisation de certains d'entre eux pourtant jugés peu pertinents. Ces situations évoquées ne peuvent alors qu'induire des distorsions sur l'ensemble des affectations proposées. Considérons deux exemples :

- supposons que le terme "*réseau*" apparaisse dans le document avec une forte occurrence, sans qu'il soit davantage qualifié explicitement. Il ne sera alors pas possible de discriminer entre des activités liées au "*réseau de distribution*", au "*réseau informatique*" ou encore au "*réseau des eaux usées*", thèmes appartenant au référentiel métier. Le terme "*réseau*" devient alors un index à la fois important et ambigu. Seule la prise en compte du contexte va permettre de lever cette ambiguïté, par exemple par la présence réitérée des expressions "*eau pluviale*", "*eau parasite*" ou "*eau industrielle*" qui permettent sans guère de doute de rattacher le document aux activités liées au "*réseau des eaux usées*".
- l'activité dénommée "*analyse des substances organoleptiques*", explicitement décrite dans le référentiel, est importante dans le domaine de l'eau, et concerne beaucoup de documents à caractère technique, au point d'en devenir le sujet essentiel. Or cette expression est rarement présente sous cette forme ou sous une forme synonymique. En revanche, la présence de paramètres tels que le "*goût*", "*l'odeur*", "*le charbon actif*", "*la couleur*", "*la saveur*", etc. évoque irrésistiblement pour les spécialistes l'activité en question. Là encore, la prise en compte appropriée du contexte devrait permettre de revaloriser l'expression terme "*substance organoleptique*" et donc d'orienter l'indexation dans ce sens.

Notre problématique est alors la suivante : comment prendre en compte l'ensemble du document, à la fois les termes du référentiel explicitement décrits, mais aussi ceux qui ne sont que suggérés, connotés, déductibles ou obtenus par associations d'idées ? Comment introduire peu ou prou le contexte, lié au domaine, et en particulier les informations communes aux auteurs, absentes du document et pourtant décisives lors de la discrimination ? et enfin comment apporter un surcroît d'efficacité à l'approche LS dont on sait qu'elle accorde une place prééminente à la fréquence des occurrences ou des co-occurrences de mots ? L'approche que nous avons retenue repose sur deux principes :

- l'enrichissement du document par *l'adjonction* de mots ou d'expressions jugés, et ce dans le but de faire apparaître les concepts jusqu'alors manquants en ajustant leur fréquence d'apparition à la mesure de leur importance.
- et pour ce faire l'utilisation d'une *ontologie* du domaine dont on attend qu'elle nous apporte certains des liens conceptuels originellement absents

des documents, et en particulier ceux qui font référence aux concepts du référentiel métier.

En somme, le traitement sémantique, précédé d'un prétraitement linguistique minimal (conversion, mise au format texte, lemmatisation), consiste ici, par une *méthode d'enrichissement*, à refléter *quantitativement* l'importance *sémantique* et *qualitative* des concepts.

4.2 Edification de l'ontologie

En Ingénierie des Connaissances, les ontologies sont des conceptualisations d'un domaine (Gruber, 95 ; Fernandez Lopez, 1999; Asucion Gomez-Perez, 1999), pour nous ici celui des métiers de l'eau et en particulier de ceux portant sur l'analyse et le traitement de l'eau. Malgré leur diversité, un invariant est qu'elles contiennent des concepts et la spécification des relations entre ces concepts.

Les ontologies sont en général définies pour un objectif donné. Il en est ainsi pour l'ontologie utilisée par notre système puisqu'il s'agit finalement d'aider à l'indexation de documents. En conséquence, cette ontologie n'a en aucun cas un caractère générique, et sa réutilisation est largement dépendante du type d'application qui y fait appel.

Diverses approches ont été proposées pour édifier des ontologies (Aussenac-Gilles et al. 2000 ; Kassel G., 2002). N'ayant pas à disposition un unique expert pour l'ensemble des activités, nous avons retenu une approche de construction de l'ontologie *à partir de textes* en suivant le processus présenté en figure 2. Les phases ont été les suivantes :

- *la phase de constitution du corpus* est importante car elle est à la base de cette approche. Des experts nous ont fourni un panel de documents techniques représentatifs du domaine : notre *corpus numérisé*.

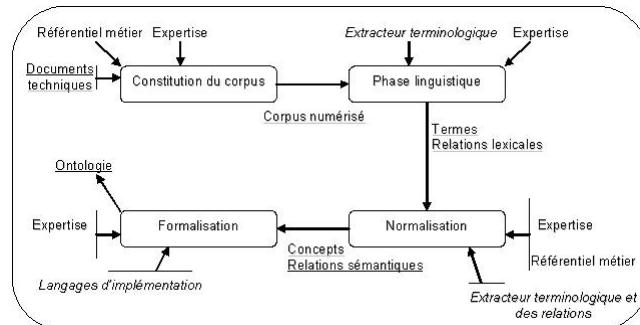


Fig. 2 – Edification d'une ontologie à partir des textes

- *l'étude linguistique* sur le corpus a été menée à l'aide d'extracteurs terminologiques, ici Intex (Silberstein, 2001) et Syntex (Bourigault, 2002). Nous avons notamment défini des grammaires de reconnaissance de certaines relations (synonymie, hyperonymie, hyponymie, etc.) (Aussenac-Gilles et Séguéla, 2000 ; Hamon et al. 1999) afin non seulement d'extraire les mots, les relations et de les désambiguïser, mais aussi de visualiser les relations dans

les contextes où elles apparaissent. Par ailleurs, nous nous sommes appuyés sur le référentiel métier (le corpus du métier), les critères de validation des concepts à la sortie des extracteurs reposant essentiellement sur l'importance que leur octroie le référentiel.

- *la normalisation* a consisté à interpréter sémantiquement et à structurer les termes à travers notre connaissance du métier, du corpus et des regroupements résultant de l'étude linguistique. Nous avons alors obtenu des concepts et des relations sémantiques, en attente d'une validation experte.
- *la formalisation* est la phase d'implémentation, d'élaboration et de validation de l'ontologie (Barry et al. 2001). Nous avons adopté une approche centrifuge (Gandon et Dieng, 2001 ; Faure et Poibeau, 2000) en identifiant les concepts clés du référentiel métier et en complétant l'ontologie par généralisation ou spécification des relations et concepts extraits du corpus. La dernière phase a consisté à implémenter l'ontologie avec Java, en raison en particulier de sa compatibilité avec l'environnement Notes de Suez.

4.3 Utilisation de l'ontologie pour l'indexation

4.3.1 Représentation de l'ontologie

La structure sémantique de l'ontologie du système d'indexation est composée de :

- *concepts* : ce sont des mots abstraits tel "*distribution*", ou concrets tel "*eau pluviale*", élémentaires ou composés. Nous aboutissons à près de 1200 concepts clés du domaine de l'eau (Lyonnaise des Eaux, 1994).
- *relations* qui représentent un type d'interaction entre les concepts du domaine (Guarino et al. 2001 ; Kassel, 2002 ; Studer et al. 1998 ; Zweigenbaum et Grabar, 2000). La version actuelle en comporte près de 1500.

Les relations binaires utilisées entre concepts, largement majoritaires, sont de types subsumption ("*l'altrazine*" est_un "*pesticide*"), synonymie, méronymie, et lien morphologique ("*chloration*" a pour morphème "*chlore*").

D'autres sont éventuellement n-aires ($n \geq 2$) : ce sont les relations de causalité ("*traitement des eaux usées*" cause "*boue*"), évocation ("*l'indice de molhman*" évoque "*épaississement des boues*"). En outre, on remarquera que ces relations sont souvent *incertaines*, ce dont on tiendra compte, et que les relations d'évocation sont fortement liées à la tâche d'indexation.

Pour représenter ces différentes et nombreuses relations, nous avons opté pour la représentation unifiée par règles de production :

- *typées* de la forme $P_1 \text{ et } \dots \text{ et } P_k \rightarrow C$, de prémisses P_i , de conclusion C , où la flèche est étiquetée par le type de la relation,
- et *pondérées* par un poids p accordé en fonction de la certitude accordée à la relation ontologique, égal à 1 si la relation est certaine sinon $0 < p < 1$.

Cette représentation, simple et largement éprouvée, nous a notamment permis de regrouper les règles en fonction de leurs types. De plus, les algorithmes développés pour les systèmes de production répondaient à nos besoins : pour enrichir automatiquement le document à indexer, il nous restait alors à les transformer en un *moteur d'annotations* de document.

4.3.2 Enrichissement du document

Cette phase de traitement sémantique par enrichissement du document, désormais partie intégrante du processus d'indexation sémantico linguistico statistique que nous noterons S-L-S, vise à mieux identifier le sujet du document analysé, par l'adjonction de concepts clés du domaine de l'eau.

Un *moteur d'annotations* à caractère déductif est dédié à cette tâche. Il ne considère *dans la version actuelle* que des fenêtres d'analyse circonscrites aux phrases. En effet, le but du traitement étant d'indexer le document en référence à une ou plusieurs idées fédératrices, nous avons estimé judicieux de focaliser l'attention sur les phrases : en effet, la circonscription à des phrases révèle des liens entre mots plus forts et plus chargés de sens que ne le révélerait une recherche d'occurrences simultanées étendue à l'ensemble du document.

Les principes d'enrichissement sont alors inspirés de ceux qui régissent le comportement des moteurs d'inférences, pour les systèmes à base de règles de première génération. En balayant le document à indexer, le moteur d'annotations déniche les prémisses des différentes relations qui sont simultanément présentes au sein d'une même phrase. Lorsque tel est le cas, la conclusion des relations concernées est ajoutée au document. Il se peut bien sûr que le concept ajouté soit lui aussi prémissse d'autres relations ontologiques. Il importe donc de réitérer le cycle précédent, et ce jusqu'à épuisement des possibilités d'ajout. A titre d'exemple, *le "chlore" est un "désinfectant"* ; *le "désinfectant" a un lien morphologique avec la "désinfection"* ; *la "désinfection" est un "traitement de finition dans le traitement des eaux usées"*. Ainsi, si "chlore" est un concept du document, "désinfectant" et "traitement de finition" seront ajoutés au document.

Parmi de nombreux choix possibles, le traitement de l'incertitude pour lequel nous avons opté, au départ à titre expérimental, est issu des méthodes à coefficients de vraisemblance en vigueur dans différents systèmes de première génération. Sans en détailler ici le mécanisme, il associe un principe de propagation de l'incertitude au sein d'un réseau de règles à un principe de renforcement de l'incertitude. En effet, si plusieurs relations permettent d'évoquer la présence d'un même concept chacune avec son propre facteur d'incertitude, il est légitime d'associer ces divers facteurs pour conclure à une pertinence accrue de ce concept. En définitive, le critère retenu pour l'inscription définitive d'un nouveau concept dans le document est qu'à l'issue de ce traitement celui-ci soit affecté d'un facteur de vraisemblance supérieur ou égal à 0.7, seuil qui expérience faite s'est révélé satisfaisant.

Par ailleurs, il est apparu que ce principe d'expansion du document par ajouts successifs de concepts clés devait être maîtrisé. En effet, si un concept apparaît n fois dans un document, alors la conclusion de la relation dont il est une prémissse sera

insérée n fois dans le cas d'une relation certaine, ces insertions étant répercutées sur ses descendants par propagation. Compte-tenu de l'objectif qui est de mieux prendre en compte les concepts à la fois importants et sous-représentés en augmentant leur fréquence d'apparition, cet effet multiplicateur était bien souhaité, mais en même temps il a fallu le contenir, sous peine de dérapage potentiel.

5 Expérimentations

Pour mener notre étude, assurer le développement du système d'indexation et enfin effectuer les expérimentations nécessaires au choix progressifs des méthodes et à la validation du système, nous avons disposé assez rapidement d'une part de documents issus de la société Suez-Environnement, écrits en français courant et souvent à caractère technique, et d'autre part des affectations au référentiel proposées par les auteurs de ces documents. Au final, nous avons effectué nos expérimentations sur un panel de près de 450 documents. L'hypothèse sous-jacente aux évaluations fut bien entendu que ces affectations de référence étaient pertinentes, même si l'analyse du contenu de certains documents nous a parfois laissé perplexes.

Généralement, les auteurs suggèrent manuellement au plus 10 activités pour un document. Le système d'indexation réduit et propose au maximum 15 activités parmi les 165 possibles, classées par ordre de pertinence, et parmi lesquelles en dernier lieu l'auteur doit choisir.

Nous avons constaté que la moyenne de mots utiles dans un document est d'environ 700 mots. Pour estimer la pertinence des indexations et ici des affectations, entre autres critères, il est usuel d'évaluer les performances par le rappel et la précision qui désignent respectivement le nombre de documents pertinents retournés par rapport au nombre total de documents pertinents et le nombre de documents pertinents retournés sur le nombre total de documents retournés.. De plus, compte tenu du caractère semi-automatique du système, il est primordial de limiter autant que faire se peut le silence, et il est plus pertinent d'évaluer le système par le rappel que par la précision. A l'issue de ces expériences, les résultats sont les suivants (Table 1) :

Table 1. Résultats des expérimentations

		Sans traitement sémantique		Avec traitement sémantique	
	Index (auteur)	Mots outils	Rappel index (système)	Mots outils	Rappel index (système)
Minimum	1	35	0%	73	0%
Maximum	10	1537	100%	1603	100%
Moyenne	4	700	77.09%	800	87%

A ce stade, et notamment au vu des réactions des premiers utilisateurs de l'entreprise, ces résultats sont apparus conformes à nos espérances : pour la méthode S-L-S, un taux de rappel fort satisfaisant, et sur les parties où l'ontologie a été validée,

aucun cas n'a donné lieu à des réponses erratiques. En outre, et tel était l'objet de cet article, il ont permis de confirmer l'intérêt qu'il y avait à ajouter un traitement sémantique au système, et à le faire par référence à une ontologie. L'adjonction de ce traitement a permis en particulier de diminuer très sensiblement le silence, ce qui était notre première préoccupation, et incidemment et indirectement de réduire le bruit en reléguant plus loin dans la liste les propositions inappropriées.

6 Conclusion et perspectives

Dans une première étape, un traitement sémantique du document permet d'annoter, d'ajouter du sens aux documents par rapport au référentiel métier du groupe et au domaine de l'eau. La méthode présentée s'appuie sur une ontologie du domaine qui a été édifiée en vue de servir à l'indexation. Ensuite, le traitement linguistique représente le document à indexer par un ensemble de concepts jugés lexicalement significatifs. En dernière étape, un traitement statistique discrimine ces concepts en considérant leurs occurrences et leurs co-occurrences, puis estime la proximité entre le document et les activités cibles du référentiel.

Globalement, les résultats obtenus sont à considérer en rapport avec la difficulté de la tâche : un référentiel assez hétérogène et imposé, une grande diversité de documents, une ontologie non stabilisée sur certains aspects faute d'expert désigné, et un jeu de tests dont la pertinence ne semblait pas toujours indiscutable.

L'évaluation comparative des méthodes, linguistico-statistique (LS) seule ou précédée d'un traitement sémantique (S-L-S), confirme le bien fondé de l'apport de la sémantique dans le domaine de l'indexation des documents à forte composante textuelle, et plaide en faveur de l'utilisation d'ontologies du domaine pour l'indexation.

Cette évaluation fait en même temps apparaître quelques lacunes en terme de précision, surtout compte tenu des exigences de qualité qu'ont légitimement les futurs utilisateurs de ce système. Nous nous efforçons de pallier ces insuffisances, notamment en affinant les mécanismes d'enrichissement des documents, par exemple par un meilleur contrôle de l'effet multiplicateur ou une meilleure utilisation de la typologie des relations. Il s'agit là d'une problématique de recherche qui nous semble porteuse de perspectives intéressantes. Nous pensons également que la prise en compte de la structure des documents permettra d'effectuer l'enrichissement avec plus de sélectivité.

Par ailleurs, nous cherchons à stabiliser l'ontologie du domaine, et à gommer les points faibles. Les experts et auteurs sont aujourd'hui sollicités dans cette phase d'amélioration et au-delà d'exploitation de l'ontologie. Cet effort doit aboutir à terme à un système intégré dans un environnement Intranet.

Références

ASUCION GOMEZ-PEREZ. (1999). Développements récents en matière de conception, de maintenance et d'utilisation d'ontologies. Revue n°19 : Terminologie et Intelligence Artificielle

(Actes du colloque de Nantes, 10-11 Mai 1999) de RINT : Réseau International de Néologie et de Terminologie.

AUSSENAC-GILLES N. , BIEBOW B. ET SZULMAN S. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In Ingénierie des Connaissances 2000

AUSSENAC-GILLES N. ET SEGUELA P. (2000). Les relations sémantiques : du linguistique au formel.

BARRY C., CORMIER C., KASSEL G. ET NOBECOURT J. (2001). Evaluation de langages opérationnels de représentations d'ontologies. Actes de la conférence IC'2001. Pages 309-327. Grenoble. France

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy, 2002, pp. 75-84

GANDON F., DIENG R. (2001). Ontologie pour un système multi-agents dédié à la mémoire d'entreprise. Actes de la conférence IC'2001, Pages 1-20. Grenoble. France

FAURE, D. POIBEAU T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In Staab, S. and Maedche, A. and Nedellec, C. and Wiemer-Hastings P., editors, Ontology Learning ECAI-2000 Workshop, pages 7-12.

GUARINO N., GANGEMI A., MASOLO C. ET OLTRAMARI A. (2001). Understanding top-level ontological distinctions. Workshop on Basic Ontological Issues in Knowledge Sharing. IJCAI'95.

GRUBER T. R. (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Of Human Computer Studies. 43 : 907-928. Stanford University

HAMON T., GARCIA D., NAZARENKO A. (1999). Détection de liens de synonymie : complémentarité des ressources générales et spécialisées. Revue n°19 : Terminologie et Intelligence Artificielle (Actes du colloque de Nantes, 10-11 Mai 1999) de RINT : Réseau International de Néologie et de Terminologie.

KASSEL G. (2002). OntoSpec : une méthode de spécification semi-formelle d'ontologies. Actes des 13ièmes journées francophones d'Ingénierie des Connaissances, IC' 2002. Pages 75-87. France

LYONNAISE DES EAUX (1994). Mémento du gestion de l'alimentation en eau et de l'assainissement. Tome 1, 2 et 3.

NJOMGUE W., FONTAINE D. (2004A). A linguistic and statistical approach for extracting knowledge from documents. TAKMA 2004 (Fifth International Workshop on Theory and Applications of Knowledge Management) in Conjunction with DEXA 2004 (Database and Expert Systems Applications); Spain

NJOMGUE W., FONTAINE D. (2004B). Identification des thèmes d'un document relativement à un référentiel métier. In Manifestation de JEunes Chercheurs Sciences et Technologies de l'Information et de la Communication. France

SILBERZTEIN, M. (2001). Intex @ manual ASSTRIL - LADL, 201p, 2000-2001.

STUDER R., BENJAMINS V. R., FENSEL D. (1998). Knowledge Engineering : Principles and Methods. Data & Knowledge Engineering. 25 : 161-197.

FERNANDEZ LOPEZ, M. (1999). Overview of Methodologies for Building Ontologies, The Proceedings of the IJCAI-99 Workshop on Ontologies and Problem Solving Methods, Sweden August-1999.

ZWEIGENBAUM P., GRABAR N. (2000). Liens morphologiques et structuration de terminologie", in IC 2000 : Ingénierie des connaissances, pp. 325-334.

Intégration de multiples ontologies en anatomie pathologique

David Ouagne¹, Christel Daniel-Le Bozec¹, Eric Zapletal¹, Maxime Thieu¹,
Marie-Christine Jaulent¹

¹INSERM, U729, Paris, F-75006 France,
{David.Ouagne, Christel.Lebozec, Eric.Zapletal, Maxime.Thieu,
Marie-Christine.Jaulent}@spim.jussieu.fr

Résumé : La variabilité diagnostique en anatomie pathologique est en partie liée à l'utilisation de systèmes de classification différents, pouvant être considérés comme des points de vue différents, pour décrire des lésions. Notre objectif est de représenter ces points de vue et de proposer une solution pour permettre leur interopérabilité. L'approche hybride décrite par Wache nous permet de développer un système multi ontologique en trois étapes 1) la représentation des points de vue au sein d'ontologies locales, 2) la construction d'un vocabulaire partagé et 3) le développement d'un outil de traduction. L'évaluation du travail, conduite sur 33 cas, a consisté à évaluer les ontologies locales grâce à un outil de validation sémantique de cas et à évaluer l'outil de traduction. Nos résultats montrent que les pathologistes produisent des descriptions qui ne suivent pas toujours les règles d'interprétation des systèmes de classification auxquels ils se réfèrent. Si 62.5% à 100% des concepts des ontologies locales sont traduisibles, nous avons constaté que la validité des cas n'était pas toujours conservée après traduction.

Mots-clés : Représentation des connaissances, interopérabilité sémantique, alignement d'ontologies, imagerie médicale, cancer du sein

1 Introduction

L'examen anatomopathologique permet d'établir un diagnostic et de donner des indications pronostiques concernant des lésions observées au niveau d'images de tissus ou cellules issues de prélèvements. Les règles qui permettent d'établir une conclusion diagnostique et/ou pronostique à partir de caractéristiques morphologiques observées dans les images créées au cours de l'examen sont publiées dans le cadre de systèmes de classification. Il existe en anatomie pathologique une variabilité diagnostique largement rapportée dans la littérature (Fleming, 1996). Celle-ci est en partie liée au fait que les systèmes de classification sont nombreux et évolutifs. Ainsi, des cas identiques peuvent conduire à des conclusions diagnostiques différentes en fonction du système de classification sur lequel repose le point de vue du pathologue (Wells, 2000).

Le système IDEM (Images et Diagnostics par l'Exemple en Médecine) a pour objectif d'assister les experts dans la constitution de descriptions consensuelles de

cas anatomo-pathologiques. Le projet est développé dans notre laboratoire depuis quelques années et propose un environnement dans lequel des descriptions de cas anatomo-pathologiques peuvent être comparées (Thieu 2004). La plateforme actuelle compare des cas exprimés selon un seul point de vue, c'est-à-dire conformément à une classification déterminée à l'avance. Dans l'optique de faire collaborer des experts distants, il est nécessaire aujourd'hui de comparer des cas exprimés selon différents points de vue.

Dans ce contexte particulier, notre objectif est de modéliser les connaissances traduisant les différents points de vue de pathologistes au sein d'ontologies locales afin de pouvoir comparer et rendre interopérables les conclusions fournies par des pathologistes se référant à des systèmes de classification différents.

Nous avons procédé en trois étapes consistant en 1) la représentation des différents points de vue sous forme d'ontologies locales, 2) la réalisation d'un vocabulaire partagé et le développement d'un environnement informatique permettant d'apparier les ontologies locales au vocabulaire partagé et 3) le développement d'un algorithme permettant de traduire la description d'un cas selon un point de vue dans un point de vue différent.

L'environnement informatique a été développé à partir d'un outil de fusion d'ontologies et trois points de vue différents ont été modélisés pour illustrer l'intérêt de cet environnement. Par la suite, une première étude a été réalisée pour l'évaluation de l'appariement entre deux de ces trois points de vue. Trente trois cas de pathologie tumorale mammaire ont été décrits par deux experts selon deux points de vue différents. Les descriptions ont été validées dans le sens où, pour chaque point de vue envisagé, leur conformité aux règles d'interprétation modélisées dans les ontologies locales a été vérifiée. En utilisant l'environnement informatique développé, chaque description a ensuite été traduite selon l'autre point de vue. Une première étude de l'interopérabilité a consisté à comparer la validité des descriptions originelles à celle des descriptions obtenues après traduction. Cette comparaison s'appuie sur la mesure du Kappa (Falissard 2001).

