Actes des
20es Journées Francophones d'Ingénierie des Connaissances
« Connaissance et communautés en ligne »
du 25 au 29 mai 2009 à Hammamet, Tunisie
avec la Plate-forme AFIA 2009

Fabien L. Gandon, INRIA Sophia Antipolis – Méditerranée, Edelweiss

http://ic2009.inria.fr

In Memoriam

La conférence « Ingénierie des Connaissances » (IC), commencée avec la journée d'acquisition des connaissances (JAC) en 1990, a vingt ans. Le Web aussi fête ses vingt ans cette année. La mémoire est un sujet au cœur de l'IC et c'est sur ce thème de la mémoire que je souhaite, à plusieurs titres, ouvrir ces vingtièmes actes, en regardant trois brins dans la tresse du passé: celui de la conférence IC, celui du Web et celui de Rose Dieng-Kuntz qui fut une fervente défenseuse du Web dans les thématiques de recherche et notamment dans la communauté IC.

En 1990, la première JAC (Journée Acquisition de Connaissances) est organisée à Lannion par le CNET (Centre National d'Etudes des Télécommunications) suivant une journée thématique organisée par le PRC-GDR IA le 4 janvier 1989. A l'époque, la JAC a pour objectif de rassembler la communauté francophone dans le domaine de l'acquisition des connaissances et de préciser les rapports entre ce domaine et celui de l'apprentissage symbolique. Cette première journée se place volontairement et d'emblée dans un cadre pluridisciplinaire insistant sur les origines variées des présentations, notamment la recherche en informatique, l'industrie et la psychologie. IC gardera comme une marque de naissance cette spécificité d'être une conférence pluridisciplinaire. La même année, Tim Berners-Lee vient de proposer pour la deuxième fois au CERN un mémo spécifiant le système *Mesh* (filet) où il suggère l'utilisation d'un hypertexte pour la gestion d'information au sein du centre en étendant les références des liens hypertextes aux adresses réseau des documents afin de tisser ce « filet » entre des documents archivés sur différentes machines. Le mémo revient avec la mention manuscrite « vague mais excitant » et ainsi commence le développement des bases techniques du Web. Chargée de recherches à l'INRIA de Sophia Antipolis et membre de l'équipe Secoia, Rose Dieng travaille à l'époque sur les systèmes experts à base de règles et notamment sur le « raisonnement sur le raisonnement » appliqué en particulier à l'explication et sur la coopération entre systèmes experts. Rose rédige cette année là le Rapport de Recherche INRIA no 1319 intitulé « Méthodes et outils d'acquisition des connaissances » qui préfigure la réorientation de ses recherches.

En 1991 *la* JAC devient *les* JAC et le programme met l'accent sur la méthodologie pour l'acquisition et la modélisation des connaissances avec notamment les méthodes Kads et Kod. Cette même année, le premier serveur Web est installé hors d'Europe. Rose publie sur un système composite pour l'acquisition des connaissances, sur les réseaux sémantiques de systèmes experts et s'intéresse aussi aux agents. Elle commence le processus administratif à l'INRIA pour créer sa propre équipe de recherche.

En 1992 les JAC s'intéressent particulièrement à l'analyse de corpus textuels pour l'acquisition des connaissances et aux composants de la connaissance pour la généricité des modèles. Au début de cette année, on recense une dizaine de serveurs Web et de nouveaux navigateurs apparaissent dans le courant de l'année. Rose crée Acacia, une équipe de recherche pluridisciplinaire dont elle est la responsable scientifique et Alain Giboin le responsable permanent. Acacia vise à développer des

aides méthodologiques et logicielles (i.e. modèles, méthodes, outils) pour l'acquisition, la modélisation et la capitalisation des connaissances, en particulier pour la construction, la gestion et la diffusion de mémoire d'entreprise. Rose publie notamment sur les thématiques d'acquisition de connaissances pour l'explication coopérative et pour les modèles de résolution de problèmes.