2 État de l'art

La prise en compte du point de vue dans les systèmes à base de connaissances permet d'indexer les connaissances afin de les rendre plus accessibles et réutilisables. Ribièvre analyse les modèles et les formalismes permettant la représentation de multiples points de vue (Ribièvre, 1999).

Parmi ces formalismes, les ontologies impliquent effectivement une certaine vue du monde par rapport à un domaine donné et sont conçues comme un ensemble des concepts d'un domaine – e.g. entités, attributs, processus –, leurs définitions et leurs interrelations (Bachimont, 2000) (Grüber, 1993) (Uschold, 1996). La définition d'une méthodologie de construction d'ontologies a fait l'objet de nombreux travaux. Uschold et Grüninger (Uschold, 1996) proposent une méthode en trois phases successives consistant à : identifier le domaine d'application et le contexte d'utilisation (la tâche) de l'ontologie, puis à construire effectivement l'ontologie,

c'est à dire à acquérir les concepts du domaine et à les coder dans un modèle conceptuel formel et enfin à valider l'ontologie construite.

En ce qui concerne l'intégration de connaissances hétérogènes (issues de points de vue différents), un certain nombre de travaux présentent des solutions qui s'appuient sur une représentation ontologique des connaissances permettant la comparaison des points de vue.

Selon Wache (Wache, 2001), on distingue généralement trois approches pour l'intégration de connaissances hétérogènes. L'approche mono ontologie vise à construire une seule ontologie exprimant l'ensemble des sources de connaissances (Noy, 2003) ; l'approche multi ontologies s'appuie sur des ontologies locales et définit des opérations d'appariement ou de transformation des ontologies locales (Mena, 2000) ; enfin l'approche hybride s'appuie sur des ontologies locales et définit un vocabulaire partagé et des relations – solutions d'appariement et de transformation – entre ce vocabulaire partagé et les ontologies locales (Buccella, 2003).

Des solutions techniques permettant de gérer de multiples ontologies se sont développées. Ces solutions peuvent être classées selon l'opération réalisée (appariement, transformation, gestion de version, etc.) ou selon les techniques de mise en correspondance utilisées (Noy, 2003). Une des difficultés des démarches multi ontologies et « hybride » est la définition de l'appariement (« mapping »). Il existe plusieurs environnements effectuant des opérations d'appariement en se basant sur des agents traducteurs (KRAFT (Preece, 1999)), les logiques de description (OBSERVER (Mena, 2000)), un modèle probabiliste faisant intervenir l'expression en langue naturelle des concepts (GLUE (Doan, 2002), Labo ISI/USC (Hovy, 1998)), des techniques lexicales (co-occurrences de termes dans des corpus) (ONION (Mitra, 2000)). L'environnement à utiliser doit être choisi en fonction de la tâche envisagée (fusion, différence, appariement, transformation, etc.) et des ontologies locales dont on dispose.

Le problème posé de l'intégration de connaissances hétérogènes s'apparente à la comparaison, l'alignement et la fusion d'ontologies. L'ambition du travail présenté n'est pas de comparer formellement les méthodologies d'alignement mais plutôt d'adapter des outils existants au domaine spécifique de l'anatomie pathologique afin d'étudier l'intérêt et la portée de ces approches dans notre domaine. Notre contribution est donc essentiellement applicative bien que par certains aspects, la réflexion puisse être généralisée. Pour des raisons pratiques, liées à l'historique du projet IDEM, le choix d'un éditeur d'ontologie s'est porté sur Protégé.

3 Construction d'un système multi ontologique pour la comparaison des points de vue en anatomie pathologique

Nous avons suivi l'approche hybride décrite par Wache qui se compose de trois étapes, 1) la construction d'ontologies locales, 2) la construction du vocabulaire partagé et 3) la définition des relations entre ce vocabulaire partagé et les ontologies locales (Buccella 2003) (cf. figure 1). Une fois le système multi ontologique construit, nous sommes capables de comparer les concepts des ontologies locales

entre elles, et plus particulièrement dans notre cas, d'exprimer un concept d'une ontologie dans une autre ontologie (opération de traduction).

3.1 Construction des ontologies locales

Nous avons défini un modèle de description des cas anatomopathologiques puis identifié les différents points de vue de pathologistes existant dans ce domaine et nous les avons représentés au sein d'ontologies locales.

- Définition des points de vue

Il existe dans la littérature, un modèle de démarche diagnostique en anatomie pathologique, formalisé par des guides de bonnes pratiques diagnostiques et appelé « anatomie pathologique basée sur le niveau de preuve » (ou « evidence-based pathology »). Selon ce modèle, la démarche diagnostique consiste à reconnaître des anomalies morphologiques au sein d'images et à les interpréter en accord avec des règles diagnostiques publiées résultant de consensus d'experts (Fleming, 2002). Des systèmes de classification diagnostique explicitant ces règles permettent de conclure à un diagnostic à partir de l'analyse d'anomalies morphologiques. Nous nous sommes restreints au domaine de la pathologie tumorale mammaire et en particulier du diagnostic du Carcinome Canalaire In Situ (CCIS) pour lequel les trois systèmes de classification diagnostique principaux sont ceux de Holland, Lagios et Van Nuys (Wells, 2000).

Le système de classification est défini par l'ensemble constitué d'une part des termes associés aux anomalies morphologiques complexes (e.g. « bas grade nucléaire ») ou élémentaires (e.g. « augmentation modérée de la taille nucléaire ») et aux diagnostiques morphologiques (e.g. « CCIS de bas grade »), et d'autre part des règles d'interprétation permettant d'inférer un diagnostic morphologique à partir d'anomalies morphologiques (par exemple, selon Van Nuys, la présence d'un « bas grade nucléaire » ou d'un « grade nucléaire intermédiaire » associé à l'« absence de nécrose » permet de conclure au diagnostic morphologique de « CCIS de bas grade »).

- Construction des ontologies locales

Nous avons utilisé la méthode de construction d'ontologie décrite par Uschold et Grüninger (Uschold, 1996). En pratique, deux experts ont extrait des trois systèmes de classification du CCIS – Holland, Lagios et Van Nuys – les anomalies morphologiques pertinentes par rapport au diagnostic de grade de CCIS.

L'organisation des concepts de l'interprétation des images anatomopathologiques a bénéficié des efforts de standardisation de la discipline ayant abouti à la constitution de ressources terminologiques électroniques telles que la SNOMED CT (Lieberman, 2003). Conformément à la SNOMED CT, nous avons organisé les anomalies morphologiques selon une hiérarchie taxinomique (relations « est un ») par région anatomique observée (par exemple, noyau, cellule, canal, etc.) puis par type de caractéristique morphologique (par exemple, taille, forme, etc). De plus, les anomalies morphologiques complexes telles que « CCIS de haut grade » et « grade nucléaire élevé » ont été mises en relation avec les anomalies morphologiques justifiant ces interprétations selon les systèmes de classification de référence choisis

grâce à la relation « est inféré par ». Les ontologies locales ont été codées en utilisant l'éditeur d'ontologie Protégé (Noy, 2003) (cf. figure 2).

- Description et validation d'un cas à partir d'une ontologie locale

Une ontologie locale permet de décrire un cas selon un système de classification donné en fournissant aux pathologistes les termes standardisés du domaine ainsi que les connaissances nécessaires à la description des cas (règles d'interprétation).

Concrètement, un plug-in Protégé a été développé afin d'afficher le formulaire permettant aux pathologistes de décrire le cas en utilisant les termes désignant les concepts de l'ontologie locale (Van Nuys ou Holland ou Lagios). Ce plug-in produit une description XML d'un cas. Une originalité du travail réalisé a consisté à développer un algorithme permettant de valider les descriptions de cas selon une ontologie locale. La validité d'un cas selon un point de vue se définit comme la conformité de ce cas aux règles d'interprétation exprimées dans l'ontologie locale. On dit alors que le cas est valide ou « prototypique ». L'algorithme de validation exploite les règles d'interprétation explicitées au sein de l'ontologie locale de référence par la relation : « est inféré par ». Cet algorithme utilise la méthode classique de validation syntaxique d'un fichier XML (ici, la description du cas) selon un schéma XML (ici, exprimant les règles d'interprétation) et permet de signaler si, au niveau de la description, les diagnostics morphologiques sont correctement inférés.

3.2 Construction du vocabulaire partagé

La méthode de construction d'un vocabulaire partagé est proche de la méthode de fusion d'ontologies qui consiste à construire une ontologie à partir de plusieurs ontologies sources. La différence, en ce qui concerne la construction d'un vocabulaire partagé, est qu'il s'agit de choisir parmi les concepts appariés des ontologies locales celui qui fournit le terme à placer dans le vocabulaire partagé et surtout de créer un lien entre ce terme (dit terme « préféré ») et chacun des termes désignant les concepts des ontologies locales. Dans le cas de la fusion d'ontologies, les concepts appariés des ontologies locales sont fusionnés en un nouveau concept (au sein d'une nouvelle ontologie) sans que soit conservé le lien entre ce concept et les concepts sources. L'idée est donc de « passer » par la terminologie (le vocabulaire partagé) pour conserver ce lien.

Nous avons par ailleurs fait l'hypothèse, qu'il était possible d'adapter une solution logicielle permettant de faire une fusion d'ontologies afin d'obtenir un vocabulaire partagé pour finalement réaliser un appariement entre les concepts. Nous avons vu que le choix d'un environnement multi ontologique dépendait de la tâche envisagée et des ontologies locales dont on dispose. Dans la mesure où les ontologies locales que nous avons construites ne contiennent pas d'instances et que les relations ont une grande importance, nous avons choisi d'utiliser la suite PROMPT (plug-in Protégé), adaptée à ce type d'ontologies et qui permet, entre autres, de comparer et de fusionner des ontologies. De plus, la suite PROMPT possède une interface de développement sous licence libre permettant de réaliser les adaptations envisagées (Noy, 2003).

Lors de l'analyse de deux ontologies, le mode fusion de la suite PROMPT (iPROMPT) affiche un tableau de suggestions sous forme de propositions d'actions de fusion. Si un même concept semble exister dans les ontologies locales, alors iPROMPT propose une fusion des concepts sources dans la nouvelle ontologie. Une fusion peut, et c'est souvent le cas, entraîner d'autres fusions.

Le mode dit « mapping » que nous avons développé à partir du mode fusion pour construire notre vocabulaire partagé s'appuie sur la première étape de suggestion de fusion puis permet de choisir un terme « préféré », placé dans le vocabulaire partagé, et de créer une relation entre ce terme « préféré » et le ou les termes « apparié(s) » désignant les concepts sources suggérés pour la fusion. La relation peut être la relation « *identique* » lorsque le terme « apparié » est celui qui a été choisi comme terme « préféré ». La relation peut-être la relation « *synonyme* » lorsque le concept désigné par le terme « préféré » est identique au concept désigné par le terme « apparié » mais que les termes « préféré » et « apparié » sont synonymes. Enfin, pour augmenter le nombre de concepts traduisibles, la relation peut-être aussi la relation « *similaire* » lorsque le concept désigné par le terme « préféré » est jugé similaire au concept désigné par le terme « apparié ». La relation entre terme « préféré » et terme « apparié » a été formalisée sous forme de relation entre les concepts désignés par ces termes. La relation « *MappedToConcept* » est définie par quatre propriétés :

- *toTerm* : contient le nom du terme de destination de la relation (terme « apparié »),
- *mapping-type* : contient le type sémantique de la relation entre terme « préféré » et terme « apparié » (identique, synonyme ou similaire),
- *toOntology* : contient le nom de l'ontologie source,
- *mapping-comment* : contient une phrase de justification de la relation.

L'intérêt de l'approche réside dans le fait qu'un terme « préféré » peut être en relation avec plusieurs termes « appariés » désignant des concepts de différentes ontologies sources grâce aux différentes instanciations de la relation « *MappedToConcept* ».

Une fois toutes les suggestions de fusion traitées, nous pouvons considérer que la construction du vocabulaire partagé est terminée.

3.3 La traduction

L'algorithme de traduction utilise l'ensemble des instances de la relation « *MappedToConcept* » pour fournir, à partir d'un cas décrit selon un point de vue, une description de ce même cas décrit selon un point de vue différent.

La méthode de traduction utilise trois paramètres – le concept à traduire, l'ontologie source et l'ontologie destination – afin de retourner le concept de l'ontologie de destination résultant de la traduction

Deux scénarios sont envisageables :

- le concept source est désigné par un terme « préféré » du vocabulaire partagé. Il suffit de trouver, parmi les termes « appariés » au terme « préféré », celui qui désigne un concept de l'ontologie de destination (ce

concept existe s'il existe une instance de la relation « MappedToConcept » entre ce concept de l'ontologie de destination et le concept désigné par le terme « préféré » dans le vocabulaire partagé).

- le concept source n'est pas désigné par un terme « préféré ». Dans ce cas, il faut d'abord identifier le concept en relation avec ce concept source qui est désigné par un terme « préféré ». On retrouve ensuite le premier scénario décrit ci-dessus.

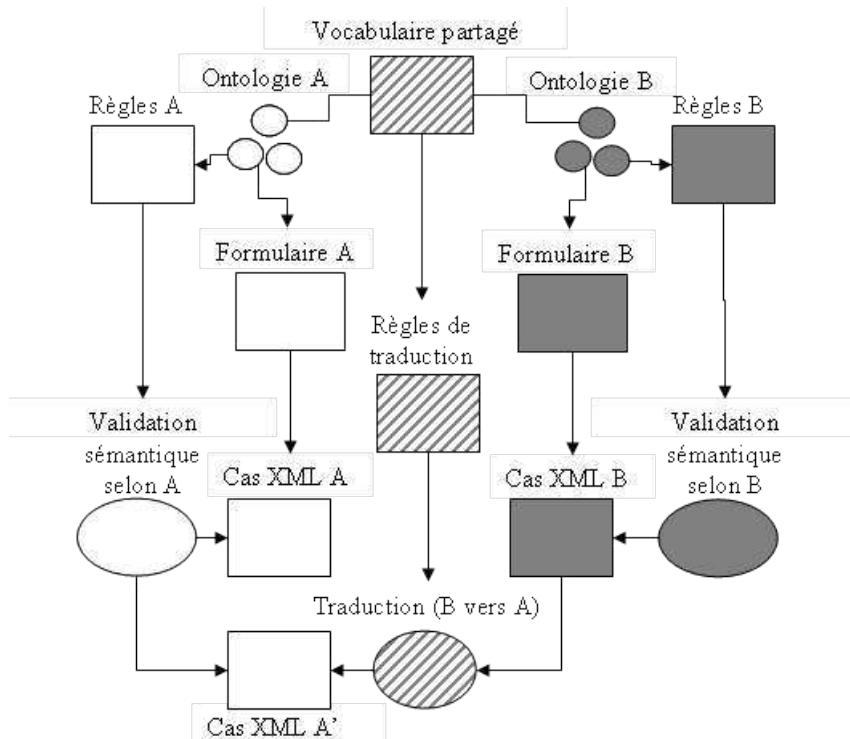


Fig. 1 – L'approche hybride permettant l'intégration de multiples ontologies. Construction et évaluation des ontologies locales correspondent aux multiples points de vue. La méthode de validation d'une description de cas selon un point de vue est basée sur un schéma XML exprimant les règles d'interprétation correspondant à l'ontologie locale de référence; construction du vocabulaire partagé et développement d'un algorithme de traduction de la description d'un cas d'un point de vue vers un autre basé sur un appariement entre le vocabulaire partagé et les ontologies locales.

4 Résultats

4.1 Ontologies locales

Les ontologies locales constituées dans Protégé à partir des classifications de Holland, Van Nuys et Lagios contiennent respectivement 27, 18 et 24 concepts correspondant aux grades de CCIS et aux caractéristiques morphologiques observées au niveau des images. Les relations « est un » et « est inféré par » ont été définies au niveau des trois ontologies locales construites.

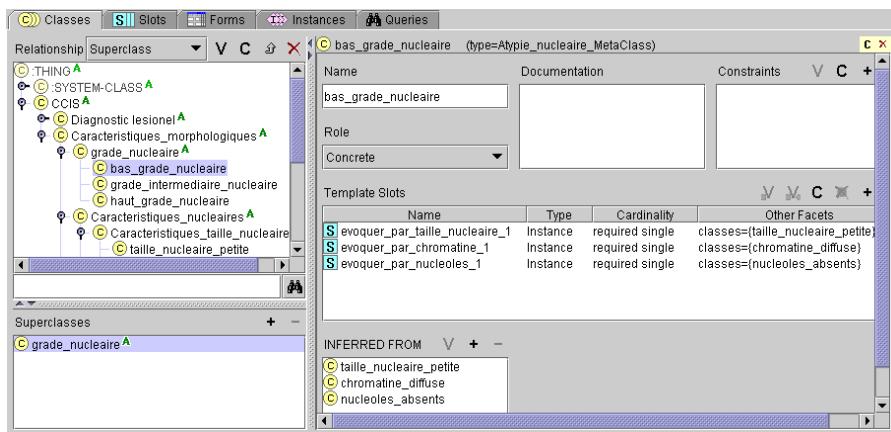


Fig 2. – Ontologie locale

4.2 Vocabulaire partagé et module de traduction

Le nouveau mode de fonctionnement, appelé « mapping » intégré dans le système multi ontologique PROMPT s'apparente dans un premier temps au mode fusion, l'utilisateur choisit un des concepts à fusionner et le terme qui le désigne devient le terme « préféré » placé dans le vocabulaire partagé. Puis le logiciel demande à l'utilisateur le type de la relation existant entre le terme « préféré » et chacun des termes désignant les concepts source et lui permet d'expliquer son choix.

Après l'analyse de toutes les suggestions de fusion des concepts des trois ontologies locales prises 2 à 2, le vocabulaire partagé constitué comporte 30 termes « préférés ». De plus, à chaque terme « préféré » de ce vocabulaire partagé sont associées les instantiations disponibles de la relation « MappedToConcept » définissant les règles d'appariement entre le concept désigné par le terme « préféré » et des concepts des différentes ontologies sources.

Le plug-in développé dans Protégé permettant d'extraire, sous forme de schéma XML, les règles d'appariement du vocabulaire partagé permet à l'algorithme de traduction de produire une représentation XML des appariements et d'exprimer un concept d'une ontologie A (classification A) dans une ontologie B (classification B).

5 Évaluation de la traduction

- Matériel

Trente-trois cas ont été décrits par un expert à deux reprises en utilisant dans le module de description de la plateforme IDEM le formulaire basé sur l'ontologie locale de Holland d'une part et celui basé sur l'ontologie locale de Van Nuys d'autre part correspondant aux deux systèmes de classification dont ils étaient familiers. La description saisie via le formulaire est ensuite soumise au module de validation, c'est-à-dire qu'on vérifie si elle est conforme aux règles d'interprétation de la classification. Ce module génère des messages d'erreur en cas de non-conformité. Par convention avec les pathologistes, un cas est valide s'il génère au plus un message d'erreur pour la classification Holland et s'il ne génère aucun message d'erreur pour la classification Van Nuys. Ces différences de convention entre les deux classifications sont dues à leur nature différente. La classification de Holland s'appuie sur des descriptions de prototypes et est naturellement plus imprécise et ambiguë que la classification Van Nuys qui s'appuie sur un arbre de décision.

Pour la classification Van Nuys, onze cas sur trente trois (soit 33%) sont conformes aux règles d'interprétation, vingt deux cas ne sont pas conformes dont dix-sept (soit 52%) génèrent un message d'erreur et cinq (soit 15%) deux messages d'erreur ou plus. En ce qui concerne la classification de Holland, aucun des trente trois cas n'est rigoureusement conforme aux règles d'interprétation. C'est à dire qu'aucune des descriptions n'est prototypique d'un des trois grades de carcinome canalaire *in situ*. Mais, dans la mesure où les cas générant un seul message d'erreur sont considérés comme valides, douze cas sur trente trois (soit 36%) sont valides. Vingt et un cas sur trente trois (soit 64%) génèrent deux messages d'erreur ou plus.

- Traduction des concepts des ontologies locales

Le taux de traduction des ontologies locales a été évalué en calculant le pourcentage de concepts d'une ontologie locale qui étaient traduisibles vers une autre ontologie locale. Nous avons calculé ce taux de traduction pour les trois classifications disponibles sous forme d'ontologies locales (VN : Van Nuys ; HL : Holland ; LA : Lagios). La traduction concept par concept d'une façon générale donne des résultats satisfaisants en terme de pourcentage de concepts traduisibles comme le montre le tableau 1.

Table 1. Pourcentage de concepts traduisibles

	Nb total de concepts	Concepts traduisibles (%)		
		VN	HL	LA
VN	18		100	83,3
HL	27	66,6		66,6
LA	24	62,5	75	

- Traduction des trente trois cas d'une ontologie locale à une autre

La traduction d'une description de cas d'un point de vue vers un autre a été évaluée en utilisant les trente trois cas décrits selon la classification de Van Nuys puis traduits selon la classification de Holland (et réciproquement). Une analyse qualitative de la traduction a été réalisée par un expert.

Par ailleurs, nous nous sommes intéressés à savoir si la traduction conservait la validité des cas. Nous avons comparé la validité des cas décrits selon une classification à la validité des cas obtenus après traduction selon la nouvelle classification. La concordance obtenue est bonne ($Kappa = 0,94$) lors de la traduction de Holland vers Van Nuys et mauvaise ($Kappa = 0,02$) en ce qui concerne la traduction de Van Nuys vers Holland.

6 Discussion

Notre problématique consistait à proposer une solution de représentation de points de vue différents de pathologistes pour la démarche diagnostique en anatomie pathologique et à offrir une solution méthodologique et technique pour rendre ces points de vue interopérables. Nous avons mis en œuvre une approche décrite par Wache et expérimentée par Buccella consistant à représenter les points de vue sous forme d'ontologies locales, à construire un vocabulaire partagé et à exprimer les correspondances existant entre les concepts des ontologies locales au niveau du vocabulaire partagé (appariement). Un intérêt important de cette approche est de pouvoir envisager de nouvelles ontologies locales qui venant s'apparier au vocabulaire partagé peuvent alors rapidement interopérer avec les autres ontologies locales. Cet appariement a permis de développer un module de traduction de descriptions structurées de cas exprimés selon une ontologie locale vers une autre ontologie locale. La méthode de traduction repose sur la modification d'une description de cas au format XML selon des règles d'appariement représentées dans le formalisme XML.

Il apparaît dans notre expérimentation que les pathologistes produisent généralement des descriptions de cas de CCIS qui ne correspondent pas rigoureusement aux cas prototypiques décrits dans les systèmes de classification selon lesquels ils interprètent les images. Le module de validation vérifiant la conformité des descriptions permet aux pathologistes de prendre conscience de cet écart entre leur pratique et la situation décrite par le système de classification. De plus, nous avons noté les descriptions sont beaucoup plus souvent conformes en ce qui concerne le système de classification de Van Nuys qui est basé sur un arbre de décision non ambigu alors que le système de classification de Holland, (comme celui de Lagios), repose sur des descriptions prototypiques des grades de CCIS. Or, les cas réels ne correspondent pas souvent à ces descriptions prototypiques.

En ce qui concerne la traduction, les résultats montrent que les concepts sont majoritairement traduisibles d'une ontologie locale à une autre. Un taux élevé de concepts traduisibles (par exemple de Van Nuys vers Holland) n'est pas suffisant pour préjuger de la qualité de la traduction. En effet, si le nombre de concepts de

L'ontologie source est inférieur à celui de l'ontologie de destination (plus riche), même si tous les concepts de l'ontologie source sont traduisibles, il manquera vraisemblablement à la description une fois traduite des concepts de description nécessaires à la bonne application des règles d'interprétation de l'ontologie de destination. A ce titre, nous avons constaté que la traduction des descriptions de cas ne conservait pas toujours la validité de ces cas selon le point de vue considéré. Dans notre expérimentation, c'est particulièrement le cas lors de la traduction de Van Nuys vers Holland. Ce résultat est lié d'une part à la plus grande richesse en concepts de description de la classification de Holland et d'autre part à une expression encore trop rigide des règles d'interprétation au niveau des ontologies locales correspondant aux systèmes de classification basés sur des descriptions prototypiques.

Des améliorations à court terme semblent importantes. En ce qui concerne la traduction, l'algorithme de fusion de PROMPT peut être enrichi afin d'améliorer la qualité des suggestions réalisées lors de la constitution du vocabulaire partagé. En ce qui concerne les ontologies locales, il est nécessaire de définir des relations d'interprétation diagnostique plus spécifiques. En effet, nous n'avons utilisé que la relation de type «est inféré par». Or, les pathologistes utilisent également d'autres types de relations que l'on pourrait implémenter (« est un critère diagnostique nécessaire et suffisant » ou « est un critère diagnostique optionnel »). Cela permettrait de mieux prendre en compte la sémantique des règles d'interprétation fournies par les classifications et d'améliorer la conformité des cas décrits

A plus long terme, nous envisageons d'intégrer les outils développés à la plate forme collaborative d'interprétation d'images anatomopathologique IDEM destinée à des pathologistes qui n'utilisent pas obligatoirement les mêmes classifications pour déterminer leurs diagnostics. La traduction des descriptions permettrait de quantifier la part de la variabilité diagnostique liée à l'utilisation de règles d'interprétation diagnostique par rapport à la variabilité d'observation de caractéristiques morphologiques dans l'image.

Références

- BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances, *Eyrolles*, pp. 305-23, 2000.
- BUCELLA A, CECLICH A. (2003). An Ontology Approach to Data Integration, *JCS&T*; 3(2), 2003; pp 62-68.
- DOAN A, MADHAVAN J, DOMINGOS P, HALEVY A. (2002). Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, pp. 662-673. ACM Press, 2002.
- FALISSARD B. (2001). Mesurer la subjectivité en santé: perspective méthodologique et statistique, Paris, *Masson*, 2001.
- FLEMING KA. (1996). Evidence-based pathology, *J Pathol*, 1996 Jun; 179(2): 127-8.
- FLEMING KA. (2002). Evidence-based cellular pathology, *Lancet*, 2002; 359 (9312): 1149-50.
- GRÜBER TA. (1993). Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition* 5, pp. 199–220, 1993.

- HOVY EH. (1998). Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses, *Proc. 1st Int'l Conf. Language Resources and Evaluation (LREC)*, European Language Resources Assoc., Paris, 1998, pp. 535-542.
- LIEBERMAN MI, RICCIARDI TN, MASARIE FE, SPACKMAN KA. (2003). The use of SNOMED CT simplifies querying of a clinical data warehouse. *AMIA Annu Symp Proc.* 2003;910.
- MENA E, ILLARRAMENDI A, KASHYAP V, SHETH AP. (2000). OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and Parallel Databases*, 8(2); pp. 223-271, 2000.
- MITRA P, WIEDERHOLD G, KERSTEN ML. (2000). A Graph-Oriented Model for Articulation of Ontology Interdependencies. In *Intl. Conference on Extending Database Technology (EDBT)*, pp. 86-100, 2000.
- NOY N, MUSEN MA. (2003). The prompt suite: Interactive tools for ontology merging and mapping. *Journal of Human-Computer Studies*, 59(6); pp. 983-1024, 2003
- NOY N. (2003). Tools for Mapping and Merging Ontologies, Handbook on Ontologies, S.Staab and R. Studer editors, pp. 365-384. *Springer-Verlag*, 2003.
- PREECE A, HUI K, GRAY A, MARTI P, BENCH-CAPON T, JONES D AND CUI Z. (1999). The KRAFT Architecture for Knowledge Fusion and Transformation, in M Brammer, A Macintosh & F Coenen (eds), *Research and Development in Intelligent Systems XVI (Proc ES99)*, Springer, New York, 1999, pp.23-38.
- RIBIÈRE M. (1999). Représentation et gestion de multiples points de vues dans le formalisme des graphes conceptuels, *Thèse de l'université de Sophia-Antipolis*; 1999.
- THIEU M, STEISHEN O, ZAPLETAL E, JAULENT MC ET CHRISTEL LE BOZEC. (2004). Mesures de similarité pour l'aide au consensus en anatomie pathologique. *IC 2004* ; pp. 225-236.
- USCHOLD M, GRÜNINGER M. (1996). Ontologies: Principles, Methods and Applications, *Knowledge Engineering Review*, 11(2), pp. 93-155, 1996.
- WACHE H, VOGELE T, VISSER U, STUCKENSCHMIDT H, SCHUSTER G, NEUMANN H AND HUBNET S. (2001). Ontology-based integration of information - a survey of existing approaches. In Stuckenschmidt, *IJCAI-01 Workshop: Ontologies and Information Sharing*, 2001, pp. 108-117.
- WELLS WA, CARNEY PA, ELIASSEN MS, GROVE MR, TOSTESON AN. (2000). Pathologists' Agreement With Experts and Reproducibility of Breast Ductal Carcinoma-in-Situ Classification Schemes, *Am J Surg Pathol*; 5(24); 2000; pp. 651-670.

Une étude approfondie pour le choix des connaissances à capitaliser en amont de la construction d'une mémoire d'entreprise

Inès Saad, Camille Rosenthal-Sabroux et Michel Grundstein

¹ Laboratoire LAMSADE, Université Paris- Dauphine
Place du maréchal de Lattre de Tassigny
Paris, 75775 Cedex 16, France,
{Saad, Sabroux, grundstein}@lamsade.dauphine.fr

Résumé : Dans cet article nous présentons une démarche d'aide au repérage et à la qualification des connaissances cruciales. Cette démarche est dédiée à l'identification des connaissances nécessitant une opération de capitalisation. Nous nous appuyons sur l'aide multicritère à la décision pour éliciter les informations préférentielles des experts concernant la qualification des connaissances déterminées à partir des exemples d'affectation qu'ils donnent pour la classification d'un échantillon de connaissances dans deux classes de décision. L'une des classes regroupe les connaissances nécessitant une opération de capitalisation et l'autre classe regroupe les connaissances qui ne nécessitent pas une opération de capitalisation. La démarche proposée a fait l'objet de plusieurs validations sur des projets de développement de produits, chez un constructeur automobile.

Mots-clés : Aide multicritère à la décision, connaissances cruciales, connaissances potentiellement cruciales, gestion des connaissances, mémoire d'entreprise, projet de développement de produit automobile.

1 Introduction

Face aux besoins accrus des entreprises de préserver et de partager la connaissance de leurs employés, la gestion des connaissances (ou *Knowledge Management*) a commencé, depuis le début des années 90, à occuper une place de plus en plus importante au sein des organisations. L'ingénierie des connaissances propose des méthodes, des concepts et des outils pour l'acquisition et la modélisation des connaissances, permettant en partie de construire la « mémoire d'entreprise ».

La construction d'une mémoire d'entreprise coûte cher à l'entreprise dans la mesure où il faut acquérir la connaissance auprès de ses détenteurs, la préserver, l'actualiser et la rendre accessible. Il devient alors indispensable de délimiter ces connaissances. Comme le souligne Dieng (Dieng *et al.*, 2001) « *au fil des ans, il est*

apparu que la plupart des entreprises étaient en général intéressées, non par la construction d'un système à base de connaissances, mais plutôt par la capitalisation de leurs connaissances cruciales ».

Dans la section 2, nous présentons le contexte et la problématique que nous traitons dans cet article. Dans la section 3, nous décrivons brièvement notre méthode d'aide à l'identification et à la qualification des connaissances cruciales. Dans la section 4, nous détaillons le modèle dédié au calcul du degré de contribution de la connaissance aux objectifs de l'entreprise.

2 Contexte et problématique

Dans l'industrie automobile, les connaissances créées au cours d'un projet de développement d'un produit automobile sont issues de l'interaction d'un nombre important d'acteurs appartenant à des corps des métiers distincts et parfois localisés sur des sites géographiques différents. Ces connaissances peuvent également être issues de collaborations avec des sous-traitants associés aux projets, dès le stade de la conception du produit. Le problème soulevé concerne à la fois la perte des connaissances détenues par les fournisseurs lors de leur départ en fin du projet, et la dispersion des acteurs de l'entreprise amenés à travailler sur de nouveaux projets. L'entreprise risque alors de perdre les savoir-faire acquis si aucune action de capitalisation n'est envisagée. Parmi les connaissances produites dans un projet (cf. Figure 1), une partie est incarnée dans la tête des acteurs, il s'agit du savoir-faire que Nonaka et Takeuchi (Nonaka & Takeuchi, 1997) appellent « connaissance tacite » en faisant référence aux travaux de Polanyi (Polanyi, 1966). Une autre partie des connaissances tacites explicitable est préservée dans la « mémoire d'entreprise».

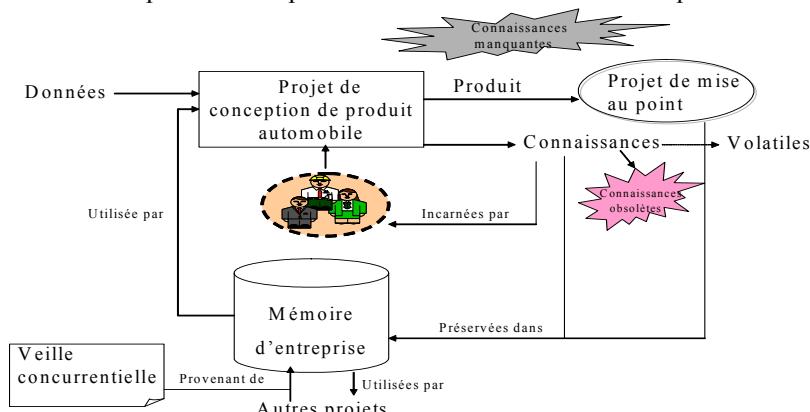


Fig. 1– Synoptique général du flux de connaissances dans un projet automobile

Nous adoptons la définition de mémoire d’entreprise proposée par Rose Dieng (Dieng *et al.*, 2004) qui la caractérise comme suit : « *la mémoire d’entreprise est la matérialisation explicite et persistante des connaissances et informations cruciales d’une organisation en vue de faciliter leur accès, partage, réutilisation par les membres de l’organisation, dans leurs tâches individuelles ou collective* ». Dans notre cas, la mémoire d’entreprise est constituée d’une part des documents techniques et d’autre part des connaissances, modélisées et préservées sous forme de systèmes à base de connaissances, de livres de connaissances, en utilisant des outils d’ingénierie des connaissances.

3 Une méthode d'aide à l'identification et à la qualification des connaissances cruciales

Dans cette section, nous décrivons la méthode que nous proposons pour identifier les « connaissances cruciales ». Cette méthode est composée de quatre étapes. La première étape consiste à définir les « connaissances potentiellement cruciales ». Le concept de « connaissances potentiellement cruciales » est introduit par analogie avec le concept des « actions potentielles » défini par Bernard Roy en aide multicritère à la décision. Selon Roy (Roy & Bouyssou, 1995), une action potentielle désigne « *une action réelle ou fictive provisoirement jugée réaliste par un acteur au moins ou présumée telle par l’homme d’étude en vue de l’aide à la décision* ». Cet ensemble peut être présupposé déjà existant, il est donné alors directement par les acteurs de terrain ou bien au cours des entretiens avec les experts ou une séance de brainstorming. Nous proposons aussi la démarche proposée par le cadre directeur GAMETH (*Global Analysis METHod*) (Grundstein, 2000) avec des adaptations nécessaires liées à la fois au type de projet, projet de développement de produit, et aux -limites de GAMETH (Saad *et al.*, 2003a ; Saad *et al.*, 2003b). La deuxième étape consiste à analyser en profondeur les connaissances repérées. Dans la troisième étape, nous proposons un modèle pour évaluer les « connaissances potentiellement cruciales ». Enfin, nous appliquons les règles de décision que nous avons déterminé, suite à des expérimentations, pour identifier les connaissances effectivement cruciales (Saad *et al.*, 2003c).

Ci-après, nous approfondissons uniquement la troisième étape de notre méthode d’identification des connaissances cruciales consistant à qualifier les connaissances. La construction de la famille de critères est le résultat des expérimentations faites au cours de plusieurs mois à la suite desquels la formulation définitive des critères a été acceptée par un comité d’évaluation. Nous n’exposons pas les différentes hésitations et restructurations des critères tout au long du processus de leur construction.

Nous nous inspirons des trois axes du triangle systémique de Le Moigne (Le Moigne, 1977) afin de construire une famille de critères selon trois points de vue (cf. Figure 2) : le point de vue fonctionnel c'est-à-dire ce que l'objet fait, le point de vue ontologique c'est-à-dire ce que l'objet est, et le point de vue génétique c'est-à-dire ce

que l'objet devient. Dans nos travaux, nous considérons la connaissance comme un système complexe et nous l'analysons selon les trois points de vue qui se présentent comme suit :

- **Le point de vue fonctionnel** consiste à déterminer le degré de contribution d'une connaissance aux objectifs de l'entreprise.
- **Le point de vue ontologique** consiste à déterminer des critères liés aux caractéristiques de la connaissance, afin d'étudier sa vulnérabilité.
- **Le point de vue génétique** consiste à prédire la durée d'usage de la connaissance dans l'entreprise, selon ses objectifs à moyen et à long terme.

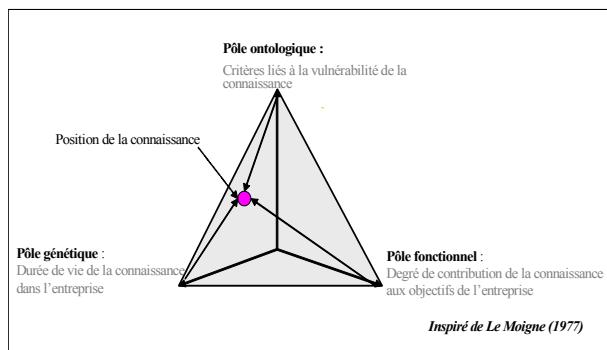


Fig.2– Le modèle de qualification d'une connaissance

Dans ce qui suit, nous détaillons uniquement le point de vue fonctionnel qui consiste à déterminer le degré de contribution de la connaissance aux objectifs de l'entreprise.

4 Le modèle pour calculer le degré de contribution d'une connaissance aux objectifs de l'entreprise

Le modèle proposé s'appuie sur notre postulat selon lequel « *la connaissance est reliée à l'action* ». Nous nous intéressons alors aux connaissances liées aux activités individuelles et collectives des acteurs dans l'entreprise. Ces connaissances sont engagées dans des processus mobilisés dans les projets, et finalisées par les objectifs de l'entreprise. Une connaissance peut ne contribuer à aucun processus et peut au plus contribuer à n processus, un processus contribue au minimum à un projet et au plus à n projets, et un projet contribue au minimum à un objectif et au plus à n objectifs (cf. Figure 3).

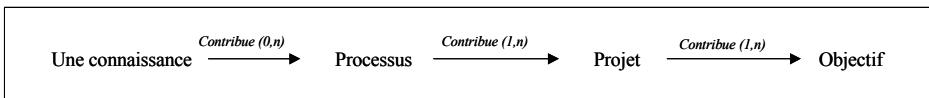
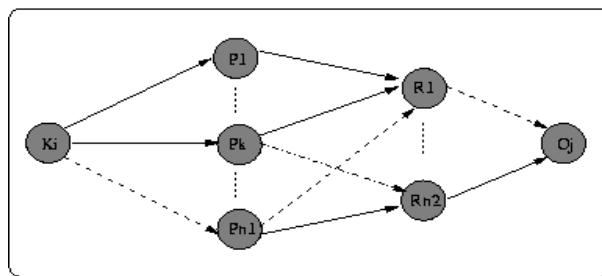


Fig. 3 – Le modèle de calcul du degré de contribution d'une connaissance aux objectifs

Nous rejoignons ainsi l'idée de Philippe Lorino, lorsqu'il explique dans ses travaux qu'il est difficile de déterminer directement l'influence d'une ressource sur les performances de l'entreprise : « Alors qu'il est difficile de répondre à la question : «comment telle ressource influe-t-elle sur telle performance stratégique ? », il est plus aisés de répondre aux deux questions : « comment tel processus influe-t-il sur telle performance stratégique ? » et « comment telle ressource influe-t-elle sur tel processus ? » (Lorino, 2000). Nous constatons qu'il est difficile de mesurer d'une façon quantitative la contribution d'une ressource, et plus particulièrement d'une connaissance, aux performances de l'entreprise. Cela nous amène à proposer un modèle permettant de calculer le degré de contribution d'une connaissance aux objectifs de l'entreprise. Le cogniticien détermine, avec les responsables du projet d'étude, la liste des objectifs, des projets et des processus essentiels mobilisés dans ces projets.

Le degré de contribution de la connaissance aux objectifs de l'entreprise est calculé en trois étapes. La première étape permet d'évaluer le degré de contribution d'une connaissance aux processus, la deuxième étape permet d'évaluer le degré de contribution de chaque processus à chaque projet et enfin la troisième étape permet d'évaluer le degré de contribution de chaque projet à chaque objectif.

Nous considérons un graphe à 4 niveaux (cf. Figure 4). Le premier niveau correspond à la connaissance notée K_i , le deuxième niveau correspond aux processus notés P_1, \dots, P_{n1} où $n1$ est le nombre de processus, le troisième niveau correspond aux projets notés R_1, \dots, R_{n2} où $n2$ est le nombre de projets et le dernier niveau correspond à l'objectif noté O_j . Il existe trois types d'arcs : des arcs reliant la connaissance aux processus ($K_i \rightarrow P_{n1}$), des arcs reliant les processus aux projets ($P_{n1} \rightarrow R_{n2}$) et des arcs reliant les projets à l'objectif ($R_{n2} \rightarrow O_j$). A chaque type d'arc correspond respectivement une valeur donnée par l'acteur concernant la contribution de la connaissance au processus (VD_{K-P}), la contribution du processus au projet (VD_{P-R}) et la contribution du projet à l'objectif (VD_{R-O}). Dans notre cas d'étude, le nombre de processus $n1$ est égal à 11, le nombre de projet $n2$ est égal à 3 et le nombre des objectifs est égal à 6.

**Fig. 4—** Le graphe considéré

Dans ce qui suit, nous présentons tout d'abord la taille de l'échantillon qui nous a permis de valider le modèle, ensuite nous décrivons la démarche d'explicitation du raisonnement des acteurs, puis nous exposons le principe de l'algorithme développé, que nous illustrons par un exemple. Enfin, nous décrivons l'interprétation des résultats.

4.1 Taille de l'échantillon

Pour valider notre modèle, nous l'avons testé sur 34 connaissances potentiellement cruciales (exemple : connaissance relative au dosage de l'additif) mobilisées dans le cadre du projet de développement des différents générations du système Filtre à Particules, notés : FAP_x, FAP_y et FAP_z. Nous avons mené 40 entretiens individuels avec 8 responsables des projets. Dans le tableau suivant (cf. Table 1), nous présentons un extrait des rôles des personnes sollicitées pour évaluer l'importance des connaissances par rapport aux objectifs du projet FAP.

Acteur	Rôle
Acteur 1	Responsable de l'implantation du système FAP sur différents types de moteur
Acteur 2	Responsable du système de dépollution diesel et de l'amélioration de la conception du système de dépollution FAP
Acteur 3	Responsable de l'intégration des nouvelles technologies dans le système FAP et de la coopération avec d'autres partenaires
Acteur 4	Responsable de la fonction FAP
Acteur 5	Responsable de l'amélioration du système FAP

Table. 1— Extrait des rôles des acteurs impliqués dans le projet de développement du système FAP

Après plusieurs réunions avec les responsables, ils ont retenu :

- 3 projets : le projet FAP_x et deux autres projets en cours de développement, considérés comme une évolution technologique par rapport au projet pilote et notés respectivement FAP_y et FAP_z ;
- 11 processus essentiels : 9 processus (exemple : conception et méthodologie de calibration du superviseur) mobilisés dans les trois projets (FAP_x, FAP_y et FAP_z) et 3 processus spécifiques à certains de ces projets (exemple de processus spécifique au FAP_x : choix de l'imprégnation) ;
- 6 objectifs de l'entreprise identifiés dans le cadre des trois projets FAP_x, FAP_y, FAP_z (exemple : minimiser le cycle de développement du système FAP)

4.2 Démarche de recueil des données

Pour calculer le degré de contribution d'une connaissance aux objectifs, nous avons procédé en trois étapes pour expliciter le raisonnement des acteurs :

Première étape : nous avons demandé à chacun des acteurs retenus de donner une évaluation directe, notée VD_{Ki-Oj}, du degré de contribution de la connaissance par rapport à chaque objectif retenu.

Deuxième étape : nous avons demandé à chaque acteur d'évaluer la connaissance par rapport à chaque processus, d'évaluer chaque processus par rapport à chaque projet, et enfin d'évaluer chaque projet par rapport à chaque objectif.

Troisième étape : Nous avons demandé à chaque acteur d'évaluer le degré de contribution, noté VD_{Ki-R}, de chaque connaissance par rapport à chaque projet et de justifier ensuite le choix d'évaluation qu'il a fait. En fait, l'acteur ne choisit un arc de type Connaissance-Processus (K_i-P) qu'en s'assurant que le processus en question a un degré de contribution au projet au minimum « moyen ».

4.3 Le principe de l'algorithme

L'algorithme proposé est construit et validé à partir d'une synthèse des différents raisonnements donnés par les acteurs, au cours des expérimentations issues de 3 projets de développement du système FAP. Cet algorithme permet de maximiser le degré de contribution minimal d'une connaissance à chaque objectif (c'est-à-dire que l'on évalue un chemin par le degré minimal le long de ses arcs, et que l'on recherche ensuite le chemin d'évaluation maximal). En effet, les acteurs cherchent le chemin correspondant à la plus grande valeur de la contribution minimale de la connaissance à un objectif donné. Nous choisissons de déterminer le degré minimal de contribution

d'une connaissance à un objectif parce que les différents niveaux du graphe ont le même niveau d'importance. Par conséquent, il suffit que la contribution d'un arc d'un chemin donné (Connaissance → Projet) soit faible, pour que le degré de contribution de la connaissance par rapport à un objectif soit également faible. De plus, le fait que nous proposons des échelles de type ordinaires rend caduque une agrégation additive. Il semble alors naturel d'utiliser l'opérateur « min » ou l'opérateur « max ». Pour ne pas tomber dans un excès d'optimisme, nous avons adopté l'opérateur « min ».

L'algorithme proposé est constitué de deux étapes :

Etape 1 : nous déterminons dans un premier temps le degré de contribution (Connaissance → Projet).

Pour chaque projet, nous procédons de la manière suivante : on énumère tous les chemins possibles (Connaissance → Processus → Projet) et on retient ensuite le chemin qui maximise le degré minimal de contribution de la connaissance par rapport à chaque projet.

Etape 2 : nous procédons de la même façon, mais en utilisant cette fois-ci le sous-graphe obtenu lors de la première étape. Nous énumérons tous les chemins possibles (Connaissance → Projet → Objectif) et nous sélectionnons ensuite celui qui maximise le degré de contribution minimal de la connaissance vers l'objectif.

4.4 Illustration

Dans ce paragraphe nous présentons un exemple d'application des deux étapes de l'algorithme.