En 1993, Rose devient membre du comité de programme des JAC et préside leur organisation à Saint-Raphaël au sein de la plateforme JAVA à l'époque déjà sous le patronage de l'AFIA. C'est aussi la première fois que l'équipe de Rose publie aux JAC, avec un article sur l'adaptation de Kads pour la construction de systèmes à base de connaissances explicatifs. Cette même année les dirigeants du CERN annoncent officiellement que la technologie du Web sera gratuite et libre de droits. Cette étape importante permettra la pénétration virale de ces technologies dans les autres systèmes d'information. En début d'année, on dénombre une cinquantaine de serveurs. De nouveaux navigateurs apparaissent mais le plus important est Mosaic qui permet pour la première fois de visualiser les images directement dans le texte d'une page. Avec ce navigateur, le Web va réellement se répandre mondialement laissant derrière lui ses ancêtres Gopher, WAIS et FTP. Autre fait marquant, alors que ce nouveau système d'échange et d'organisation de document commence à peine à se répandre, l'approche dite CGI (*Common Gateway Interface*) est proposée pour permettre aux serveurs Web de ne plus simplement envoyer des pages statiques mais aussi d'exécuter un programme pour retourner un contenu généré. Ce détail technique va ouvrir une gigantesque avenue au Web, celle permettant d'aller au-delà du service documentaire et d'offrir un moyen d'accès universel à des services applicatifs.

En 1994, Rose préside le comité de programme de la plateforme JAVA à Strasbourg au sein de laquelle les JAC ont lieu. Cette année là, le programme intègre un nouveau mot aux côtés des modèles de connaissances : les « ontologies ». L'équipe de Rose présentera un article sur la simulation du comportement macroscopique d'un modèle conceptuel d'expertise. Cette même année, l'un des co-auteurs de cet article, Paul-André Tourtier, ramène des Etats-Unis une copie du navigateur Web Mosaïc et le présente dans l'équipe Acacia, laissant ses membres pensifs, se demandant ce que cet outil allait bien pouvoir changer. Plus de 600 serveurs Web sont maintenant en ligne mais, plus important, la conférence invitée de Tim Berners-Lee à la première édition de la conférence du Web (WWWC) présente sa vision de ce qui deviendra le Web sémantique. En parallèle, la multiplication des navigateurs a amorcé la « guerre des navigateurs » et pour éviter les dangers du morcellement ou du monopole, un organisme de standardisation est créé pour le Web : le W3C. Cette même année, Rose publie sur l'utilisation des graphes conceptuels dans le cadre de l'acquisition des connaissances à partir de multiples experts. Cette famille de formalismes deviendra le choix privilégié pour l'opérationnalisation des solutions proposées par l'équipe Acacia.

En 1995, aux côtés des thèmes maintenant classiques des JAC (acquisition des connaissances, explicitation, modélisation, représentation et langages) apparaissent explicitement deux thèmes qui resteront : les retours d'expériences avec notamment le problème de l'intégration des solutions et les liens bidirectionnels entre modélisation des connaissances d'un côté et apprentissage et raisonnement à partir de cas de l'autre.

Plus de 10 000 serveurs Web sont maintenant disponibles. En introduisant les feuilles de style CSS, le W3C commence à regarder comment standardiser le découplage entre le contenu du Web et sa présentation notamment afin de faciliter les traitements et les exploitations multiples d'un même contenu. Rose continue de s'intéresser à la méthode *CommonKads* et aux agents, et publie aussi sur la gestion des conflits lors de l'acquisition des connaissances.

En 1996, le programme des JAC met en avant les thèmes : ontologies, raffinement des modèles et méthodes, construction coopérative et industrialisation des approches. Il y a maintenant plus de 100 000 serveurs Web. Cette année voit aussi l'une des premières actions en matière de sécurité sur le Web avec la recommandation PICS permettant la protection des enfants contre des contenus inappropriés. Rose s'intéresse aux graphes de règles d'inférences graduelles pour exprimer des *topoi* et à la mémoire d'entreprise, notamment dans le domaine de l'accidentologie.