- **Application de l'étape 1 de l'algorithme**

Pour évaluer le degré de contribution de la connaissance K₁, nous énumérons son degré de contribution à tous les processus et nous énumérons également le degré de contribution de chacun des processus à tous les projets. Ensuite, nous calculons les différents chemins qui amènent de la connaissance K₁ vers le projet et nous choisissons celui qui a la meilleure contribution de la connaissance K₁ vers chaque projet (FAP_X, FAP_Y, FAP_Z). En considérant le graphe de la figure 5, nous constatons qu'il y a un chemin maximal de la connaissance par rapport à chaque projet :

- La contribution de la connaissance est « *très importante* » pour le processus P₁ « conception et méthodologie de calibration du superviseur » et le processus P₁ est « *moyen* » pour le FAP_X.
- La contribution de la connaissance est « *faible* » pour le processus P₂ « choix de support filtrant » et le processus P₂ est « *important* » pour le FAP_Y.

- La contribution de la connaissance est « très importante » pour le processus P_k et le processus P_k est « très faible » pour le FAP_Z .

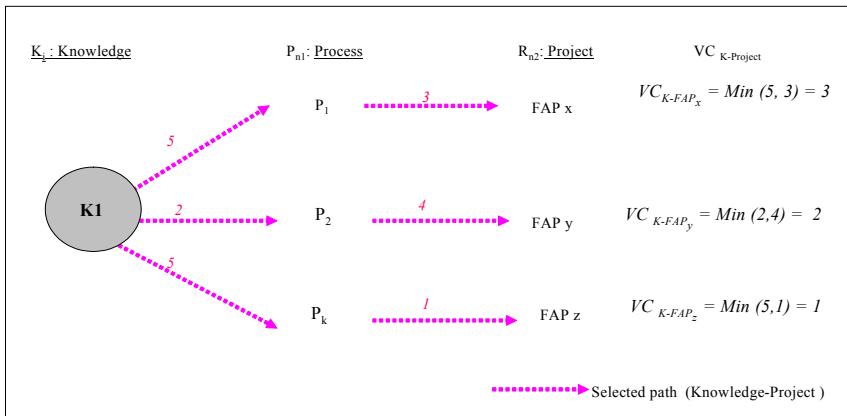


Fig. 5– Le chemin retenu par rapport à chaque projet

- **Application de l'étape 2 de l'algorithme :**

Reprenons le sous-graphe obtenu à l'issue de l'étape précédente. En fonction des chemins retenus à l'étape précédente et en fonction des évaluations des degrés de contribution de chaque projet à un objectif donné, nous choisissons l'évaluation maximum de la connaissance par rapport à l'objectif (cf. Figure 6).

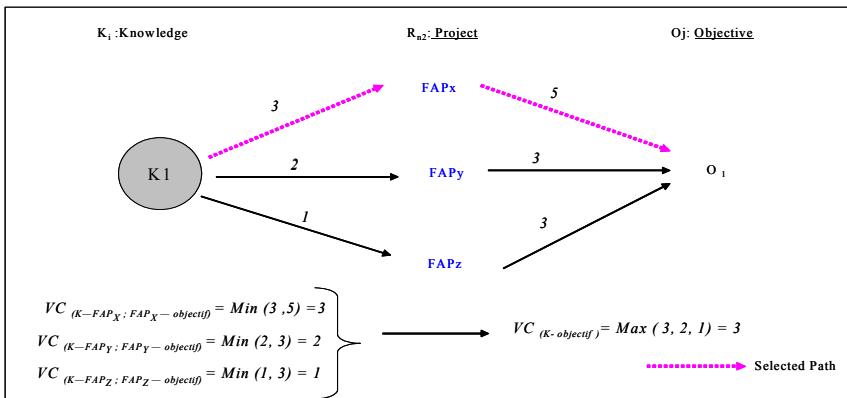


Fig.6– Le chemin VC_{K1-O1} maximal retenu

4.5 Interprétation des résultats

Dans ce qui suit, nous présentons notre analyse des résultats suite à la comparaison entre les valeurs données par les acteurs de terrain et les valeurs calculées en utilisant l'algorithme proposé dans la section précédente :

- **La comparaison, pour chaque acteur, des écarts entre les valeurs qu'ils ont données et les valeurs calculées** : il y a un écart relativement important entre les valeurs données (VD_{k-O}) et les valeurs que nous avons calculées. À titre d'exemple, nous présentons les valeurs calculées (à partir des valeurs données : VD_{k-P} , VD_{P-R} , et VD_{R-O} par l'acteur 5) et les valeurs données VD_{K-O} par l'acteur 5 (cf. Figure 7) pour le calcul de l'importance de la connaissance k : « Connaissance relative au dosage de l'additif », par rapport à chacun des objectifs retenus. Les écarts sont dus à la difficulté, pour chaque expert, de projeter une évaluation liée à la contribution de la connaissance par rapport aux objectifs de l'entreprise ;
- **La comparaison, entre les différents acteurs, des écarts entre les valeurs données et les valeurs calculées** : en fonction des acteurs, les écarts entre les valeurs données (VD_{K-O}) et les valeurs calculées (VC_{K-O}) ne sont pas de même importance. Ceci est dû au degré d'implication des acteurs dans les projets. Par exemple, pour l'acteur 2, nous constatons que les écarts sont tout à fait négligeables (cf. Figure 8) entre la valeur donnée et la valeur calculée concernant la contribution de la connaissance par rapport à chaque objectif. En fait, l'acteur 2 a participé, dès le début du projet, à l'utilisation et au développement des connaissances liées au développement du système FAP. Il s'est ainsi forgé, au fur et à mesure, une vision globale du projet qui s'est renforcée lorsqu'il a évolué du point de vue hiérarchique, sans perdre sa vision opérationnelle. Il continue en effet à intervenir pour résoudre des problèmes techniques et il a participé à des réunions pour résoudre des problèmes liés au FAP¹. Par ailleurs, l'acteur 5 participe au projet, mais en tant que fournisseur de solutions. Il a donc une vision partielle du projet, même s'il est capable de déterminer, pour certaines connaissances, leur degré d'importance par rapport à certains processus.

¹Lors des prochaines évolutions du FAP, il va perdre cette vision technique des nouvelles technologies liées au FAP car son niveau hiérarchique a considérablement augmenté.

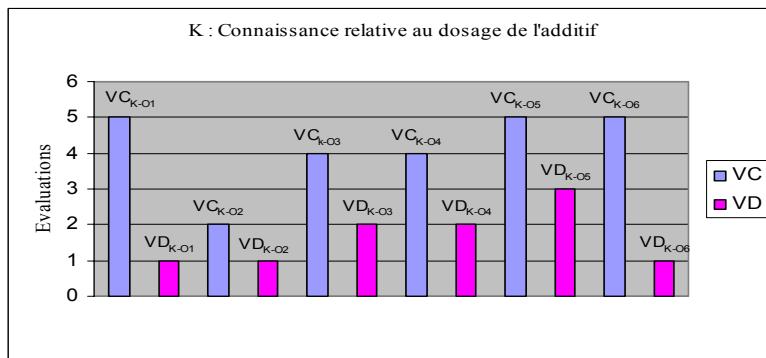


Fig. 7– Comparaison pour l'acteur 5 des écarts entre VC_{K-Oj} et VD_{K-Oj}

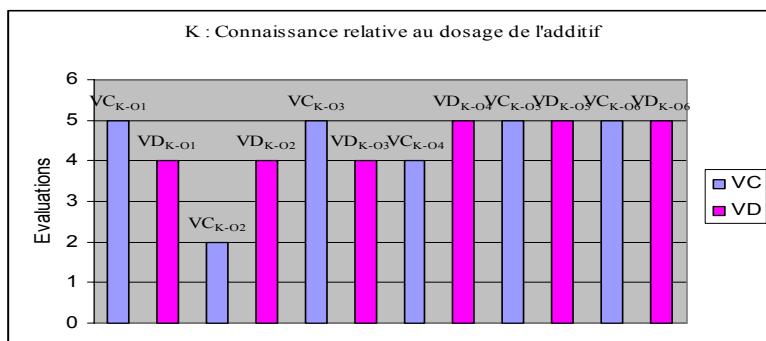


Fig. 8– Comparaison pour l'acteur 2 des écarts entre VC_{K-Oj} et VD_{K-Oj}

Suite à l'analyse, avec les responsables des projets, des écarts entre les valeurs élicitées et les valeurs calculées par l'algorithme, ils ont choisis de conserver les valeurs calculées. Ils ont en effet considéré qu'il n'est pas facile pour eux de donner directement une valeur de degré de contribution de la connaissance aux objectifs de l'entreprise et qu'il est préférable d'avoir un modèle, tel que celui que nous avons présenté, pour faire les calculs. De plus, ils considèrent qu'ils ne sont pas capables de donner des valeurs sur tous les niveaux du modèle. Ils s'accordent à dire qu'il y a des acteurs plus aptes à procéder à des évaluations sur certains niveaux du modèle que d'autres.

5 Conclusion

Dans cet article nous avons présenté une démarche d'identification et de qualification des connaissances nécessitant une opération de capitalisation. Nous avons principalement détaillé le modèle permettant d'évaluer l'importance des connaissances via leurs contributions aux processus contribuant aux projets contribuant eux-mêmes aux objectifs de l'entreprise. Ce modèle permet en partie d'identifier les connaissances mobilisées dans un projet et dont le transfert est vital pour des projets qui traitent des problèmes similaires. Nous considérons que ce modèle, conçu et validé dans le domaine automobile, peut être applicable dans d'autres industries qui fonctionnent en mode projet c'est-à-dire, qui adoptent une vision transversale par les processus (aéronautique, informatique, etc.). Néanmoins, ce modèle devrait faire l'objet d'une validation future dans d'autres projets.

Références

- DIENG R., CORBY O., GANDON F. & GOLEBIOWSKA J. *Ontologies pour la construction d'un web sémantique*. In B. Eynard, M. Lombard, N. Matta et J. Renaud. *Gestion dynamique des connaissances industrielles*, p. 27-42, Paris : Hermès Science.
- DIENG R., CORBY O., GANDON F., GIBOIN A., GOLEBIOWSKA J., MATTA N. & RIBIÈRE M. (2001). *Méthodes et outils pour la gestion des connaissances*, Paris : Dunod.
- GRUNDSTEIN M. (2000). *From capitalizing on Company Knowledge to Knowledge Management*. In M. Morey, M. Maybury & B. Thuraisingham. *Knowledge Management, Classic and Contemporary*. p.261-287. Massachussets : The MIT Press.
- LE MOIGNE J.L. (1977). *La théorie du système générale, théorie de la modélisation*, Paris : P.U.F.
- LORINO Ph. (2000). *Méthodes et pratiques de la performance*. Paris : Editions d'Organisation.
- NONAKA I. & TAKEUCHI H. (1997). *La connaissance créatrice*. Paris : DeBoeck Université (traduction de la version américaine de 1995).
- POLANYI M. (1966). *The tacit dimension*. London: Routledge & Kegan Paul Ltd.
- ROY B. & BOUYSSOU D. (1995). *Aide multicritère à la décision*. Paris : Economica.
- SAAD I., GRUNDSTEIN M. & ROSENTHAL-SABROUX C. (2003a). Méthodes d'identification des connaissances cruciales pour le choix d'investissement dans des opérations de capitalisation des connaissances, Ed., CITE'2003, p. 335- 348, Troyes.
- SAAD I., GRUNDSTEIN M. & ROSENTHAL-SABROUX C. (2003b). *Locating The Company's Crucial knowledge to Specify Corporate Memory : A Case Study in Automotive Company*, Workshop Knowledge Management and Organisational Memory, Ed., IJCAI'2003, p.75-84, Acapulco.
- SAAD I., BASSERAS F. & ROSENTHAL SABROUX C. (2003c). *Les démarches de gestion des connaissances : vers une meilleure conception et exploitation des systèmes d'information coopératifs dans l'entreprise étendue*. In Revue RSTI, Ingénierie des Systèmes d'Information, Volume 8, n° 2/2003, p.33-56.

COWOS : un modèle des situations de travail coopératif pour l'apprentissage de l'organisation

Neil Taurisson, Pierre Tchounikine

LIUM - Laboratoire d'Informatique de l'Université du Maine
{neil.taurisson, pierre.tchounikine}@lium.univ-lemans.fr

Résumé : Cet article propose un modèle des situations de travail coopératif explicitant le processus d'organisation du travail. Ce modèle a été élaboré pour supporter l'apprentissage de l'organisation à travers un environnement de simulation. Le modèle est présenté à travers l'exemple d'une situation modélisée en se focalisant sur l'apprentissage qu'on peut supporter à travers la manipulation des notions du modèle par des apprenants.

Mots-clés : Modélisation du travail coopératif, Apprentissage de l'organisation, Théorie de l'activité, Système multi agents, Environnement de simulation

1 Introduction

Le travail présenté se situe dans le domaine de la conception d'environnements informatiques supportant l'apprentissage du travail coopératif. Notre démarche consiste à proposer aux apprenants un environnement de simulation basé sur un modèle explicite des situations de travail coopératif. En effet l'apprentissage de l'organisation nécessite de la part de l'apprenant un travail de modélisation de la situation à laquelle il est confronté, modélisation devant être supportée par un processus d'essai / erreur, ne pouvant être envisagé qu'à travers une simulation de la situation en question. Le modèle sur lequel repose l'environnement de simulation doit donc être élaboré de façon à supporter le processus d'apprentissage de l'organisation, et rendre explicite des notions spécifiquement liées à l'organisation.

Nous avons élaboré le modèle COWOS à partir de la définition du travail coopératif de (Schmidt 1990) et des notions issues de la théorie de l'activité (Leont'ev 1978), (Engeström 1987). Ce modèle explicite l'activité métad'organisation liée de façon inhérente à toute activité de travail coopératif, et qui constitue notre cible en termes d'apprentissage. Le modèle est présenté à travers un exemple de mise en œuvre en se focalisant sur les types d'apprentissages qu'il permet de supporter.

La section 2 définit les « situations de travail coopératif » et l'« activité d'organisation ». La section 3 présente le modèle et ses fondements théoriques. La section 4 expose une mise en œuvre du modèle dans une situation concrète. La section 5 définit les catégories d'apprentissage de l'organisation et explore la situation modélisée pour montrer en quoi la manipulation du modèle relève de ces catégories. Enfin, la section 6 décrit la réification du modèle dans une plateforme

multi agents.

2 Les situations de travail coopératif et leur organisation

D'après Schmidt (Schmidt 1990) il ne suffit pas d'observer des interactions sociales dans un contexte de travail pour parler de situation de travail coopératif. Cette notion a un caractère plus spécifique : il s'agit d'une situation où les acteurs engagés sont interdépendants dans leur travail. De façon cohérente avec la définition de (Schmidt 1990), nous considérons une situation de travail coopératif comme un système d'activité (au sens de la théorie de l'activité (Leont'ev 1978)) dans lequel les acteurs sont consciemment engagés dans un processus collectif de production, leur travail y étant interdépendant (i.e., les actions des uns dépendent de celles des autres) ce qui demande de leur part de mener une activité secondaire visant à coordonner leurs actions individuelles : l' « activité d'organisation ».

Dans des travaux ultérieurs (Schmidt & Simone 1996, Simone & al. 1996), Schmidt et Simone se sont intéressés à l'articulation du travail collectif dans une démarche visant à l'automatiser (tout ou partie) de façon à faciliter le travail coopératif. Les auteurs définissent pour cela le concept de mécanisme de coordination permettant de réifier et de stabiliser l'organisation. Notre objectif est très différent. En effet, nous nous intéressons à l'activité d'organisation en tant que telle, sans chercher à l'automatiser ou à la masquer. Notre démarche consiste au contraire à amener les acteurs d'une situation de travail coopératif à expliciter cette organisation.

L'activité d'organisation, en tant qu'activité au sens de la théorie de l'activité, est un processus de production. Schmidt définit le produit de cette activité comme l' « organisation du travail ». L'organisation du travail est un instrument, artefactuel et/ou psychologique, qui cristallise le motif et les moyens de l'activité collective sous-jacente. L'activité d'organisation est une activité de niveau méta, motivée par la mise en place et le maintien d'une structure coopérative entre les individus leur permettant de réaliser ce travail coopératif dans les meilleures conditions (Schmidt 1994).

L'interdépendance des actions de chacun fait d'une situation de travail coopératif un système complexe au sens de la théorie des systèmes complexes (Le Moigne 1990). Ce sont des situations dont les différents éléments sont interreliés, dont on peut exhiber différents niveaux d'organisation (organisation individuelle de chaque acteur, organisation globale du groupe) et qui sont instables, irréversibles et irréductibles : instables car le système évolue d'un état stable de l'organisation du travail vers un état instable (état de l'organisation ne permettant plus le bon déroulement de l'activité) nécessitant que les acteurs passent au niveau de l'activité d'organisation pour établir un nouvel état stable ; irréversibles car l'activité crée un contexte qu'elle ne cesse de modifier (Suchman 1987); irréductibles car il est impossible d'en donner une description simple, le système acquiert des propriétés globales qui ne sont pas présentes dans les éléments qui le composent. Leont'ev décrit cette irréductibilité en définissant l'activité comme une unité atomique de vie, unité que la somme des actions des acteurs ne suffit pas à décrire (Leont'ev 1978).

3 COWOS : cadres théoriques et modèle

D'après Le Moigne (Le Moigne 1990), on ne peut appréhender le fonctionnement d'un système complexe que par la modélisation et la simulation. Les caractéristiques d'un système complexe font qu'il est impossible d'en comprendre la dynamique autrement que par la pratique, en s'y confrontant pour comprendre la portée de ses actions sur la globalité du système. En terme d'environnement informatique pour l'apprentissage, on peut proposer à l'apprenant un environnement de simulation lui permettant d'expérimenter des situations virtuelles de travail coopératif.

Un environnement de simulation repose sur une modélisation computable du phénomène simulé. Les choix de modélisation ne sont pas pédagogiquement neutres puisqu'ils définissent les notions perçues et manipulées par les apprenants. Afin d'amener les apprenants à travailler sur l'organisation du travail coopératif nous avons élaboré un modèle des situations de travail coopératif explicitant l'activité d'organisation dénotant l'organisation du travail telle que Schmidt la définit. Ce modèle se veut un modèle de référence en permettant à la fois de définir des situations d'apprentissage et de proposer une « grille de lecture » aux apprenants pour comprendre l'organisation par la manipulation de ces notions à travers une simulation.

3.1 Cadres théoriques pour la conception du modèle COWOS

Le modèle COWOS est fondé sur la description qui est faite par la théorie de l'activité des représentations qu'un individu mobilise lors de son travail. Nous reprenons les trois niveaux de représentations activité / action / opération (Leont'ev 1978), ainsi que les notions du triangle de l'activité (Engeström 1987) qui nous permettent de modéliser la représentation de l'organisation du travail.

Le modèle que nous proposons se veut computable : permettre de simuler dans un environnement informatique une situation de travail coopératif. Pour cela, nous avons réinterprété le paradigme général multi agents BDI (Bratman & al. 1988) avec les notions proposées par la théorie de l'activité. Le modèle BDI permet d'expliquer les processus et représentations liés au raisonnement orienté par les buts (adoption d'un but, mise en place d'un plan pour l'atteindre), c'est-à-dire des représentations concernant les connaissances procédurales et leur mise en œuvre. Nous avons également utilisé des notions proposées par le paradigme « tâche / méthodes » (Trichet & Tchounikine 1999) pour affiner les représentations des connaissances que l'agent mobilise pour l'organisation du travail coopératif.

3.2 Les principales notions du modèle COWOS

Le modèle propose d'expliquer les représentations qu'un agent particulier se fait du travail collectif qu'il réalise à travers les trois niveaux activité / action / opération.

La notion d'activité permet d'expliquer le contexte qui motive la réalisation d'action par l'agent.

La notion d'action permet de dénoter la représentation qu'a un agent du processus

qu'il met en place pour réaliser un but. L'adoption d'un but (une tâche par exemple) est dénotée par la représentation d'une action. Une action permet d'associer au but une méthode opérationnelle pour l'atteindre, les instruments mobilisés ainsi qu'une représentation du résultat produit. La notion d'action dénote le processus qui va de l'adoption du but, du choix de la méthode opérationnelle et des instruments (phase d'orientation) à l'exécution de la méthode (phase de réalisation).

La notion d'opération permet de dénoter les constituants des méthodes opérationnelles. Les opérations sont des routines « internalisées » par l'agent : elles ne sont pas soumises à une phase d'orientation. Une opération dénote les conditions dans lesquelles elle peut être réalisée. La notion d'opération est liée à la notion d'instrument. Un instrument est la représentation de l'association à un objet (matériel ou pas) d'opérations que l'agent sait pouvoir réaliser avec lui (l'instrument « cristallise » ces opérations).

La notion de méthode opérationnelle permet de dénoter la représentation qu'a un agent de ses connaissances procédurales (son savoir faire). Une méthode opérationnelle associe à un couple type de but / contexte d'activation un plan reflétant la réalisation d'opérations par l'agent pour atteindre son but dans les conditions spécifiques de l'environnement. A un même but peuvent être associées plusieurs méthodes opérationnelles en fonction du contexte d'activation des méthodes.

Etroitement associée à l'activité, la notion de structure d'activité permet de dénoter la représentation que se fait un agent, à un moment donné, de l'organisation du travail. La notion de structure d'activité dénote les éléments médiateurs de l'activité proposé dans (Engeström 1987) : sujets, objets, instruments, production, règles, communauté et division du travail. Dans notre modèle, la structure d'activité représente le produit de l'activité d'organisation.

Dans l'optique d'expliciter les processus et les représentations mobilisées par les agents pour organiser le travail collectif, les connaissances des agents concernant l'organisation du travail sont modélisées de façon explicite à travers la notion de méthode d'organisation. Une méthode d'organisation est une représentation plus ou moins abstraite qu'un agent s'est construit de la résolution d'un problème d'organisation dans le but de la préserver et de la transférer face à un nouveau problème. Une méthode d'organisation dénote les modifications à apporter à la structure d'activité pour résoudre un problème d'organisation.

Les problèmes d'organisation sont de deux types : 1) une tâche doit être organisée, c'est-à-dire que la structure d'activité ne permet plus de réaliser la tâche et qu'il faut la modifier (décomposer la tâche, ajouter des règles, etc.), et 2) une tension doit être résolue, c'est-à-dire qu'un évènement conflictuel apparaît (refus d'un acteur de s'engager dans une tâche, non respect d'une règle, etc.). La notion de tension dénote un évènement révélateur d'un problème d'organisation. Il est associé à un type de tension, c'est-à-dire une représentation abstraite du problème d'organisation que représente la tension. Dans notre modèle, un type de tension peut-être associé à des observables discriminant son apparition.

4 Mise en œuvre du modèle

Les exemples présentés dans cette section ainsi que dans la section suivante sont issus de la modélisation du processus d'organisation d'une conférence scientifique. Le point de vue modélisé est celui de l'agent Pierre qui a le rôle de président du comité de programme. Cet exemple a constitué notre base de travail pour l'élaboration du modèle COWOS.

4.1 Représentation de l'activité et de l'organisation du travail

La table 1 présente la représentation de l'agent Pierre de l'activité « processus de sélection des articles pour la conférence » et de son organisation (champs « structure d'activité). C'est la représentation qu'a l'agent Pierre de l'activité au moment où le comité de programme est constitué (la tâche « constitution du comité de programme » de la division du travail est réalisée) et où la soumission des articles est terminée (la règle « date limite de soumission » est dépassée).