En 1997, les JICAA (Journées d'ingénierie des connaissances et d'apprentissage automatique) sont la charnière entre les JAC et IC. On peut lire dans l'introduction de ces journées qu'elles « font suite aux JAVA'96. Ce changement de nom était devenu nécessaire du fait de la présence envahissante du langage du Web », l'auteur faisant ici référence au langage de programmation Java de plus en plus présent dans les applications du Web. Ce changement de nom traduit aussi un changement de contenu : l'acquisition des connaissances vue comme le processus permettant d'acquérir et de construire la base de connaissances d'un système du même nom a vécu. Les systèmes à bases de connaissances (SBC) s'inscrivent maintenant dans le cadre plus large des systèmes d'information et l'acquisition des connaissances et la validation doivent alors s'inscrire dans le cadre de l'ingénierie des connaissances. La tendance est à ouvrir les différents domaines des conférences pour que cohabitent des préoccupations théoriques et appliquées et un maximum d'approches différentes et pertinentes. Parmi les sessions, on retrouve les ontologies, la terminologie et l'acquisition de connaissances à partir de textes et on note aussi : les hypertextes et la modélisation documentaire, les mémoires d'entreprise et les systèmes coopératifs, la validation des connaissances pour les SBC, la fouille de données, le raisonnement à partir de cas et les approches cognitives.

En 1998, le mot 'Web' apparait pour la première fois dans le titre d'un article d'IC. Il était temps car la barre du million de serveurs Web est franchie cette même année. Si les langages de représentation ou de modélisation des connaissances et les mémoires d'entreprise sont très présents au programme cette année-là, hypertextes et hypermédias assoient aussi leur présence. Une session est aussi dédiée aux systèmes d'information pour le travail coopératif médiatisé. Vincent Quint, conférencier invité à IC, parle d'aller au-delà de HTML vers des documents riches sur le Web. Cette même année est effet publiée la première recommandation sur XML. C'est aussi cette année-là qu'à Pont-à-Mousson je fais la connaissance de la communauté IC en venant présenter un article que mes directrices de DEA ne peuvent venir présenter elles-mêmes. La conférence invitée de Richard Benjamins ouvre comme perspectives de transformer une région du Web en base de connaissances. L'ontologie qu'il propose d'élaborer doit en effet constituer une référence pour annoter des documents sur le Web et permettre un accès intelligent à ces documents. Cette même année, Tim

Berners-Lee publie une note appelée « *Semantic Web Road map* » qui sera lue dans l'équipe de Rose, notamment par Oliver Corby qui établira le rapprochement entre les serveurs de connaissances à base de graphes et les approches Web sémantiques. Rose commence alors à formuler une nouvelle direction de recherche qu'elle appellera « Web sémantique d'entreprise ». Elle publie aussi sur MULTIKAT, un outil utilisant les graphes conceptuels pour comparer les points de vue de différents experts.

En 1999, IC consacre à nouveau une session aux ontologies et une autre à l'acquisition de connaissances à partir de textes mais une thématique nouvelle est introduite sur l'ergonomie cognitive et l'ingénierie des connaissances face à l'ingénierie des besoins. Un problème notamment soulevé est celui des liens de plus en plus étroits qui se tissent entre l'étude des systèmes d'information, l'ingénierie des systèmes à base de connaissances et la gestion des organisations au sein desquels ces systèmes sont mis en place. La problématique est celle de nos systèmes plongés dans des usages. Les exposés montrent que l'introduction de nouveaux procédés exploitant les connaissances individuelles et collectives soulève un ensemble de problèmes imbriqués qui ne doivent plus être traités isolément mais conjointement par des informaticiens, des cogniticiens, des ergonomes, des sociologues du travail, des spécialistes de l'organisation du travail et des gestionnaires chargés de l'organisation de l'entreprise. C'est la première année que je publie à IC. Il s'agit de mon premier article scientifique et je tremble dans un amphithéâtre de l'école polytechnique. J'obtiens aussi un rendez-vous avec Rose et son équipe pour candidater sur une bourse de thèse. Quand je suis reparti le soir après cet entretien, j'avais déjà pris les éclats de rire de Rose en plein cœur et je savais que c'était avec elle que je voulais faire ma thèse. Les groupes de travail au W3C à cette époque généralisent l'insertion des objets multimédias et des scripts dans une page, ce qui permettra au Web dans les années suivantes de se doter de contenus de plus en plus riches, dynamiques et réactifs ; il s'agit là des bases techniques qui permettent le développement du Web 2.0 actuel.

En 2000, l'accent est mis à IC sur Intranet et Internet, les systèmes d'information et l'ingénierie éducative. Une part importante est aussi accordée aux réalisations, à travers une session de démonstrations (20 logiciels) et deux sessions parallèles sur les réalisations. Il est question de l'épistémologie de l'IC et une session relie ontologies et hypertextes. La terminologie et l'acquisition de connaissances à partir de textes sont toujours présentes comme des piliers du programme mais on aborde aussi l'évolution du capital « connaissances » d'une entreprise, la modélisation et la représentation de connaissances à l'aide d'objets et la gestion des connaissances. Cette même année, un groupe du W3C s'intéresse à proposer un langage pour échanger du graphisme vectoriel (SVG). Rose Dieng devient Rose Dieng-Kuntz à l'église d'Antibes. L'équipe Acacia publie aussi ses premiers résultats sur l'utilisation des graphes conceptuels pour l'opérationnalisation du Web sémantique et en particulier RDF. Enfin Rose dirige et finalise la publication d'un ouvrage collectif de l'équipe Acacia intitulé « Méthodes et outils pour la gestion des connaissances » tout en présidant la même année la conférence internationale EKAW sur l'ingénierie et la gestion des connaissances.

En 2001, le Web sémantique fait son entrée dans le programme d'IC. Il sera le thème d'une session aux côtés de thèmes comme : les systèmes d'Information pour l'aide à la décision, l'épistémologie, les expériences pratiques, la coopération et le raisonnement à partir de cas. Le Web est devenu inamovible avec 26 millions de serveurs. Des groupes du W3C s'intéressent à généraliser la notion de lien dans les documents structurés (XLink) et à permettre l'accès au Web par la voix et l'audition (Voice XML), notamment au téléphone. Rose, infatigable, et à l'étonnement de son éditeur, finalise une deuxième édition de l'ouvrage collectif et dirige quatre thèses abordant différents aspects du Web sémantique et de ses applications en entreprise.

En 2002, IC est toujours marquée par l'interdisciplinarité et la diversité : systèmes de résolution de problèmes, systèmes hypermédia, ingénierie éducative, extraction terminologique à partir de textes ou encore gestion des connaissances confrontent leurs approches tout en se rassemblant autour des problématiques liées à la gestion des connaissances. La thématique des ontologies est devenue centrale. Et le Web sémantique s'installe comme thème à part entière dans les actes. La conception (industrielle) est cette année-là une nouvelle thématique dans les applications de l'IC. A la même époque, le W3C propose une première recommandation (P3P) pour favoriser le respect de la vie privée des internautes et ces aspects commencent à inquiéter de plus en plus de personnes. En parallèle, un autre groupe travaille sur l'accessibilité du Web et commence à émettre des guides de bonnes pratiques. Rose préside un atelier à ECAI sur la gestion des connaissances et les mémoires d'entreprises et un atelier à EKAW sur le Web sémantique d'entreprise. Elle lance aussi un nouvel axe de recherche combinant le traitement automatique de la langue et le Web sémantique pour l'annotation et l'exploitation de textes en biologie.

En 2003, Rose est présidente du comité de programme d'IC. Parmi les thèmes, on retrouve maintenant classiquement : modélisation, textes et ontologies, Web sémantique, gestion des connaissances, conception, application. Le W3C commence son implantation en Chine et propose une évolution des formulaires du Web pour les généraliser à toute la famille des langages XML.