Table 1

Activité		
nom : processus de sélection des articles pour la conférence		
<u>organisation du travail</u>		
structure d'activité		
sujets	Agent (s)	Role (s)
	Pierre	Président du comité de programme
	Edouard, Bernadette, Constant,...	Membre du comité de programme
	Justine, Ludovic, Bertrand, ...	Membre du steering committee
objet	Sélectionner les meilleurs articles pour présentation et parution dans les actes	
instruments	<ul style="list-style-type: none"> - courrier électronique - articles soumis - outil de gestion des articles 	
règles	<ul style="list-style-type: none"> - date de la conférence (<i>non atteinte</i>) - date de clôture des soumissions (<i>dépassée</i>) 	
communauté		
division du travail	<ul style="list-style-type: none"> - tâche « choix d'un lieu pour la conférence » (<i>terminée</i>) - tâche « choix d'une date pour la conférence » (<i>terminée</i>) - tâche « constitution du comité de programme » (<i>terminée</i>) - tâche « attribution des articles aux relecteurs » - tâche « gestion des relectures » - tâche « réunion du comité de programme » - ... 	
production	liste d'articles sélectionnés	

La structure de l'activité reflète l'état de l'organisation du travail à un instant donné. Cette représentation est individuelle, elle décrit la représentation qu'a Pierre de l'organisation du travail, celle des autres pouvant en être très éloignée. L'activité d'organisation consiste notamment à faire en sorte que les représentations de chacun soient suffisamment proches pour que le travail collectif soit réalisable.

4.2 Réalisation du travail

La table 2 décrit la représentation de Pierre de l'outil de gestion des articles. Il s'agit d'une représentation individuelle ; le même artefact, l'interface web, n'aura pas la même représentation pour un autre agent qui ne lui associera pas les mêmes opérations. Un relecteur lui associera des opérations relatives au renseignement de sa fiche de présentation ou d'une fiche de lecture, un auteur lui associera des opérations relatives à la soumission d'un article.

Table 2

Instrument		
<u>nom</u>	<u>entrées</u>	<u>sorties</u>
<u>opérations cristallisées</u>		
Compter le nombre d'articles		nb-articles
Compter le nombre de relecteurs		nb-relecteurs
Attribuer un article à un relecteur	article, relecteur	
Retirer l'attribution d'un article à un relecteur	article, relecteur	
Paramétriser l'algorithme d'attribution	nb -naïfs, nb -spécialistes	
Appliquer l'algorithme d'attribution		[attributions]
Lister les articles attribués à un relecteur	relecteur	[articles]
Compter le nombre d'articles attribués à un relecteur	relecteur	nb-articles
Lister les relecteurs à qui a été attribué un article	article	[relecteurs]
Compter le nombre de relecteurs à qui a été attribué un article	article	nb-relecteurs

Une action est réalisée en mobilisant une méthode opérationnelle. La table 3 présente la méthode opérationnelle qui permet à Pierre de réaliser la tâche « établir la liste d'attribution des articles aux relecteurs ». Pour mettre en place cette méthode, l'agent doit disposer de l'outil de gestion des articles qui est décrit dans la table 2 (champs « instruments mobilisés »). Le contexte d'activation dénote que pour appliquer cette méthode, les règles d'attribution des articles doivent être définies. Dans l'état de la structure d'activité de la table 1, cette méthode n'est pas applicable. Pour qu'elle le devienne, Pierre va devoir faire évoluer la structure d'activité.

Table 3

Méthode opérationnelle		
<u>type de tâche réalisée</u> : établir la liste d'attribution des articles aux relecteurs		
<u>instruments mobilisés</u>		
outil de gestion des articles		
<u>contexte d'activation</u>		
<u>règle</u> d'attribution des articles (nb-relecteurs-naïfs, nb-relecteurs-spécialistes) <i>définie</i>		
<u>opérations</u>		
<u>nom</u>	<u>entrées</u>	<u>sorties</u>

Paramétrier l'algorithme d'attribution	nb-relecteurs-naïfs, nb-relecteurs-spécialistes	
Appliquer l'algorithme d'attribution		[attributions]

4.3 L'activité d'organisation

La réalisation d'actions d'organisation (participant à l'activité d'organisation) est médiatisée par un type particulier d'instruments : les méthodes d'organisation. La table 4 en est un exemple. Avec cette méthode, Pierre est capable de rendre la tâche « attribution des articles aux relecteurs » réalisable en apportant à la structure d'activité les modifications dénotées par le champ « structure à adopter ».

Table 4

Méthode d'organisation	
<i>type organiser la tâche « attribution des articles aux relecteurs »</i>	
contexte d'activation	
règle date de clôture des soumissions dépassées tâche « constitution du comité de programme » terminée tâche « attribution des articles aux relecteurs » non organisée	
structure à adopter	
règles	règles d'attribution des articles aux relecteurs : <ul style="list-style-type: none"> - nombre de relecteurs naïfs par article - nombre de relecteurs spécialistes par article
division du travail	tâche « attribution des articles aux relecteurs » décomposée : <ul style="list-style-type: none"> - tâche « établir la liste d'attribution des articles aux relecteurs » - tâche « demander l'avis du steering comitee sur la liste d'attribution » - tâche « transmettre les articles à relire aux relecteurs »

Ne disposant pas de méthode opérationnelle lui permettant de réaliser la tâche d'attribution des articles dans le contexte de la table 1, Pierre est face à un problème d'organisation. Puisque le contexte d'activation de la méthode ci-dessus correspond au contexte dans lequel il se trouve, Pierre peut l'appliquer. Cette application aura pour effet la modification de la structure décrite en table 5. L'organisation du travail nouvellement adoptée permet à Pierre de réaliser la tâche « établir la liste d'attribution des articles aux relecteurs » en mobilisant la méthode opérationnelle décrite en table 3 qui est devenue applicable.

Table 5 En grisé, les modifications apportées à la structure décrite en table 1

règles	<ul style="list-style-type: none"> - Date de la conférence (<i>non atteinte</i>) - Date de clôture des soumissions (<i>dépassée</i>) - règles d'attribution des articles aux relecteurs : <ul style="list-style-type: none"> o nombre de relecteurs naïfs par article o nombre de relecteurs spécialistes par article
division du travail	<ul style="list-style-type: none"> ... - tâche « attribution des articles aux relecteurs » <ul style="list-style-type: none"> o tâche « établir la liste d'attribution des articles aux relecteurs »

- | | |
|--|--|
| | <ul style="list-style-type: none"> ○ tâche « demander l'avis du steering comitee sur la liste d'attribution » ○ tâche « transmettre les articles à relire aux relecteurs » |
|--|--|

...

5 L'apprentissage de l'organisation par la manipulation du modèle

5.1 Catégories d'apprentissage de l'organisation

On se réfère, pour catégoriser les différents types d'apprentissages que nous visons à supporter à travers l'utilisation du modèle, aux catégories d'apprentissage définies par Bateson (Bateson 1977) et à leur interprétation dans le cadre de la théorie de l'activité proposée par Engeström (Engeström 1987). Bateson définit cinq catégories d'apprentissage (apprentissage zéro à apprentissage IV), celles qui nous intéressent sont les apprentissages I, II et III (l'apprentissage zéro étant de très bas niveau, et l'apprentissage IV inatteignable par l'être humain). L'apprentissage de l'organisation à travers le modèle COWOS réside dans la manipulation de la notion de méthode d'organisation. Nous identifions les types de manipulation de cette notion aux catégories d'apprentissage de Bateson.

L'apprentissage I consiste à apprendre à discriminer un contexte. Il y a apprentissage I si le sujet, ayant identifié un comportement comme adéquat, produit la même réponse lorsque confronté au même contexte. Une méthode d'organisation représente la « marche à suivre » face à un problème d'organisation. A ce niveau, la décision de mobiliser une méthode d'organisation est un processus d'apprentissage I : il s'agit de reconnaître dans le contexte (état de la structure d'activité et problème d'organisation), un contexte auquel l'agent a déjà été confronté dénoté par le contexte d'activation d'une méthode d'organisation.

L'apprentissage II consiste à être capable de transférer dans un nouveau contexte un apprentissage I. Le processus de l'apprentissage II implique de la part du sujet la construction d'une représentation mentale de la tâche qu'il a réalisée pour pouvoir la transférer dans un nouveau contexte (Wartofsky 1979). Le sujet construit cet « instrument secondaire » (Engeström 1987) en élaborant un modèle de la tâche qu'il réalise et de la façon dont il la réalise.

Engeström décrit deux types d'apprentissage II : reproductif et productif. Le premier consiste à adapter l'instrument secondaire à un nouveau contexte, le second à adapter (ou « inventer » dans les termes de l'auteur) l'instrument secondaire pour le mobiliser dans un nouveau contexte. Nous ne pensons pas que les méthodes d'organisation du modèle COWOS sont des instruments secondaires puisqu'elles ne sont pas internalisées par l'apprenant, mais qu'elles permettent aux apprenants, en les explicitant, de se construire les instruments secondaires y correspondant.

Le processus de l'apprentissage II reproductif est supporté par la manipulation du contexte d'activation des méthodes d'organisation. Lorsqu'un agent ne dispose pas d'une méthode d'organisation permettant de résoudre un problème, le modèle lui

permet d'adapter le contexte d'activation de l'une de ses méthodes d'organisation pour la rendre applicable dans le contexte courant. Le processus d'apprentissage II productif est supporté par la possibilité pour un apprenant d'adapter une méthode (au-delà de l'adaptation du contexte d'activation, de modifier la structure à adopter, etc.) ou d'en créer une nouvelle.

L'apprentissage III entraîne une modification des comportements que le sujet juge acceptables socialement. Engeström appelle l'apprentissage III, apprentissage par expansion. Son processus réside dans un changement durable des structures d'activité acceptées au sein d'une communauté (c'est l'expansion). Cette apprentissage naît des contradictions qui apparaissent entre la structure d'activité habituellement admise et l'impossibilité d'achever le travail dans cette structure (c'est une double contrainte dans les termes de Bateson). Dans le cadre de l'utilisation du modèle COWOS, l'apprentissage III consisterait à généraliser l'utilisation d'une méthode d'organisation adoptée face à un problème d'organisation récurrent (exprimant une double contrainte) avec le modèle au contexte réel de la situation modélisée. Le modèle n'embarque pas les mécanismes de l'apprentissage par expansion, mais nous pensons qu'il peut en constituer un instrument s'il est utilisé pour modéliser une situation réelle par les acteurs de cette situation. Il peut, dans ce cadre, servir de révélateur des situations de double contrainte et permettre l'expérimentation d'organisation innovante. Dans (Virkkunen & Kuutti 2000), les auteurs décrivent la manière dont ils ont accompagné une organisation dans l'évolution de ses pratiques de travail en proposant aux employés de modéliser leur activité en analysant sa structure et les situations de double contrainte auxquelles ils étaient confrontés. Nous pensons que le modèle COWOS et l'environnement de simulation peuvent appuyer une telle démarche.

5.2 Exemples de processus d'apprentissage avec le modèle

5.2.1 Apprentissage I : appliquer une méthode d'organisation

La méthode d'organisation de la table 4 permet d'organiser une tâche qu'il n'est pas possible de réaliser sinon. L'autre cas de mobilisation d'une méthode d'organisation est celui de l'apparition de tensions. La table 6 décrit le type de tension « refus d'attribution d'un article par un de ses relecteurs ».

Table 6

Type de tension
nom : refus d'attribution d'un article par un de ses relecteurs
Observables : Réception d'un message d'un relecteur qui refuse l'attribution d'un article

Une tension de ce type ne comporte pas assez d'informations pour permettre au président du comité de réagir car il ne mobilisera pas la même méthode d'organisation si le relecteur refuse parce qu'il est surchargé où parce qu'il se sent incompté pour le thème de l'article. Pour pouvoir appliquer l'une ou l'autre de ces méthodes, l'agent devra expliciter la raison de la tension. Ce processus d'explicitation

relève de l'apprentissage I : il s'agit de « reformuler » la tension de manière à y reconnaître le contexte d'activation d'une méthode. Nous réifions ce processus en définissant une tâche d'« explicitation d'une tension ». Cette tâche consiste à déterminer la motivation du refus puis à spécialiser le type de tension en conséquence. A cette tâche est associée la méthode opérationnelle décrite dans la table 7.

Table 7

Méthode opérationnelle
<u>type de tâche réalisée</u> : expliciter une tension « refus d'attribution d'un article par un de ses relecteurs »
<u>instruments mobilisés</u>
<ul style="list-style-type: none"> - courrier électronique
<u>opérations</u>
<p>si le refus n'est pas motivé</p> <ul style="list-style-type: none"> - opération « envoyer un message au relecteur lui demandant de motiver son refus » - opération « attendre la réception de la motivation du refus » <p>si le refus est motivé par la croyance « relecteur surchargé »</p> <ul style="list-style-type: none"> - opération « spécialiser la tension en refus d'un relecteur s'estimant surchargé » <p>sinon si le refus est motivé par la croyance « relecteur incomptent »</p> <ul style="list-style-type: none"> - opération « spécialiser la tension en refus d'un relecteur s'estimant incomptent » <p>sinon si le refus est motivé par la croyance « relecteur trop proche d'un des auteurs »</p> <ul style="list-style-type: none"> - opération « spécialiser la tension en refus d'un relecteur s'estimant trop proche d'un auteur »

5.2.2 Apprentissage II : reproduction et production de méthodes d'organisation

Face à un problème d'organisation, lorsque l'agent ne dispose pas de méthode d'organisation apte à le résoudre, l'agent doit réorganiser ses représentations pour être capable de résoudre le problème. Dans la suite on exemplifie cette réorganisation à travers les processus d'apprentissage II reproductif (modification du contexte d'activation d'une méthode) et productif (modification de la structure à adopter).

Apprentissage II reproductif : modification du contexte d'activation d'une méthode

Avant de transmettre les articles à lire aux relecteurs, Pierre soumet la liste d'attribution au steering committee pour avis. Un de ses membres peut s'opposer à une attribution parce que le relecteur travaille sur le même projet que les auteurs de l'article. Même s'il ne s'agit pas du même type de tension, le président du comité de programme peut assimiler la situation à un refus par un relecteur pour des raisons de proximité. Disposant d'une méthode pour résoudre ce type de tension, il peut en modifier le contexte d'activation pour la rendre applicable ici (Table 8).

Table 8 Modification apportée au contexte d'activation de la méthode (en gris)

Méthode d'organisation
<u>type résoudre une tension d'attribution d'article à un relecteur</u>
<u>contexte d'activation</u>
tension de type « refus d'un relecteur s'estimant trop proche d'un auteur »

tension de type « opposition à une attribution par un membre du steering committee en raison d'une trop grande proximité d'un relecteur et des auteurs »

Apprentissage II productif : création d'une nouvelle méthode d'organisation

La méthode opérationnelle décrite par la table 3 permet au président du comité de programme de réaliser la tâche « établir la liste d'attribution des articles aux relecteurs ». L'attribution des articles aux relecteurs est faite de façon automatique en appliquant un algorithme embarqué dans l'outil de gestion des articles. Il est inévitable que l'application d'un tel algorithme soit source de tensions, en particulier, il est prévisible que certains relecteurs soient surchargés en raison du manque de relecteurs disponibles dans leur domaine d'expertise. Face à cette situation, le président du comité de programme peut modifier la méthode d'organisation de la tâche « attribution des articles aux relecteurs » présentée dans la table 4 pour ajouter une tâche « lisser le nombre d'articles à relire », tâche qui sera réalisée après la tâche « établir la liste d'attribution des articles aux relecteurs » pour égaliser le nombre de relectures entre les relecteurs. Il s'agit là d'un processus lié à l'apprentissage II productif : la méthode est modifiée pour s'adapter au contexte, celui de l'apparition récurrente de tensions lors de l'application de la méthode.

6 Implémentation

La plate forme multi agents a été développée en respectant la correspondance structurelle (Reinders & al 1991) avec le modèle COWOS. Cette plateforme, développée en Java, correspond à une sur couche à la plateforme multi agents Jade¹. Le noyau de chaque agent correspond à une base de connaissances réifiée selon le modèle proposé en utilisant l'éditeur d'ontologie Protégé (Musen & al. 2000).

La programmation d'un agent consiste à 1) associer du code aux opérations, 2) décrire ses méthodes opérationnelles et d'organisation, 3) décrire les tensions observables. La plate forme est construite de telle sorte qu'il est possible de modifier / ajouter, en cours d'exécution des méthodes et opérations. A chacune des notions du modèle, on peut associer des composants graphiques. De cette façon, il est possible de construire différents environnements réifiant à l'interface tout ou partie du modèle.

Nous avons développé cette plateforme afin de pouvoir implémenter différents types d'agents allant d'agents complètement autonomes à des agents entièrement pilotés par l'utilisateur. Il est possible de donner le contrôle à l'utilisateur soit sur le choix des méthodes opérationnelles, des méthodes d'organisation, le processus d'explicitation. On peut ainsi construire un environnement d'apprentissage supportant les différents types d'apprentissage I et II : apprentissage I, l'apprenant a le contrôle de l'explicitation des tensions et du choix des méthodes opérationnelles, apprentissage II reproductif, l'apprenant a le contrôle du choix des méthodes d'organisation et peut modifier le contexte d'activation de celles-ci, apprentissage II productif, il peut modifier les méthodes d'organisation et en créer de nouvelles.

¹ <http://jade.tilab.com>

7 Conclusion

Ce travail s'inscrit dans une démarche à long terme d'étude de l'instrumentation de l'apprentissage de l'organisation du travail coopératif. Les contributions proposées dans cet article sont 1) un modèle des situations de travail coopératif orienté vers l'apprentissage de l'organisation, 2) une conceptualisation des catégories d'apprentissage que la manipulation du modèle peut médiatiser. La poursuite de cette recherche passe par un travail de développement, au dessus de l'architecture multi agents abordée en section 6, d'environnements d'apprentissage pour différentes situations et différents publics, afin d'étudier l'impact en termes d'apprentissage de la manipulation du modèle. Une autre direction est celle dessinée dans la section 5.1 : l'utilisation du modèle et de l'environnement de simulation pour mettre en place un processus d'apprentissage par expansion au sein d'une situation réelle de travail, c'est-à-dire promouvoir, au sein d'une structure, l'analyse des pratiques et leur évolution.

Références

- BATESON, G. (1977). Vers une écologie de l'esprit. Paris, France, Éditions du Seuil.
- BRATMAN, M. E., D. J. ISRAEL, & al. (1988). "Plans and Resource-Bounded Practical Reasoning." Computational Intelligence 4(4): 349-355.
- ENGESTRÖM, Y. (1987). Learning by Expanding. An activity-theoretical approach to developmental research. Orienta - Konsultit, Helsinki.
- LE MOIGNE, J.-L. (1990). La modélisation des systèmes complexes. Paris, Dunod.
- LEONT'EV, A. N. (1978). Activity, Consciousness, and Personality, Prentice-Hall.
- MUSEN, M. A., R. W. FERGESSON, & al. (2000). Component-Based Support for Building Knowledge-Acquisition Systems. Conference on Intelligent Information Processing, Beijing.
- REINDERS, M., E. VINKHUYZEN, & al. (1991). "A conceptual modelling framework for knowledge-level reflection." AI Communications 4(2/3): 74-87.
- SCHMIDT, K. (1990). Analysis of Cooperative Work. A Conceptual Framework. Roskilde, Denmark, Riso National Laboratory.
- SCHMIDT, K. (1994). Modes and Mechanisms of Interaction in Cooperative Work. Outline of a Conceptual Framework. Roskilde, Denmark, Riso National Laboratory: 76.
- SCHMIDT, K. & SIMONE, C. (1996). Coordination mechanisms: Toward a conceptual foundation of CSCW system design, Computer Supported Cooperative Work: The Journal of Collaborative Computing 5(2/3): 155-200.
- SIMONE, C., SCHMIDT, K., CARTENSEN, P., & DIVITINI, M. (1996). Ariadne: Towards a technology of coordination, Riso National Laboratory
- SUCHMAN, L. A. (1987). Plans and Situated Actions, Cambridge University
- TRICHET, F. & TCHOUNIKINE, P. (1999). "DSTM: a Framework to Operationalize and Refine Problem-Solving Methods modeled in terms of Tasks and Methods." International Journal of Expert Systems With Applications (ESWA) 16: 105-120.
- VIRKKUNEN, J. & K. KUUTTI (2000). "Understanding organisational learning by focusing on activity systems" Accounting Management and Information Technologies 10: 291-319.
- WARTOFSKY, M. (1979). Models: Representation and scientific understanding. Dordrecht.

Alignement d'ontologies pour OWL-Lite : l'apport d'un classifieur sémantique*

Raphaël Troncy¹, Umberto Straccia¹, Henrik Nottelmann²

¹ ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
{troncy,straccia}@isti.cnr.it
<http://nmis.isti.cnr.it/troncy/>

² Institute of Informatics and Interactive Systems,
University of Duisburg-Essen, 47048 Duisbourg, Germany
nottelmann@uni-duisbourg.de

Résumé : Cet article introduit un cadre logique pour apprendre automatiquement à aligner des ontologies formelles, une tâche cruciale pour le Web Sémantique. Notre approche est basée sur la théorie des probabilités, ce qui permet de gérer à la fois l'incertitude intrinsèque au processus de comparaison et les correspondances entre entités ontologiques qui ne sont pas exactes. Plusieurs composants spécialisés sont combinés pour trouver les différents alignements possibles (avec leurs poids). En particulier, un nouveau classifieur utilisant pleinement la sémantique des définitions est proposé. Les ensembles de règles de correspondance (ou *mapping*) sont alors évalués, la probabilité maximale désignant finalement le “meilleur” alignement.

Mots-clés : Alignement d'ontologies, logique probabiliste, apprentissage, classifieur sémantique, OWL

1 Introduction

La recommandation récente par le W3C des langages OWL (OWL, 2004) et RDF pour représenter des ontologies et des annotations formelles sur le Web est une étape supplémentaire pour parvenir au Web Sémantique (Berners-Lee *et al.*, 2001). Alors que des efforts importants sont désormais déployés pour accélérer l'adoption de ces nouveaux langages à travers la mise en ligne de ressources et de guides de bonnes pratiques¹, et pour interroger les structures formelles déjà existantes via un langage de requêtes et des mécanismes de règles², il est un autre domaine primordial pour permettre une large utilisation des ontologies :

*Ce travail a été soutenu et financé dans le cadre d'une bourse ERCIM.

¹<http://www.w3.org/2001/sw/BestPractices/>

²<http://www.w3.org/2001/sw/DataAccess/>

leur comparaison. En effet, les ontologies déjà développées pour un même domaine sont multiples, par exemple dans le cas de la médecine (Charlet *et al.*, 2005). De plus, on peut très facilement imaginer que les ontologies vont être développées dans des environnements distribués qui vont requérir, entre autres fonctionnalités, la possibilité d'intégrer et de fusionner des extraits d'ontologies déjà existantes. Dans ce cadre, l'alignement d'ontologies, c'est-à-dire la comparaison des différentes entités définies dans celles-ci, est une tâche centrale.

Les méthodes de comparaison d'ontologies vont permettre de traduire des requêtes formulées selon une base de connaissance *source* vers une base de connaissance *cible*. Ainsi, des documents annotés formellement selon cette dernière base de connaissance pourront être retrouvés. D'une manière générale, l'alignement d'ontologies sera une des premières tâches pour les agents les consommant pour pouvoir rendre des services sur le web. A l'extrême, ces méthodes de comparaison pourraient être utilisées entre les différentes versions d'une même ontologie afin de retrouver une trace des changements intervenus lors de son évolution et de détecter des éventuelles inconsistances. Enfin, cette tâche doit pouvoir s'effectuer de manière complètement automatique dans la mesure où il n'est pas réaliste de maintenir des mises en correspondance d'ontologies effectuées à la main, a fortiori si ces ontologies dépassent une certaine taille ou complexité. Dans cet article, nous proposons un cadre logique et probabiliste permettant de comparer automatiquement des ontologies. En particulier, nous présentons un classifieur utilisant pleinement la sémantique des définitions et des axiomes OWL pour établir des relations d'équivalence et de subsomptions entre les concepts et les relations (ou propriétés) des ontologies.