En 2004, IC compte parmi les thèmes mis en avant : la relation entre l'IC et l'ingénierie documentaire, les activités coopératives et les systèmes d'assistance. A la fouille de textes s'ajoute explicitement la fouille de données. On trouve aussi la thématique de la résolution de conflits et la recherche de consensus. Une table ronde fait le point sur le transfert vers l'industrie des technologies de la connaissance. Jérôme Euzenat présente le langage OWL recommandé cette même année par le W3C et qui étend l'expressivité des formalismes du Web sémantique. Le nombre de serveurs Web dépasse à cette date les 46 millions. L'accès au Web par les téléphones et PDA mobiles devient une activité importante au W3C. Rose renforce le développement dans son équipe des applications du Web sémantique aux domaines médical et biologique. Après un post-doc, je rejoins l'équipe Acacia en tant que chargé de recherche junior.

En 2005, IC a lieu au sein de la plateforme AFIA organisée par l'équipe Acacia à Nice. La conférence regroupe de façon assez équilibrée les problématiques de construction et d'exploitation d'ontologies d'une part et les problématiques d'ingénierie des connaissances au sein d'organisations 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. On observe un développement du thème de l'alignement d'ontologies existantes. L'indexation et l'annotation à l'aide d'ontologies pour la recherche intelligente d'informations sont également très bien représentées dans cette édition. Plusieurs articles témoignent de l'ouverture vers d'autres disciplines comme la théorie des organisations, les systèmes de travail coopératif ou l'ingénierie éducative. Les adresses du Web (URL) deviennent multilingues (IRI) permettant d'utiliser d'autres caractères que les caractères ASCII dans nos adresses. Le W3C ouvre aussi une activité pour favoriser le Web dans les pays en voie de développement et un autre groupe est créé pour travailler sur les apports du Web sémantique dans le domaine médical. Rose reçoit le prix Irène Joliot-Curie pour l'ensemble de ses travaux dont seule une petite partie a été mentionnée ici.

En 2006, les annotations sont un thème de session à IC, ainsi que la cartographie et visualisation des connaissances. La conférence s'insère dans la semaine de la connaissance à Nantes. La question « Le Web Sémantique peut-il être social ? » est isolée dans une session avec un débat. Le W3C met en place une nouvelle structure (*Incubator Groups*) pour permettre d'avancer sur des sujets novateurs sans viser la standardisation à court terme. La problématique de l'internationalisation s'étend à plusieurs groupes de travail et le W3C inaugure un bureau en Chine continentale. Rose est nommée Chevalier de la Légion d'Honneur en France et Chevalier dans l'Ordre National du Lion au Sénégal. L'équipe Acacia devient l'équipe Edelweiss qui se réoriente vers la gestion de connaissances pour les communautés virtuelles (de pratiques, d'intérêts, etc.) interagissant à travers des ressources du Web.

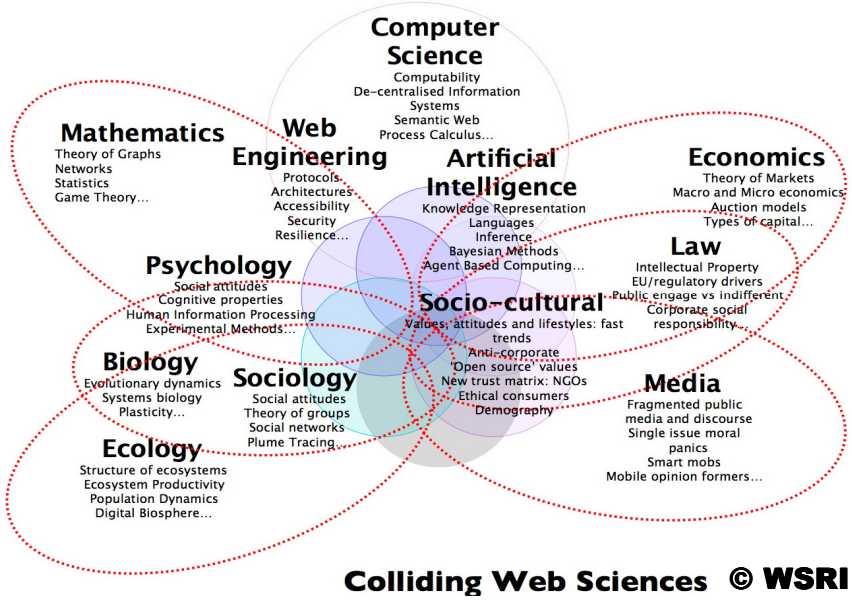
En 2007, IC aborde le Web sémantique et la recherche d'information au sein d'une même session, notamment autour de la synergie entre annotations et SBC. Parmi les thèmes on retrouve l'analyse de textes et les ontologies, les applications de l'IC ainsi que la coopération et le partage de connaissances au sein de collectifs humains. Au W3C, le langage XML se dote enfin d'un langage de requête (XQuery) et d'un moyen (GRDDL) de faire la passerelle entre le Web structuré (documents XML) et le Web sémantique (graphes RDF). Le W3C inaugure un bureau en Afrique du Sud. Rose conclue l'encadrement d'une thèse sur le Web sémantique appliqué à l'enseignement et coédite les Actes de la 7ème Conférence Terminologie et Intelligence Artificielle (TIA) et un rapport de recherche sur la fouille de textes pour l'annotation d'articles dans le domaine biomédical.