L'article est structuré de la manière suivante. Nous introduisons dans la section 2 notre cadre logique et probabiliste qui a pour but de trouver les meilleures correspondances (ou *mappings*) entre entités définies dans des ontologies représentées en OWL. La section 3 détaille notre approche théorique pour apprendre les différents alignements possibles, l'évaluation finale reposant sur la combinaison de prédictions de divers classifieurs. Nous donnons alors une liste non exhaustive de classifieurs pouvant être utilisés. Ceux-ci se basant essentiellement sur les données terminologiques des entités ontologiques à comparer ou sur leurs instances, nous présentons un nouveau classifieur utilisant pleinement la sémantique formelle des définitions et des axiomes OWL (section 4). La section 5 revient sur les différents travaux liés à notre approche. Finalement, la section 6 nous permet de conclure et d'ouvrir quelques perspectives.

2 Un cadre logique pour l'alignement d'ontologies

Cette section introduit notre cadre logique pour comparer automatiquement des ontologies. Celui-ci s'inspire des travaux formels menés autour de l'échange d'information (Fagin *et al.*, 2003) et emprunte à d'autres approches comme GLUE (Doan *et al.*, 2003) l'idée de combiner plusieurs composants spécialisés pour obtenir le meilleur résultat. De plus, notre approche gère l'incertitude intrinsèque au processus de comparaison en se basant sur Datalog probabiliste,

pour lequel des outils sont disponibles. Il reprend donc la formalisation proposée dans (Nottelmann & Straccia, 2005) et constitue une extension du système **sPLMAP** (*Schema Probabilistic Learning Mappings*) pour l'alignement d'ontologies.

2.1 Datalog probabiliste

Datalog probabiliste (Fuhr, 2000) (ou *pDatalog*) est une extension de *Datalog* qui est lui même une variante de la logique des prédictats, basé sur des clauses de Horn sans fonction en argument des prédictats (*i.e.* les arguments sont des constantes ou des variables). Dans pDatalog, tous les faits et règles sont préfixés par un poids probabiliste $0 < \alpha \leq 1$:

$$\alpha A \leftarrow B_1, \dots, B_n$$

où A (la tête de la règle), et B_1, \dots, B_n ($n \geq 0$) (les sous-but du corps de la règle) sont des formules atomiques. Le poids $\alpha = 1$ est facultatif. Une règle αr s'interprète alors comme “*la probabilité qu'une instantiation de la règle r soit vraie est α*”. Par exemple, le programme pDatalog ci-dessous indique qu'une personne est une femme avec une probabilité de 50%.

$$\begin{aligned} \text{personne(ana)} &\leftarrow \\ 0.8 \text{ personne(rapha\"el)} &\leftarrow \\ 0.5 \text{ femme(X)} &\leftarrow \text{personne(X)} \end{aligned}$$

On peut alors calculer $Pr(\text{femme(ana)}) = 0.5$, et $Pr(\text{femme(rapha\"el)}) = 0.8 \times 0.5 = 0.4$. Formellement, une structure d'interprétation³ dans pDatalog est un couple $\mathcal{I} = (\mathcal{W}, \mu)$ où \mathcal{W} représente les mondes possibles (*i.e.* l'instanciation de la partie déterministe d'un programme pDatalog plus un sous-ensemble de la partie probabiliste où toutes les probabilités sont enlevées) et μ est une distribution des probabilités sur \mathcal{W} . Une interprétation est un couple $I = (\mathcal{I}, w)$ tel que $w \in \mathcal{W}$. La valeur de vérité d'une formule, compte tenu d'une interprétation et d'un monde possible, peut être définie récursivement de la manière suivante :

$$\begin{aligned} (\mathcal{I}, w) \models A &\text{ssi } A \in w, \\ (\mathcal{I}, w) \models A \leftarrow B_1, \dots, B_n &\text{ssi } (\mathcal{I}, w) \models B_1, \dots, B_n \Rightarrow (\mathcal{I}, w) \models A, \\ (\mathcal{I}, w) \models \alpha r &\text{ssi } \mu(\{w' \in \mathcal{W} : (\mathcal{I}, w') \models r\}) = \alpha. \end{aligned}$$

Une interprétation est alors un modèle pour un programme pDatalog si et seulement si elle implique tous les faits et règles.

2.2 Ontologies et alignement

Ontologies. Par ontologie, nous entendons la représentation conceptuelle et formelle d'un domaine pour une application donnée. Les ontologies fournissent donc le vocabulaire propre à un domaine et définissent formellement des concepts et des relations entre ceux-ci. Ces concepts (et ces relations) sont organisés *via la*

³Dans pDatalog, seules les interprétations dans l'univers d'Herbrand sont considérées.

relation de subsomption pour former une taxonomie. L'ontologie peut contenir en outre des axiomes formels pour permettre des calculs d'inférences additionnels, et des individus (*i.e.* des instances des concepts représentés). Nous considérons dans la suite que ces ontologies sont représentées dans les langages OWL-Lite ou OWL-DL (OWL, 2004) qui sont dérivés de logiques de descriptions bien étudiées (resp. $\mathcal{SHIF}(\mathcal{D})$ et $\mathcal{SHOIN}(\mathcal{D})$) (Horrocks *et al.*, 2003). Le tableau 1 donne ainsi un extrait de deux ontologies **S** et **T** en utilisant une syntaxe usuelle en logiques de descriptions.

Ontologie source S :
<code>Directions ≡ (≤ 1 town xsd:string)</code>
<code>Congress ≡ (and (≤ 1 id xsd:string) (≤ 1 acronym xsd:string))</code>
<code>(all place Directions)</code>
Ontologie cible T :
<code>Address ≡ (≤ 1 city xsd:string)</code>
<code>Conference ≡ (and (≤ 1 name xsd:string) (≤ 1 shortName xsd:string))</code>
<code>(all location Address)</code>

TAB. 1 – Extrait de deux ontologies **S** et **T**

La partie terminologique $\mathcal{T}\mathcal{F}$ d'une ontologie contient des axiomes de la forme $A \sqsubseteq C$ et $A \doteq C$, où A est un nom de concept et C une expression. Nous transformons alors ces axiomes en forme normale pour la négation (NNF) en utilisant les règles de ré-écriture suivantes sur les différentes composantes d'une définition jusqu'à ce que le connecteur de négation ne soit appliqué qu'à des formules atomiques :

$$\begin{array}{lll} \neg T \text{ devient } \perp & \neg \perp \text{ devient } T & \neg \neg C \text{ devient } C \\ \neg(C \sqcap D) \text{ devient } \neg C \sqcup \neg D & & \neg(C \sqcup D) \text{ devient } \neg C \sqcap \neg D \\ \neg \exists R.C \text{ devient } \forall R.\neg C & & \neg \forall R.C \text{ devient } \exists R.\neg C \\ \neg(\geq n R.C) \text{ devient } (\leq n-1 R.C) \text{ si } n \geq 1. \text{ Pour } n=0 \neg(\geq n R.C) \text{ devient } \perp & & \\ \neg(\leq n R.C) \text{ devient } (\geq n+1 R.C) & & \end{array}$$

De plus, les définitions de concept de la forme $A \sqsubseteq C$ sont remplacés par $A \doteq C \sqcap A^*$, où A^* est un nouveau nom de concept. Finalement, $\mathcal{T}\mathcal{F}$ est *saturée* (*i.e.* complétée par l'ensemble des déductions possibles compte tenu de la sémantique du langage OWL) pour donner $\mathcal{T}\mathcal{F}^* = \{A \doteq C\}$ où A est un nom de concept et C une expression en NNF.

Alignement. Notre but est de déterminer automatiquement des relations d'équivalence (ou de subsomption) entre des entités définies dans des ontologies OWL. Par exemple, étant donné les deux ontologies définies dans le tableau 1, nous voudrions établir une correspondance entre les concepts **Congress** et **Conference**. Théoriquement, l'alignement d'ontologies dans notre cadre suit l'approche GLAV⁴ détaillée dans (Lenzerini, 2002). Un *mapping* est un tuple $\mathcal{M} = (\mathbf{T}, \mathbf{S}, \Sigma)$ où **S** et **T** sont respectivement les ontologies source et cible, et

⁴ Approche mixte *Global and Local as View* développée pour l'interopérabilité des bases de données.

Σ est un ensemble de contraintes (*i.e.* des règles pDatalog) de la forme :

$$\alpha_{j,i} \ T_j(x) \leftarrow S_i(x).$$

Cette règle signifie que le concept (resp. la relation) S_i dans l'ontologie source correspond au concept (resp. la relation) T_j dans l'ontologie cible, et que la probabilité que cet appariement soit effectivement vrai est $\alpha_{j,i}$. Notons qu'aucune contrainte supplémentaire n'est posée sur ces mises en correspondance : un concept de l'ontologie source peut correspondre à 0 ou plusieurs concepts dans l'ontologie cible et réciproquement.

Soit un *mapping* $\mathcal{M} = (\mathbf{T}, \mathbf{S}, \Sigma)$ et un modèle I pour \mathbf{S} , un modèle J pour \mathbf{T} est une *solution* pour I sous \mathcal{M} si et seulement si $\langle J, I \rangle$ (l'interprétation combinée sur \mathbf{T} et \mathbf{S}) est un modèle pour Σ . La solution minimale, notée $J(I, \Sigma)$, est l'instance de la relation correspondante avec $\mathbf{T}(I, \Sigma)$. En utilisant les règles pDatalog, la solution minimale $\mathbf{T}(I, \Sigma)$ est exactement le résultat de l'application des règles Σ sur les entités de \mathbf{S} .

3 Apprendre automatiquement à comparer des ontologies

L'apprentissage des appariements possibles entre ontologies dans notre approche s'effectue en quatre étapes : (*i*) nous devinons un alignement potentiel, c'est-à-dire un ensemble de règles Σ_k de la forme $T_j(x) \leftarrow S_i(x)$ (règles sans poids) ; (*ii*) nous estimons alors la qualité de Σ_k ; (*iii*) parmi tous les ensembles possibles Σ_k , nous sélectionnons le “meilleur” selon une certaine mesure de qualité; et finalement (*iv*) nous estimons le poids α pour les règles appartenant à l'alignement sélectionné.

3.1 Estimation de la qualité d'un ensemble de correspondances entre entités ontologiques

Soit une ontologie source $\mathbf{S} = \langle S_1, \dots, S_s \rangle$, une ontologie cible $\mathbf{T} = \langle T_1, \dots, T_t \rangle$, et deux interprétations I pour \mathbf{S} et J pour \mathbf{T} . Notre but est de trouver le “meilleur” ensemble Σ de règles établissant des correspondances entre entités ontologiques, c'est-à-dire l'ensemble qui maximise la probabilité $Pr(\Sigma, J, I)$ que les objets appartenant à la solution minimale $T(I, \Sigma)$ avec $\mathcal{M} = (T, S, \Sigma)$ et les objets de T soient similaires. $Pr(\Sigma, J, I)$ estime donc la probabilité qu'un tuple de $T(I, \Sigma)$ soit une valeur plausible pour T et vice versa. En utilisant la théorie de Bayes, nous obtenons⁵ :

$$\begin{aligned} Pr(\Sigma, J, I) &= Pr(T(I, \Sigma) | J) \cdot Pr(J | T(I, \Sigma)) \\ &= Pr(T(I, \Sigma) | J)^2 \cdot \frac{Pr(T)}{Pr(T(I, \Sigma))} \\ &= Pr(T(I, \Sigma) | J)^2 \cdot \frac{|J| / |U|}{|T(I, \Sigma)| / |U|} \end{aligned}$$

⁵Où U est l'ensemble des tuples et $Pr(T)$ est la probabilité qu'un tuple donné soit dans J .

$$= \Pr(T(I, \Sigma) | J)^2 \cdot \frac{|J|}{|T(I, \Sigma)|} \quad (1)$$

Nous appelons ensuite $T_j(I, \Sigma)$ la restriction de $T(I, \Sigma)$ à T_j . L'ensemble Σ peut être partitionné en sous-ensembles Σ_j formés des règles de *mapping* ayant T_j comme objet commun en tête. Nous avons alors :

$$\Pr(T(I, \Sigma) | J) = \sum_j \Pr(T_j(I, \Sigma_j) | J_j) \cdot \frac{\Pr(J_j)}{\Pr(J)} \quad (2)$$

Pour s entités dans l'ontologie source et un nombre j fixe, il y a aussi s ensembles possibles $\Sigma_{j,i}$, et donc $2^s - 1$ combinaisons non vides (union) de ces ensembles qui forment tous les sous-ensembles Σ_j possibles⁶. Pour simplifier la notation, nous posons dans la suite $S_i = T_j(I, \Sigma_{j,i})$. Ainsi, si Σ_j est formé par les r sous-ensemble de règles $\Sigma_{j,i_1}, \dots, \Sigma_{j,i_r}$, nous obtenons :

$$\Pr(T_j(I, \Sigma_j) | J_j) = \sum_{k=1}^r \Pr(S_{i_k} | J_j) \quad (3)$$

La probabilité $\Pr(\Sigma, J, I)$ peut donc être obtenue à partir des $\mathcal{O}(s \cdot t)$ probabilités $\Pr(S_i | J_j)$. La section suivante montre comment calculer la probabilité de ces règles.

3.2 Estimation de la probabilité d'une règle

De la même manière que GLUE (Doan *et al.*, 2003), nous estimons la probabilité $\Pr(S_i | J_j)$ en combinant les prédictions de différents classifieurs CL_1, \dots, CL_n . Chaque classifieur calcule un poids⁷ $w(S_i, J_j, CL_k)$ qui est ensuite normalisé et transformé en $\Pr(S_i | J_j, CL_k) = f(w(S_i, J_j, CL_k))$, l'approximation de $\Pr(S_i | J_j)$ pour CL_k . Nous utilisons les fonctions f de normalisation suivantes :

$$\begin{aligned} w(S_i, J_j, CL_k) &\mapsto \Pr(S_i | J_j) , \\ f_{id}(x) &= x , \\ f_{sum}(x) &= \frac{x}{\sum_{i'} w(S_{i'}, J_j, CL_k)} , \\ f_{lin}(x) &= c_0 + c_1 \cdot x , \\ f_{log}(x) &= \frac{\exp(b_0 + b_1 \cdot x)}{1 + \exp(b_0 + b_1 \cdot x)} . \end{aligned}$$

Les fonctions f_{id} , f_{sum} et f_{log} retournent des valeurs dans l'intervalle $[0, 1]$. Pour la fonction linéaire, les résultats inférieurs à 0 (resp. supérieurs à 1) doivent être normalisés à 0 (resp. 1). La fonction f_{sum} assure que chaque valeur appartient à l'intervalle $[0, 1]$ et que leur somme vaut 1. Son principal avantage est qu'elle n'utilise pas de paramètres extérieurs. En revanche, les paramètres des fonctions linéaire et logistique doivent être appris par régression sur un jeu d'entraînement. Cette phase n'étant nécessaire qu'une seule fois, le résultat peut ensuite

⁶ Nous discutons de cette combinatoire apparemment rédhibitoire dans la section 4.3.

⁷ Dans la suite, w représentera toujours un poids.

être utilisé pour comparer n'importe quelles ontologies. Les fonctions de normalisation peuvent naturellement être combinées. Par exemple, il est souvent utile d'amener les poids des classificateurs dans un même intervalle (en utilisant f_{sum}), puis d'appliquer une autre fonction de normalisation (*e.g.* f_{log}).

Pour la probabilité finale $Pr(S_i|J_j, CL_k)$, nous avons la contrainte

$$0 \leq Pr(S_i|J_j, CL_k) \leq \frac{\min(|S_i|, |J_j|)}{|J_j|} = \min\left(\frac{|S_i|}{|J_j|}, 1\right)$$

Par conséquent, la valeur normalisée (qui appartient à $[0, 1]$) est multipliée par $\min(|S_i|/|J_j|, 1)$ dans une seconde étape de normalisation.

Les prédictions finales $Pr(S_i|J_j, CL_k)$ sont alors combinées en utilisant le Théorème des Probabilités Totales (Mood *et al.*, 1974) qui donne la somme pondérée suivante :

$$Pr(S_i|J_j) \approx \sum_{k=1}^n Pr(S_i|J_j, CL_k) \cdot Pr(CL_k) \quad (4)$$

La probabilité $Pr(CL_k)$ reflète la confiance et donc l'importance que l'on donne au classificateur CL_k . Dans la suite, nous avons simplement utilisé $Pr(CL_k) = \frac{1}{n}$ pour $1 \leq k \leq n$, considérant donc que tous les classificateurs sont équiprobables. Nous présentons dans la section 4 différents classificateurs en mettant l'accent sur un nouveau classificateur utilisant pleinement la structure et la sémantique des définitions des concepts et des relations de l'ontologie.

4 Les différents classificateurs utilisés

Nous commençons par présenter quelques classificateurs classiques pouvant être utilisés dans notre cadre (section 4.1). Cependant, la plupart de ces classificateurs nécessitent, pour fonctionner, d'avoir des instances des deux ontologies que l'on cherche à comparer. Nous introduisons donc un nouveau classificateur n'ayant pas ce pré-requis et utilisant pleinement la sémantique des concepts et des relations définis dans les ontologies (section 4.2). Nous montrons finalement à l'aide d'un exemple son fonctionnement et nous discutons de son implémentation (section 4.3).

4.1 Les classificateurs classiques

4.1.1 Même nom de classe ou de relation

Ce classificateur binaire CL_S retourne le poids 1 si et seulement si les deux concepts (ou relations) que l'on cherche à comparer ont le même nom ou le même *stem* (la racine du terme privée de ses suffixes), et 0 dans les autres cas :

$$w(S_i, J_j, CL_S) = \begin{cases} 1 & \text{si } S_i, J_j \text{ ont le même stem,} \\ 0 & \text{sinon} \end{cases}$$

4.1.2 Classifieur kNN

L'algorithme des k plus proches voisins (kNN) repose sur le calcul des distances entre une forme inconnue et toutes les formes d'une base de référence. Il est particulièrement populaire pour la classification de textes (Sebastiani, 2002). Dans notre classifieur CL_{kNN} , chaque concept ou relation S_i représente ainsi une catégorie et des ensembles d'entraînement sont formés à partir des instances i' de S_i :

$$Train = \bigcup_{l=1}^s \{(S_l, i'): i' \in S_l\}$$

Une variante probabiliste du produit scalaire est alors utilisée pour calculer les valeurs de similarités. Soient t et t' des ensembles de mots, $Pr(m|S_i)$ et $Pr(m|J_j)$ sont calculés comme la fréquence normalisée de ces mots dans les instances :

$$RSV(t, t') = \sum_{m \in t \cap t'} Pr(m|S_i) \cdot Pr(m|J_j)$$

4.1.3 Classifieur naïf de Bayes

Le classifieur naïf de Bayes CL_B est lui aussi très prisé pour la classification de textes (Sebastiani, 2002). De la même manière que précédemment, chaque concept ou relation S_i représente une catégorie, et leurs instances sont considérées comme des ensembles de mots (la fréquence des mots est normalisée pour les estimations de probabilité). La formule de calcul est finalement :

$$w(S_i, J_j, CL_B) = Pr(S_i) \cdot \sum_{x \in J_j} \prod_{m \in x} Pr(m|S_i)$$

4.2 Le classifieur structurel et sémantique

Outre ces algorithmes bien connus en classification, nous introduisons un autre classifieur, CL_{Sem} , capable d'utiliser pleinement la sémantique des définitions OWL et guidé par la syntaxe de ces définitions. Son utilisation dans notre cadre peut s'effectuer a posteriori (auquel cas, il utilisera en entrée le résultat des autres classificateurs), ou indépendamment (dans le cas où les ontologies n'ont pas d'instances par exemple). Dans la suite, nous notons $Pr'(A_T|A_S, \Sigma)$ (resp. $Pr'(R_T|R_S, \Sigma)$) la probabilité de la règle $A_T(x) \leftarrow A_S(x)$ (resp. $R_T(x, y) \leftarrow R_S(x, y)$) estimée par les autres classificateurs, où A_S (resp. A_T) est un nom de concept de l'ontologie source (resp. cible), et R_S (resp. R_T) est un nom de propriété de l'ontologie source (resp. cible). Nous rappelons également que la probabilité $\alpha_{j,i}$ d'une règle $T_j(x) \leftarrow S_i(x)$ est donnée par :

$$\alpha_{j,i} = Pr(T_j|S_i) = Pr(S_i|T_j) \cdot \frac{Pr(T_j)}{Pr(S_i)} = Pr(S_i|T_j) \cdot \frac{|T_j|}{|T_j(I, \Sigma)|}$$

La définition formelle et récursive de CL_{Sem} est alors :

- Si R_S et R_T sont des noms de propriétés :

$$w(R_S, R_T, \Sigma) = \begin{cases} 0 & \text{si } R_T \leftarrow R_S \notin \Sigma \\ Pr'(R_T|R_S, \Sigma) & \text{sinon} \end{cases}$$

2. Si A_S et A_T sont des noms de concepts : soit $\mathcal{D} = \mathcal{D}(A_S) \times \mathcal{D}(A_T)$ ⁸

$$w(A_S, A_T, \Sigma) = \begin{cases} 0 & \text{si } A_T \leftarrow A_S \notin \Sigma \\ Pr'(A_T | A_S, \Sigma) & \text{si } |\mathcal{D}| = 0 \text{ et } A_T \leftarrow A_S \in \Sigma \\ \frac{1}{(|\mathcal{D}|+1)} \cdot \left(Pr'(A_T | A_S, \Sigma) + \max_{Set} \left(\sum_{(C_i, D_j) \in Set} w(C_i, D_j, \Sigma) \right) \right) & \text{sinon} \end{cases}$$

3. Soit $C_S = (QR.C)$ et $D_T = (Q'R'.D)$, où Q, Q' sont les quantificateurs \forall ou \exists ou une restriction sur la cardinalité, R, R' sont des noms de propriétés et C, D des expressions, alors :

$$w(C_S, D_T, \Sigma) = w_Q(Q, Q') \cdot w(R, R', \Sigma) \cdot w(C, D, \Sigma)$$

4. Soit $C_S = (op C_1 \dots C_m)$ et $D_T = (op' D_1 \dots D_m)$, où les concepts C_S, D_T sont en notation préfixée et op, op' sont des constructeurs de concept parmi \sqcap, \sqcup, \neg et $n, m \geq 1$, alors :

$$w(C_S, D_T, \Sigma) = w_{op}(op, op') \cdot \frac{\max_{Set} \left(\sum_{(C_i, D_j) \in Set} w(C_i, D_j, \Sigma) \right)}{\min(m, n)} \quad \text{avec}$$

- $Set \subseteq \{C_1 \dots C_m\} \times \{D_1 \dots D_n\}$, et $|Set| = \min(m, n)$,
- $(C, D) \in Set, (C', D') \in Set \Rightarrow C \neq C', D \neq D'$

Plutôt que de prendre la valeur maximale, la valeur moyenne peut être utilisée pour simplifier les calculs. En effet, dans ce cas là, l'estimation du classifier est donnée par la résolution d'un seul système d'équations linéaires, alors que dans le premier cas, il faut résoudre autant de systèmes qu'il y a de choix pour Set . Des valeurs pour les poids w_{op} et w_Q sont données à titre indicatif dans le tableau 2.

w_{op} est défini par :

	\sqcap	\sqcup	\neg
\sqcap	1	1/4	0
\sqcup		1	0
\neg			1

w_Q est défini par :

	\exists	\forall	
\exists	1	1/4	$\leq n$
\forall		1	$\geq n$
$\leq m$		1	1/3
$\geq m$			1

TAB. 2 – Valeurs possibles pour les poids w_{op} et w_Q

4.3 Exemple et implémentation

Nous illustrons maintenant le fonctionnement de ce classifier en reprenant une version simplifiée de l'exemple donné dans le tableau 1. Supposons que les autres classifiers aient fournis les estimations suivantes :

$$\begin{aligned} w(town, city, \Sigma) &= Pr(city | town, \Sigma) &= 0.8 \\ w(id, name, \Sigma) &= Pr(name | id, \Sigma) &= 1 \\ w(place, location, \Sigma) &= Pr(location | place, \Sigma) &= 0.6 \end{aligned}$$

⁸ $\mathcal{D}(A_S)$ représente l'ensemble des concepts directement parents de A_S .

Le classifieur CL_{Sem} estimera $w(Directions, Address, \Sigma)$ par :

$$\begin{aligned} w(D_S, A_T, \Sigma) &= \frac{1}{2} \cdot (\alpha + w_Q(\leq 1, \leq 1) \cdot w(town, city, \Sigma) \cdot w(string, string, \Sigma)) \\ &= \frac{1}{2} \cdot (\alpha + 1 \cdot 0.8 \cdot 1) = \frac{\alpha + 0.8}{2} = 0.9 \text{ (si } \alpha = 1) \end{aligned}$$

Nous définissons ensuite Set sur $\{C_1 \dots C_m\} \times \{D_1 \dots D_n\}$:

$$\begin{aligned} Set &= \{(Cong_1, Conf_1), (Cong_2, Conf_2)\} \text{ ou} \\ Set &= \{(Cong_1, Conf_2), (Cong_2, Conf_1)\} \end{aligned}$$

On obtient alors :

$$\begin{aligned} w(Cong_1, Conf_1, \Sigma) &= 1 \cdot w(id, name, \Sigma) \cdot w(string, string, \Sigma) &= 1 \\ w(Cong_1, Conf_2, \Sigma) &= 1 \cdot w(id, location, \Sigma) \cdot w(string, Add, \Sigma) &= 0 \\ w(Cong_2, Conf_1, \Sigma) &= 1 \cdot w(place, name, \Sigma) \cdot w(Add, string, \Sigma) &= 0 \\ w(Cong_2, Conf_2, \Sigma) &= 1 \cdot w(place, location, \Sigma) \cdot w(Add, Add, \Sigma) &= \frac{15 \cdot \alpha + 12}{50} \\ &= 1 \cdot 0.6 \cdot \frac{\alpha + 0.8}{2} &= \frac{15 \cdot \alpha + 12}{50} \end{aligned}$$

Finalement, CL_{Sem} estimera $w(Congress, Conference, \Sigma)$ par :

$$\begin{aligned} w(C_S, C_T, \Sigma) &= \frac{1}{2} \cdot \left(\beta + w_{op}(\sqcap, \sqcap) \cdot \frac{\max_{Set} \left(\sum_{(Cong_i, Conf_j) \in Set} w(Cong_i, Conf_j) \right)}{\min(2, 2)} \right) \\ &= \frac{1}{2} \cdot \left(\beta + 1 \cdot \frac{\max \left((1 + \frac{15 \cdot \alpha + 12}{50}), (0 + 0) \right)}{2} \right) \\ &= \frac{1}{2} \cdot \left(\beta + \frac{15 \cdot \alpha + 62}{100} \right) = \frac{100 \cdot \beta + 15 \cdot \alpha + 62}{200} = 0.88 \text{ (si } \alpha = \beta = 1) \end{aligned}$$

où α et β représentent la moyenne pondérée des estimations fournies par les autres classificateurs, et valent 1 si le classifieur sémantique est utilisé seul.

Comme nous l'avons vu dans la section 3.1, notre approche commence par considérer tous les appariements possibles (supposons $m \times n$), et génère ensuite tous les sous-ensembles Σ_k contenant ces règles, soit $2^{m \times n}$. Cette combinatoire est évidemment très coûteuse en temps de calcul. Cependant, il est possible de la réduire fortement en ajoutant des contraintes sur la manière dont les sous-ensemble Σ_k sont formés. Par exemple, nous imposons que chaque ensemble Σ_k contienne exactement une règle appartenant chacune des entités ontologiques, le nombre de Σ_k est alors réduit à $\frac{Max(m,n)!}{|m-n|!}$. D'autres contraintes utilisant plus fortement les taxonomies des ontologies à comparer ont aussi été implémentées pour réduire le nombre de ces sous-ensembles.

5 Travaux existants

Parmi les travaux liés à l'alignement d'ontologies, (Ehrig & Staab, 2004) proposent également de combiner différentes mesures de similarité calculées à partir de règles générales établies à la main. Cependant, outre le fait que ces règles sont relativement discutables, cette approche n'est pas complètement automatique, contrairement à la notre. (Noy & Musen, 2001) ont développé une approche très intéressante : ils partent de paires de concepts qui semblent proches (découverts

de façon automatique ou proposés manuellement) et calculent leur similarité *hors contexte* (les ancrés de la recherche, peuvent être éloignées) en étudiant les chemins (dans la taxonomie) qui relient les paires de concepts. Cette méthode, implémentée dans l'outil ANCHOR-PROMPT, affiche pour l'instant les meilleurs résultats. (Euzenat & Valtchev, 2004) ont adapté, eux, des travaux sur le calcul de similarité dans la représentation de connaissance par objets aux langages web actuels pour représenter des ontologies. Une mesure de similarité globale prenant en compte l'ensemble des caractéristiques du langage OWL-Lite est ainsi proposée, capable à la fois de traiter les définitions circulaires et les collections. Pour un état de l'art complet sur les différentes méthodes proposées jusqu'à présent, nous renvoyons le lecteur à (KW, 2004).

Ces techniques d'alignement d'ontologies étant donc relativement différentes, il est difficile de les comparer théoriquement. Une approche empirique, sous la forme de *benchmarks* communs, est préférable pour évaluer tous ces algorithmes. C'est dans cet esprit qu'a eu lieu le challenge EON *Ontology Alignment Contest* fournissant des jeux de tests communs avec les alignements à deviner (Sure *et al.*, 2004). Bien qu'hétérogènes, les résultats fournis par les participants sont très intéressants et de nouvelles idées ont été proposées pour améliorer les futures compétitions du même type. Signalons enfin les travaux de (Euzenat, 2004) qui propose une API commune pour exprimer les résultats des méthodes d'alignement d'ontologies. Nous utilisons également ce format ce qui devrait faciliter les futures comparaisons avec les autres approches.

6 Conclusion et perspectives

Dans cet article, nous avons introduit un cadre logique pour comparer automatiquement des ontologies formelles. Les alignements possibles y sont définis comme des règles dans Datalog probabiliste. Nous avons également présenté un nouveau classifieur utilisant pleinement la sémantique des définitions OWL-Lite des concepts et des relations et des axiomes de l'ontologie. L'implémentation de ce classifieur et son intégration dans le cadre global est en cours de finalisation, mais les premiers résultats sont déjà très encourageants.

Il nous reste désormais à prendre en compte la définition des concepts en extension (par l'énumération de ses instances) pour que notre approche puisse considérer tout OWL-DL. Une autre piste à l'étude consiste à ajouter des classificateurs supplémentaires (utilisation de ressources terminologiques ou d'autres mesures de distance) et à multiplier les tests en les combinant. Nous planifions enfin de participer activement aux prochaines campagnes d'évaluation des différentes approches existantes pour l'alignement d'ontologies, dans le même esprit que le dernier atelier EON (Sure *et al.*, 2004) par exemple.

Références

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The Semantic Web.

- Scientific American*, **284**(5), 34–43.
- CHARLET J., BACHIMONT B. & TRONCY R. (2005). Ontologies pour le Web sémantique. *Revue I3, numéro HS “Web sémantique”* (à paraître).
- DOAN A., MADHAVAN J., DHAMANKAR R., DOMINGOS P. & HALEVY A. (2003). Learning to Match Ontologies on the Semantic Web. *The VLDB Journal*, **12**(4), 303–319.
- EHRIG M. & STAAB S. (2004). QOM - quick ontology mapping. In *3rd International Semantic Web Conference (ISWC'04)*, p. 683–697, Hiroshima, Japon.
- EUZENAT J. (2004). An API for ontology alignment. In *3rd International Semantic Web Conference (ISWC'04)*, p. 698–712, Hiroshima, Japon.
- EUZENAT J. & VALTCHEV P. (2004). Similarity-based ontology alignment in OWL-Lite. In *15th European Conference on Artificial Intelligence (ECAI'04)*, p. 333–337, Valence, Espagne.
- FAGIN R., KOLAITIS P. G., MILER R. J. & POPA L. (2003). Data Exchange : Semantics and Query Answering. In *9th International Conference on Database Theory (ICDT'03)*, p. 207–224, Sienne, Italie.
- FUHR N. (2000). Probabilistic Datalog : Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, **51**(2), 95–110.
- HORROCKS I., PATEL-SCHNEIDER P. F. & VAN HARMELEN F. (2003). From SHIQ and RDF to OWL : The making of a web ontology language. *Journal of Web Semantics*, **1**(1), 7–26.
- KW C. (2004). State of the Art on Ontology Alignment. Deliverable Knowledge Web 2.2.3, FP6-507482.
- LENZERINI M. (2002). Data integration : A theoretical perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'02)*, p. 233–246, Madison, Wisconsin, USA.
- MOOD A. M., GRAYBILL F. A. & BOES D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.
- NOTTELMANN H. & STRACCIA U. (2005). sPLMap : A probabilistic approach to schema matching. In *27th European Conference on Information Retrieval (ECIR '05)*, p. 81–95, Santiago de Compostela, Espagne.
- NOY N. F. & MUSEN M. A. (2001). Anchor-PROMPT : Using non-local context for semantic matching. In *Workshop on Ontologies and Information Sharing at IJCAI'01*, Seattle, Washington, USA.
- OWL (2004). Web Ontology Language Reference. W3C Recommendation (10 Février). <http://www.w3.org/TR/owl-ref/>.
- SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.
- Y. SURE, O. CORCHO, J. EUZENAT & T. HUGHES, Eds. (2004). *3rd International Workshop on Evaluation of Ontology-based Tools (EON'04)*, Hiroshima, Japon.

Introduction aux ontologies sémiotiques dans le Web Socio Sémantique

Manuel Zacklad¹

¹Equipe Tech-CICO / ISTIT FRE CNRS 2732
Université de Technologie de Troyes

Manuel.Zacklad@utt.fr

Résumé : Dans cette communication, nous présentons le positionnement du *Web Socio Sémantique* (W2S) par rapport au Web Sémantique (WS) et nous discutons en particulier la problématique de l'ingénierie des *ontologies sémiotiques*. Nous présentons différents critères pour l'établissement des *accords définitionnels* dans la construction des ontologies issus des approches *logique, contextuelle et situationnelle* (ou pragmatique).

Mots-clefs : Web Socio Sémantique, Web Sémantique, Ontologie, Ontologie Sémiotique, Coopération, Transaction Communicationnelle, Document

1 Introduction

Les infrastructures du Web dont l'importance ne cesse de croître répondent à la fois à des besoins de communication et de partage d'information. Relevant du premier besoin, on trouve des applications variées allant de la messagerie à la visio-conférence en passant par le chat et la téléphonie. Relevant du second besoin on trouve des applications elles aussi très diverses allant des sites Web statiques aux bibliothèques numériques de grandes tailles en passant par les weblogs et les sites interactifs de toutes natures offrant des accès à de nombreuses applications administratives (au sens des « legacy system » anglo-saxons). Dans cette typologie, les forums constituent certainement l'intermédiaire le plus fameux entre communication à distance et mise à disposition publique du corpus constitué par des interactions scripturales.

Portées par des ambitions d'ingénierie, différentes visions du Web ont émergé avec un succès plus ou moins grand. La plus célèbre est celle du Web Sémantique bien connue en ingénierie des connaissances et résolument tournée vers les besoins informationnels. Malgré les ambitions assez larges affichée par Tim Berners Lee (Berners Lee et al 2001) qui visait autant la médiation des activités humaines que la recherche automatique d'information, force est de constater que l'essentiel des efforts de recherche se réclamant de cette approche ont porté sur les moyens d'une formalisation logique d'un certain nombre de contenus ou de systèmes d'index de ces contenus permettant à des agents logiciels de répondre automatiquement aux requêtes complexes de leurs utilisateurs. Le projet actuel du Web Sémantique est avant tout

celui d'un Web Sémantique Formel principalement tournées vers les besoins d'exploitation automatiques servis par des programmes informatiques.

Moins structuré, mais correspondant à un nombre très important d'applications, le Web social, parfois associé au « social computing », s'adresse aux besoins de communication en visant des usages associés à la conduite d'interactions éphémères entre utilisateurs distants tout en offrant des représentations, souvent de nature graphique, des réseaux sociaux ainsi constitués.

Enfin, la troisième vision du Web, dans laquelle nous nous inscrivons, est celle du Web Socio Sémantique (W2S) (Zacklad 2003c, Cahier et al. 2003) une approche qui combine les deux précédentes en postulant l'existence d'une co-détermination des usages informationnels et communicationnels du Web. Cette vision est notamment promue actuellement par l'équipe Tech-CICO à l'Université de Technologie de Troyes, une équipe spécialisée dans l'étude des usages coopératifs des TIC, en relation avec d'autres communautés comme le groupe « Pratiques Collectives Distribuées » dont l'animateur, Bill Turner, est au LIMSI (CNRS). Le Web Socio Sémantique prolonge pour partie les premières critiques que nous avions formulées à l'égard de certaines attitudes hégémoniques du courant du Web Sémantique Formel, en proposant l'approche alternative du Web Cognitivement Sémantique (Caussanel et al. 2002).

Alors que les travaux se réclamant du *Web Sémantique* cherchent surtout à développer les possibilités d'inférence automatique d'agent logiciels portant sur des contenus stables et assez simples pour être décrits en logique formelle, le *Web Cognitivement Sémantique*, vise à soutenir les activités de recherche d'utilisateurs humains dans des corpus complexes et évolutifs. Tout en prolongeant cette approche, le *Web Socio Sémantique* (W2S) inscrit les pratiques de recherche et d'élaboration informationnelle des usagers du Web dans des activités de « coopération structurellement ouvertes » (Zacklad 2003a) qui reposent également sur des pratiques communicationnelles intensives. Les collectifs visés sont de taille très variée comme l'est leur degré de structuration qui va de communautés de pratiques informelles à des projets sous pression temporelle en passant par des communautés d'action. Leur caractéristique commune principale est qu'une partie de leur activité est « distribuée », spatialement, temporellement voire socialement, ce qui justifie leur usage d'un système d'information et de communication (cf. infra).

2 Un cadre théorique pour l'analyse des activités collectives distribuées

Dans la théorie psycho-socio-économique d'orientation pragmatique que nous sommes entrain de développer, les activités interpersonnelles, notamment professionnelles, sont analysées sous l'angle du déroulement de *transactions communicationnelles symboliques* (Zacklad 2003b, 2004). Celle-ci sont réalisées dans des *situations transactionnelles* caractérisées par la présence de réalisateurs et bénéficiaires poursuivant un projet commun dans un cadre spatio-temporel et des conditions environnementales donnés en fonction des relations sociales qui les lient (Zacklad 2004).

Du fait de la distribution des activités collectives, réalisateurs et bénéficiaires des transactions communicationnelles peuvent ne pas être présents dans le même cadre spatio-temporel. Cela implique la mise en place de stratégies permettant de prolonger ces transactions de manière à ce qu'elles puissent être initialisées, interrompues, réactualisées, répétées, dans toutes les configurations de présence absence du bénéficiaire et du réalisateur. Nous définissons ainsi plusieurs stratégies de distribution spatio-socio-temporelle des transactions : la *normalisation de la situation transactionnelle*, la *formalisation de l'expression*, la *ritualisation mnémotechnique*, *l'abstraction*, la *médiation substitutive* et la *documentarisation*.

Le Web Socio Sémantique est ainsi surtout appréhendé comme visant à faciliter la prolongation et la préservation des transactions communicationnelles à l'intérieur de communautés de pratiques ou d'action spatio-socio-temporellement distribuées. Alors que l'approche du Web Social priviliege souvent la *médiation substitutive* pour faciliter les interactions synchrones à distance sans recourir à des supports pérennes, les activités coopératives des usagers du Web Socio Sémantique les amènent le plus souvent à privilégier la stratégie de *documentarisation* stratégie consistant à transcrire ou à enregistrer une production sémiotique sur un support pérenne puis à l'équiper « *d'attributs spécifiques visant à faciliter les pratiques liées à son exploitation ultérieure dans le cadre de la préservation de transactions communicationnelles distribuées* » (Zacklad 2004).

Ce faisant, ils s'appuient sur de nouveaux types de média leur permettant de concilier l'interactivité et la pérennité offerte par les supports électroniques que nous avons appelés des « Documents pour l'Action » (Zacklad 2004). Les forums, les usages publics de la messagerie, les systèmes d'annotation disponibles à partir des éditeurs de texte ou de sites Web (blogs, wikis, mais également annuaires évoluant sur la base des contributions de participants...) en constituent les exemples les plus connus. Les usagers des applications relevant du W2S, dont les transactions sont médiatisées par des Documents pour l'Action, sont ainsi confrontés à la gestion d'un nombre très important de documents évolutifs, souvent composés de fragments multiples non nécessairement intégrés, pris en charges par des auteurs différents dont les droits et les prérogatives à l'égard des fragments sont eux-mêmes sujets à négociation.

Face à cette diversité et à la complexité des relations entre les documents et souvent entre les fragments eux-mêmes (les annotations), le Web Socio Sémantique doit proposer des outils facilitant l'investissement documentaire de ses usagers et donc la qualité des transactions ultérieures. Si les moteurs de recherche offrent indéniablement une technologie de premier plan, celle-ci n'est généralement pas suffisante dans le contexte du W2S. En effet, la proximité sémantique des fragments, des documents et des dossiers, est potentiellement élevée et l'objectif est avant tout de les discriminer finement sur la base d'attributs aussi divers que les auteurs impliqués, le contexte de production et les thématiques abordées dans tel ou tel fragment. Qui plus est, la dimension coopérative de l'activité des usagers, signifie que les transactions rassemblent des acteurs divers dont les compétences sont parfois hétérogènes, ce qui implique le recours à des systèmes de classification partagés par le collectif et représentées explicitement dans le système et son interface que nous

appelons des *ontologies sémiotiques*. Loin de s'opposer à l'utilisation d'un moteur de recherche, la mise à disposition d'une ontologie sémiotique est en mesure d'accroître la précision des réponses fournies.

3 Critères pour l'établissement des accords définitionnels dans la construction des ontologies sémiotiques

Les ontologies sémiotiques sont des productions sémiotiques cohérentes qui regroupent des expressions stéréotypées extraites des transactions communicationnelles et organisées selon des axes paradigmatic et syntagmatique. Les expressions sélectionnées sont considérées comme des *concepts sémiotiques* à l'issu d'un processus *d'investissement définitionnel*. Cet investissement définitionnel, qui peut se limiter à un agencement différentiel¹ de concepts sémiotiques dans des hiérarchies de proximités sémantiques basées sur leur fonction opératoire dans la situation transactionnelle, a pour objectif de faciliter la résolution des problèmes rencontrés dans cette situation. L'aide apportée par une ontologie sémiotique dans la recherche et la sélection d'informations comme dans la navigation sémantique est ainsi toujours relative à une classe de situations problèmes dans un cadre transactionnel donné qui justifie la cohérence d'ensemble de l'ontologie².

La construction d'un concept sémiotique peut être plus ou moins délibérée, mais son intégration dans l'ontologie fait toujours l'objet d'un *accord définitionnel* explicite, contrôlé par un individu ou une communauté responsable de sa signification opératoire. L'accord définitionnel consiste à associer à un ensemble de signes un contenu sémiotique, c'est-à-dire à s'accorder sur une représentation commune et sur les effets que celle-ci est susceptible de déclencher. Il existe différentes modalités de l'accord qui renvoient plus ou moins aux grandes tendances utilisées par la sémantique pour aborder le signifié (par exemple Mounin 1968, p. 159). Celles-ci peuvent être caractérisées comme relevant de théories *logiques*, *contextuelles*, *situationnelles* (ou pragmatiques) du signifié même si ces approches visent souvent à caractériser des mots alors que nous visons principalement le signifié d'expressions.

L'approche logique qui est privilégiée dans les ontologies formelles, veut asseoir la signification sur des critères qui soient à la fois indépendant du contexte sémiotique de l'expression (les autres signes composants la production sémiotique) et de la situation transactionnelle. L'accord définitionnel s'appuie sur la sélection d'un petit nombre de sèmes, pour partie hérités du graphe dans lequel se trouve le concept, qui confère à l'expression (au mot dans le cas du Web sémantique) une sorte « d'autonomie logique » lui permettant ainsi de se combiner librement avec d'autres expressions dans des arrangements syntaxiques inédits. L'accord sur les sèmes considérés comme pertinents peut-être sensé refléter une raison universelle, comme

¹ On discutera plus loin de la distinction entre les ontologies différentielles de Bruno Bachimont (2004) et les ontologies sémiotiques.

² On reproche parfois aux thésaurus leur incohérence... cela ne n'est pas le cas pour les ontologies sémiotiques qui s'appuient sur des critères de classification explicites organisés en points de vue et justifiés par des situations problèmes génériques.

c'est le cas dans certaines tendances du Web sémantique, ou peut-être le fait d'une communauté plus restreinte. Ce qui caractérise l'approche logique c'est la revendication d'une indépendance de la signification vis-à-vis du contexte de la production sémiotique c'est-à-dire vis-à-vis de la situation transactionnelle³.

L'approche contextuelle ancre largement la signification de l'expression dans celle de la production sémiotique envisagée dans sa globalité, ou tout du moins, dans les parties de la production sémiotique, texte ou discours par exemple, qui « font sens ». En effet, comme l'expliquent les traducteurs professionnels⁴, non seulement toute traduction mot à mot est impossible mais le plus souvent toute traduction phrase à phrase l'est également. C'est en considérant des fragments signifiant de texte eux-mêmes délimités par les structures « formelles » du document plus ou moins flexibles (saut de paragraphe, chapitre...) qu'une délimitation des unités signifiantes peut être effectuée.

L'accord définitionnel s'inscrit ici dans le contexte d'une production sémiotique donnée ou d'un « genre » de production sémiotique et s'appuie sur les relations de l'expression avec d'autres expressions voisines ou éloignées (selon leur proximité sémantique) sans qu'une analyse sémique ait nécessairement à être conduite. Prise au sérieux, l'approche contextuelle vient également relativiser l'importance de la syntaxe. La pertinence des combinaisons d'expressions dépend du genre mais dépend également de la proximité d'autres expressions dans une aire de signification donnée. La syntaxe apparaît d'avantage comme un soutien heuristique à l'interprétation que comme une condition préalable de toute signification.

De la même façon, l'approche contextuelle tendrait à remettre en cause l'indépendance de la signification de l'expression vis-à-vis de la matérialité de son support ou du moins de la manière dont le support serait appréhendé perceptivement dans le cas du recours à une médiation substitutive. En effet, les indices permettant d'interpréter la proximité sémantique sont souvent eux-mêmes reliés à des caractéristiques de proximité spatiale sur le support, voire de facilité manipulatoire associée aux activités de navigation sémantique dans celui-ci.

L'approche situationnelle (ou pragmatique) ancre la signification dans la situation transactionnelle elle-même et dans ses différentes composantes. Les expressions ne sont pas seulement rapportées à l'environnement interne de la production sémiotique considérée, à leur proximité sémantique, mais également aux paramètres de la situation transactionnelle dans leur diversité : projet commun, nature des relations sociales entre les participants et caractéristiques de ces derniers, cadre spatio-temporel et conditions environnementales, terrain représentationnel commun... Selon cette approche, la signification se construit sur la base d'une « pratique sociale continue, au cours de milliers et de milliers d'expériences d'usage des termes en situation, corrigées successivement par la méthode des essais et erreurs » (Mounin 1968, p. 163).

Les accords définitionnels s'ancrent dans une approche fonctionnelle de la signification dépendant des caractéristiques de la situation. Il ne s'agit d'ailleurs pas

³ Nous expliquons plus bas que l'effectivité du Web Sémantique formel ne réside pas dans cette indépendance impossible à réaliser mais dans un usage limité à des situations transactionnelles très standardisées.

⁴ Monique Slodzian, communication personnelle, voir aussi (Rastier, 2004).

d'une approche *référentielle* qui rabattrait le sens sur l'existence « choses autonomes dans un monde objectif » servant de référent indépendamment des activités coopératives des acteurs et de leurs pratiques linguistiques de dénomination dans le cadre des situations transactionnelles problématiques (cf. Dewey et Bentley 1949). L'existence des « choses » en tant qu'entités appréhendables par la pensée et la langue dépend des projets poursuivis par les acteurs, projets dont le déroulement entraîne une dénomination des ingrédients des situations problématiques à l'aide d'expressions, expressions qui structureront la pensée et qui influenceront en retour le déroulement de l'activité.

Dans l'approche situationnelle, l'accord définitionnel, la signification de l'expression, dépend donc de la compréhension de la finalité de l'action dans laquelle elle est employée, de son réalisateur (énonciateur dans le cas du discours), du contexte matériel de la production, des relations sociales entre les réalisateurs et les bénéficiaires, etc... Selon que l'expression sera utilisée dans une transaction visant à produire des soins, assembler un dispositif technique, juger un prévenu, former un élève, piloter un navire, recruter un collaborateur, convaincre des électeurs potentiel, vendre un produit... sa définition ne sera pas la même et il n'y aura pas de compréhension possible entre des partenaires sans une expérience, même indirecte, de ces situations de référence.

Selon cette approche, dans une situation donnée, l'expression n'aura pas le même sens selon qu'elle émanera de l'un ou l'autre des participants, voire même n'aura tout simplement aucun sens. Ainsi, l'expression « Vous êtes embauché ! » émanant du candidat à l'emploi, est une expression *a priori* sans signification dans les transactions associées au recrutement. Son interprétation nécessiterait la référence à une autre situation transactionnelle où les rôles seraient inversés.

Dans l'approche situationnelle, l'investissement ontologique implique une entente entre les participants quant à la fonction des expressions dans une situation transactionnelle donnée, c'est-à-dire quant à leur contenu sémiotique effectif, leur pouvoir évocation et leurs effets potentiels. C'est non seulement l'autonomie combinatoire des expressions vis-à-vis de la syntaxe et du support de la production sémiotique, c'est-à-dire du média, qui est mise à mal, mais encore leur autonomie par rapport à l'ensemble des paramètres des situations transactionnelles dont elles sont issues et auxquelles elles sont destinées.

4 Fermeture sémiotique vs ouverture sémiotique, formalité machinale vs formalité sémiotique

En conclusion, nous présentons une hypothèse provocatrice selon laquelle l'effectivité de certaines applications relevant du Web Sémantique ne tient pas à l'approche « logique » de la signification, ancrée dans une syntaxe formelle et une sémantique référentielle, pourtant souvent mise en avant, mais à la fermeture sémiotique opérée dans les applications, fermeture sémiotique qui ne fonctionne qu'en regard de la fermeture corollaire des situations transactionnelles de référence. Cette fermeture sémiotique est également celle qui permet à la plupart des applications

automatisées de fonctionner efficacement sans qu'aucune référence ne soit faite à la « logique formelle ». Or, paradoxalement, cette caractéristique semble largement ignorée par les promoteurs du Web Sémantique Formel qui, en s'identifiant au courant logiciste du programme de l'Intelligence Artificielle, imputent leur effectivité à la remonté vers les couches hautes du célèbre cake, couches de la preuve et de la vérité (sic !).

Quelle est la condition de l'effectivité des calculs réalisés par les programmes informatiques dans les situations sociales ou techniques dans lesquelles ils sont employés dès lors que l'utilisation visée n'est pas un soutien heuristique à la réflexion mais une prise de décision automatisée ? Elle dépend entièrement de la fermeture de la situation transactionnelle dans laquelle le programme informatique transmettra des informations et appliquera des règles décisionnelles. Comme le savent tous les programmeurs, c'est uniquement à condition d'avoir listé toutes les alternatives possibles, que celles-ci résultent d'un ensemble de règles stockées dans une base de données ou d'un calcul basé sur des données issues de capteurs, qu'un programme ne « boguera » pas, dans un sens étroitement technique ou lié aux implications pragmatiques associées aux dégâts consécutifs pouvant être causés dans la situation d'usage.

Du point de vue de la situation transactionnelle, ce fonctionnement renvoie à des situations interactionnelles dites « fermée » dans lesquelles la dimension sémiotique de la communication s'efface devant le recours à des automatismes comportementaux. Le signe est alors réduit à un signal qui déclenche une réponse conditionnée sans recours à l'interprétation. Alors que dans une base de donnée automatisée la sémiotique des expressions utilisées est fermée, dans une base documentaire, les activités de recherche et de navigation doivent pouvoir s'appuyer sur des signes dont la sémiotique est ouverte en cohérence avec l'ouverture des situations transactionnelle associées. La recherche et la navigation tirera donc partie d'expressions du domaine à la fois pertinentes et contrôlées, des « concepts sémiotiques », organisés dans des ontologies du même nom et permettant des parcours guidées par des finalités, certes identifiées, mais non réductible à des procédures automatisées.

Ces concepts sémiotiques et leurs ontologies relèvent bien d'une formalisation. Mais cette formalisation est celle donnée à l'expression pour accroître son ouverture sémiotique vers le contexte et la situation au moyen de l'investissement définitionnel dans un cadre transactionnel donnée. La *formalité sémiotique* vient ouvrir le sens en multipliant les ancrages possible du concept sémiotique à partir de différents points de vue tandis que la *formalité machinale* doit fermer le sens pour que le signe devienne autant que possible un signal univoque. Dans la formalité machinale, la syntaxe est rigide, les signes sont précisément dénombrés et infiniment combinables et leur organisation sur le support totalement indifférente.

Ainsi, ce n'est pas la formalité sémiotique qui explique l'effectivité des « langages syntaxiquement formels » car ceux-ci ne possèdent aucun potentiel sémiotique, mais la standardisation des situations transactionnelles de référence, celles dont ils sont issus et auxquelles ils se destinent (une vision proche des réflexions du deuxième Wittgenstein). L'effectivité de ces « langages » renvoie en fait à l'automatisme

comportemental qui fait que le signe est traité comme un stimulus attendant une réponse conditionnée. Cette effectivité est elle-même renforcée par les sanctions sociales qui permettent de normaliser les comportements : l'énoncé $2+2=5$ vaut une sanction à l'élève et un licenciement au technicien de même que l'emploi inconsidéré d'une carte bleue dans les transactions commerciales envoie l'usager inconsistant en prison.

C'est dans l'automaticité des comportements conditionnés et des sanctions que réside le secret de la formalité non sémiotique, de la formalité machinale (littéralement de formules qu'il nous faut interpréter comme des machines, qui n'est pas celle formules des mathématiques avancées où le *style* de la démonstration importe). On ne contestera d'ailleurs pas l'intérêt de cette automaticité pour un fonctionnement organisationnel ordonné et le développement de pratiques d'ingénierie efficaces et fiables, mais on se souviendra que la force de la formalité syntaxique repose sur celle des conventions sociales qui restreignent le potentiel sémiotique des signes dans certaines situations transactionnelles et pas sur une quelconque preuve ou vérité résultant des activités introspectives des « ontologies formels » accédant aux couches hautes du Web Sémantique.

5 En conclusion : positionnement par rapport aux autres tentatives de dépassement du WS formel⁵

Le rejet théorique des ontologies

La thèse «du rejet théorique des ontologies» défendue par F. Rastier (1999 et 2004) s'appuie sur une vaste mise en perspective épistémologique et vise à la fois les ontologies formelles au sens strict et certains usages des réseaux sémantiques (F. Rastier considère d'ailleurs avec raison, selon nous, que la plupart des réseaux sémantiques actuellement construits sont nommés ontologies). Bien que nous ne puissions présenter ici une discussion des thèses riches mais complexes de cet auteur il nous semble que nous pourrions être assez en accord avec la plupart de ses propositions épistémologiques. Notre appel aux «ontologies sémiotiques», schémas de classification constitués d'expressions contextualisées dédiés à des types de problèmes impliquant l'exploration systématique de corpus, diffère assez profondément des positions philosophiques traditionnelles invoquées lors du recours aux ontologies formelles. Reste notre emploi de l'expression «d'ontologie sémiotique» alors que F. Rastier plaide pour une sémiotique sans ontologie (1999) qui pourrait sembler problématique dans la perspective de cet auteur. Nos pratiques d'ingénierie nous font préférer un renversement de perspective sur la formalité et les ontologies à une opposition frontale conduisant à détourner de manière radicale des travaux sur les ontologies dans lesquels les approches semi formelles et informelles gagnent progressivement du terrain.

⁵ Nous remercions nos relecteurs pour leurs remarques pertinentes nous ayant incité à ces développements.

Les engagements ontologiques croissants

Cette position défendue par B. Bachimont (2004) est associée à la méthodologie Archonte qui permet de passer progressivement de corpus de textes à des ontologies formelles opérationnelles. Les concepteurs définissent progressivement des ontologies différentielles, référentielles puis computationnelles. Les ontologies différentielles, issues d'une normalisation définitionnelle des termes extraits des corpus, sont relativement proches des ontologies sémiotiques sauf que ces dernières n'ont pas un caractère exclusivement intralinguistique. En effet, alors que les ontologies différentielles semblent principalement relever de l'approche contextuelle de la signification les ontologies sémiotiques relèvent d'une approche situationnelle qui englobe la précédente. Les « différences » entre les concepts sémiotiques tiennent aussi à de multiples critères extra documentaires, notamment représentés dans les différents points de vue, qui ne sont pas uniquement issus d'une analyse de corpus textuels⁶ mais qui reflètent également les débats entre les membres des communautés partie prenantes.

Si nous ne contestons aucunement la pertinence du passage vers des ontologies référentielles et computationnelles dans le cas où il est nécessaire de réaliser une application relevant du WS formel, nous souhaitons ici insister sur la possibilité d'offrir un soutien « informatique » aux utilisateurs des ontologies sémiotiques, sans que ce soutien n'implique un passage par les ontologies référentielles et computationnelles. Dans tous les cas où l'ouverture des situations transactionnelles de référence implique de maintenir une relative ouverture sémiotique, il est souhaitable d'utiliser les ontologies sémiotiques pour guider la navigation à l'intérieur des corpus. Leur plus grande expressivité permet de guider plus efficacement l'utilisateur dans la catégorie pertinente en lui offrant des « prises » ou des « affordances » nombreuses, lors de la recherche d'une information : trouver des lieux touristiques correspondant à un projet culturel, par exemple. A l'inverse, le recours à la formalité machinale et aux ontologies référentielles est nécessaire quand la signification véhiculée par les situations transactionnelles est étroitement confinée : trouver des places d'avion entre deux dates précises, par exemple.

Le syncrétisme indifférencié

Nous avons déjà évoqué le fait que T. Gruber, un des principaux promoteurs des ontologies notamment avec l'initiative formelle Ontolingua, défendait aujourd'hui une vision des ontologies comme étant un « *un traité – un accord social – entre des gens ayant une raison commune de partager* » en accordant une place essentielle aux ontologies non formelles. Tout en étant politiquement en accord avec l'évolution de cet auteur, qui n'est peut être pas sans lien avec les préoccupations liées au développement du logiciel « Intraspect » de la société dont il est un des fondateurs, nous pensons que ses thèses manquent actuellement de cohérence tant en ce qui concerne la relation entre formel et informel qu'en ce qui concerne l'extension

⁶ Il est toujours possible de considérer qu'un débat est un texte, mais la démarche présentée dans Archonte ne permet pas en l'état de véritablement s'appuyer sur une analyse des échanges en contexte.

considérable donnée au domaine des ontologies, terme qui devient un quasi synonyme du terme de modèle.

En ce qui concerne la relation entre formel et informel nous pensons qu'elle est à la fois un peu naïve et contradictoire sur un plan épistémologique. En effet, il nous semble difficile de donner à la fois la primauté à « l'accord social » et de continuer à considérer qu'il existerait une infériorité des langues naturelles utilisées pour la spécification des ontologies informelles, considérées comme vagues ou imprécises, par rapport aux langues formelles qui seules permettent d'atteindre un haut niveau d'expressivité et de sémantisme (cf. Sheth & Ramakrishnan 2003, p. 5 et Gruber 2004 qui considère que les «*parties formelles*» de l'ontologie sont des énoncés qui peuvent être utilisés pour «*pour déduire ou faire appliquer les significations*» p.2).

Bien que nous soyons en accord avec ces auteurs quand ils soulignent les limitations inhérentes aux ontologies formelles pour des raisons de faisabilité et d'utilité (p.e. Sheth & Ramakrishnan 2003 p.6), nous considérons quant à nous que la langue et les ontologies sémiotiques sont porteuses d'un potentiel sémiotique et donc d'une expressivité bien supérieure à celle de formalité machinale des ontologies de cette tendance. D'ailleurs, tout accord social « sur le fond » est toujours réalisé dans le cadre de transactions communicationnelles synchrones et « en présentiel » qui sont la condition d'une réelle intercompréhension et corollairement d'une meilleure « rationalité » (au sens philosophique du terme).

Le deuxième point sur lequel nous nous différencions de ce syncrétisme indifférencié tient à l'extension extrême qui est donnée au domaine des ontologies. Dans sa présentation de 2003, Gruber considérait qu'une ontologie devait de manière nécessaire définir un ensemble de concepts (p. 5) pour assimiler finalement des entrées de catalogue Yahoo, ou CommerceOne à des ontologies (p. 13). Dans son interview de 2004, celui-ci fait un pas supplémentaire en considérant que tout le Web est par nature basé sur une ontologie. La raison en est que «*dans son cœur, le concept d'hyperlien est basé sur un engagement ontologique sur l'identité de l'objet. Pour hyperlier un objet il faut une notion stable de cet objet et que son identité ne dépende pas du contexte (la page sur laquelle je suis actuellement, ou la période de temps, ou qui je suis)* » (p.2).

La machinerie du Web originel est considérée comme reposant sur des documents de spécification débattant de la notion d'identité des objets et constituant donc l'ontologie de celui-ci. Or, en assimilant toute spécification de standard à une ontologie, on perd de vue l'une de leur spécificité qui est, selon nous, de constituer un schéma de classification organisant des éléments de corpus au moins pour partie partiellement préexistant, en s'appuyant, dans le cas des ontologies sémiotiques, sur des points de vue multiples et potentiellement divergents. Peut-être faut-il comprendre la position de Gruber en considérant que, dans un processus de spécification, les délibérations associées à l'examen des alternatives en présence constituent un corpus qui doit être organisé comme dans le courant du « design rationale ». Mais, on doit être alors en mesure de fournir une explicitation classificatoire des alternatives en présence ce qui est loin d'être encore le cas de la plupart des documents de spécification.

Le renversement dialectique de la formalité

La position épistémologique défendue dans l'initiative du W2S et des ontologies sémiotiques consiste à la fois :

- à défendre l'importance des modèles de représentation des connaissances ontologiques qui classifient des corpus divers et facilitent les transactions communicationnelles au sein de communautés distribuées ;
- à considérer que le recours aux ontologies sémiotiques ne relevant pas de la logique formelle est la réponse adaptée pour la construction de ces modèles dès que les signification visées par les corpus ne sont pas réductibles à des conventions simples et univoques dans le contexte de transactions standardisées.

Ainsi, s'il est bien nécessaire de recourir à des « investissements de forme » pour concevoir des ontologies (Thévenot 1985), en l'occurrence des investissements définitoires, ceux-ci relèvent d'une formalité sémiotique qui facilite les traductions entre le sens en contexte actualisé dans les transactions communicationnelles conversationnelles et les significations stabilisées dans le cadre d'accords conventionnels plus larges portés par des communautés de parole, de pratique, d'action ou d'intérêt (en s'inspirant pour partie de la distinction de Rastier 1999).

Cette thèse nous semble partagée par différents auteurs tels que Veltman (2004), défenseur d'un Web Sémantique pour les contenus culturels, ou encore Ribes et Bowker (2004) qui plaident pour un ancrage du travail ontologique dans la sociologie des sciences. K. H. Veltman critique l'épistémologie sous jacente aux langages de représentation formels dans une veine proche de celle de Rastier et plaide pour des accès distribués, dynamiques, multilingues, historiques et culturels à la connaissance, permettant de suivre les changements spatio-temporels rapides des définitions sur des plans locaux, régionaux, nationaux, internationaux. Si la logique formelle est suffisante pour les transactions simples de machine à machine où dans le contexte des routines bien établies du monde des affaires une autre perspective est nécessaire quand la dimension culturelle et symbolique entre en scène.

Ribes et Bowker (2004) défendent quand à eux l'idée selon laquelle, pour réaliser l'alignement épistémologique entre les spécialistes du domaine et les spécialistes en représentation des connaissances, ceux-ci doivent se transformer délibérément en sociologues des sciences pour se rapprocher plus intimement du domaine à représenter. Au lieu d'imposer des catégories formelles a priori ceux-ci sont ainsi invités à les transformer de l'intérieur et à opérer une double traduction, du domaine vers les catégories logiques et des catégories logiques vers le domaine, créatrice de nouveaux points de vue. Mais ce rapport à la logique tend à nous éloigner de l'épistémologie de la logique formelle tant dans son lien à la sémantique qu'en ce qui concerne la place donnée à la syntaxe pour se rapprocher des visions pragmatistes de celle-ci (Dewey 1938).

6 Bibliographie

- BACHIMONT, B (2004). *Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle*. Habilitation à diriger des recherches de l'Université Technologique de Compiègne.
- BERNERS-LEE ET AL. (2001). Berneers Lee T., Hendler J., Lassila O. The Semantic Web, *Scientific American*, May 2001, <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>
- CAHIER ET AL. (2004). Cahier, J.-P., Zacklad, M., Monceaux, A., Une application du Web socio sémantique à la définition d'un annuaire métier en ingénierie, *Actes des journées Ingénierie des Connaissances IC2004*, mai 2004, Lyon.
- CAUSSANEL ET AL. (2002). Caussanel J., Cahier J.-P., Zacklad M., Charlet J., " Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? ", *Conférence Ingénierie des Connaissances IC2002*, Rouen Mai 2002.
- Dewey J., (1938), Logic : *The Theory of Enquiry*, Henry Holt and Company, New York, trad. Fcse. (1993). Logique : La théorie de l'enquête, PUF Paris.
- DEWEY, J. & BENTLEY, A. F. (1949). *Knowing and the known*. In J. A. Boydston (1989), John Dewey: The later works, 1925-1953 (Vol. 16, pp. 2-294). Carbondale: Southern Illinois University Press.
- MOUNIN, G., (1968). *Clefs pour la linguistique*, Seghers, Paris
- RASTIER F. (1999). De la signification au sens – pour une sémiotique sans ontologie, www.revue-texto.net, 1999.
- RASTIER F. (2004). Ontologie(s), *Revue des sciences et technologies de l'information*, série : *Revue d'Intelligence artificielle*, Vol. 18, n°1, 2004, p. 15-40
- RIBES J. & BOWKER G. C. (2005). "Ontologies and the Machinery of Difference," à paraître in *Journal of the Association of Information Systems (JAIS)* for the Special Edition on Ontologies (2005).
- SHETH & RAMAKRISHNAN C. (2003). Semantic (Web) Technology in Action : Ontology Driven Information Systems for Search, Integration and Analysis, *IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real*, Volume 26, Number 4, December 2003, 40-48.
- THEVENOT L. (1985). Les investissements de forme. In : *Conventions économiques*, Paris, CEE/PUF, pp.21-71
- VELTMAN K. H. (2004), Semantic Web for Culture, *Journal of Digital Information*, Volume 4 Issue 4, Article No. 255, 2004-03-15. <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Veltman/>
- ZACKLAD M., (2003a). Un cadre théorique pour guider la conception des collecticiels dans les situations de coopération structurellement ouvertes, in Bonardi, C. GEORGET, P. ROLAND-LEVY, C. ROUSSIAU, *Psychologie Sociale Appliquée, Economie, Médias et Nouvelles Technologies*, In : Press, Coll Psycho, Paris, 135-164.
- ZACKLAD M., (2003b). Transactions communicationnelles symboliques et communauté d'action : réflexions préliminaires, présentation au colloque de Cerisy de septembre 2003 organisé par P. Lorino et R. Teulier, Archive SIC, <http://archivessic.ccsd.cnrs.fr/>
- ZACKLAD ET AL. (2003c). Zacklad, M., Cahier, J.P., Pétard, X. Du Web Cognitivement Sémantique au Web Socio Sémantique, *Journée « Web Sémantique et SHS » du 7 mai 2003*, <http://www.lalic.paris4.sorbonne.fr/stic/as5.html>
- ZACKLAD, M. (2004). Processus de documentarisation dans les Documents pour l'Action (DopA) : statut des annotations et technologies de la coopération associées. *Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire*, 13-15 Octobre 2004, Montréal. Montréal: 19 p.

INDEX DES AUTEURS

Yann Abd-el-Kader	1	Philippe Laublet	25, 157
Iulian Alecu	13	Myriam Lewkowicz	169
Florence Amardeilh	25	Gaelle Lortal	169
Bruno Bachimont	157	Jean-Luc Minel	25
Audrey Baneyx	37	Wilfried Njomgue Sado	181
Catherine Barry	73	Jérôme Nobécourt	97
Giuseppe Berio	49	Henrik Nottelmann	229
Jacques Bouaud	61	David Ouagne	193
Cédric Bousquet	13	Liliane Pellegrin	109
Sandra Bringay	73	Camille Rosenthal-Sabroux..	205
Corine Cauvet	145	Michel Roux	109
Jean Charlet	37, 73	Inès Saad	205
Hervé Chaudet	109	Brigitte Séroussi	61
Olivier Corby	133	Umberto Straccia	229
Christel Daniel-Le Bozec	193	Sylvie Szulman	85
Sylvie Despres	85	Neil Taurisson	217
Catherine Duclos	97	Pierre Tchounikine	217
Manal EL Zant	109	Maxime Thieu	193
Dominique Fontaine	181	Amalia Todirascu-Courtier...	169
Frédéric Furst	121	Francky Trichet	121
Fabien Gandon	133	Raphaël Troncy	229
Alain Giboin	133	Alain Venot	97
Nicolas Gronnier	133	Jean-Jacques Vieillot	61
Michel Grundstein	205	Manuel Zacklad	241
Cécile Guigard	133	Eric Zapletal	193
Gwladys Guzélian	145		
Mounira Harzallah	49		
Antoine Isaac	157		
Marie-Christine Jaulent	13, 193		