En 2008, IC consacre des sessions à la diffusion de connaissances médicales, l'interrogation de graphes de connaissances, les SBC ontologiques et la conception d'ontologies. Ivan Herman, responsable de l'activité Web sémantique au W3C, fait un état des lieux du Web sémantique qui sera l'occasion d'une discussion très riche sur les interrelations entre Web sémantique et Web social. Une session Web sémantique et Web 2.0 fait écho à l'atelier « IC 2.0 - Vers une ingénierie sociale des connaissances » qui s'était tenu la veille de la conférence en se demandant dans quelle mesure les usages du Web 2.0 font évoluer les pratiques d'IC. D'autres sessions abordent l'extraction de connaissances à partir des textes, l'instrumentation de pratiques à base de connaissances, les traces et les inscriptions de connaissances ou

l'apprentissage et l'adaptation. Un autre débat pose la question de l'Ingénierie des Connaissances française dans ses rapports avec le monde anglo-saxon. Cette année-là, Rose est trop fatiguée pour se joindre aux débats d'IC.

Aujourd'hui, en 2009, IC sort pour la première fois de l'hexagone et essaie de se doter d'une étendue francophone au-delà de sa portée nationale française. J'ai aussi choisi cette année de doter la conférence d'une session spéciale sur le thème « connaissance et communautés en ligne ». Nous retrouvons dans le programme bien évidement les problématiques du contenu, de la construction, du cycle de vie et du peuplement des ontologies. L'annotation sera à nouveau présente ainsi que la conception des interfaces et des interactions. S'ajoutent cette année deux sessions originales : une session dédiée à l'évaluation des similarités sémantiques et à l'adaptation des représentations d'ontologies à l'utilisateur ; une session consacrée à la modélisation des démarches, des pratiques et des cas. Une représentation très utilisée actuellement est celle du nuage de termes. Voici son application aux titres et mots clefs des articles de cette vingtième édition d'IC:

Cette année a vu aussi naître la conférence *Web Science* qui pose le Web comme un objet d'étude scientifique et qui est conçue comme une conférence résolument pluridisciplinaire, ce qui n'est pas sans intérêt pour la communauté IC et qui posera aussi la question du repositionnement de ces communautés en pleine évolution.



IC et le Web ont vingt ans cette année et continuent le chemin de leur évolution. Rose, elle, s'est arrêtée au bord du chemin. Rose était plus qu'une figure de la communauté IC, c'était une amie de cette communauté. Rose était aussi une éternelle émerveillée du Web, un objet complexe émergeant d'une technologie simple et qui a irrémédiablement imprimé sa présence dans le domaine de l'IC. C'est pour toutes ces raisons que j'ai voulu tresser ces trois brins d'histoire ensemble dans cette préface.

Rose avait un incroyable sens des valeurs et du contact humain, une culture impressionnante nourrissant un esprit passionné, une humilité naturelle malgré une aura mondiale et le reflexe de toujours se projeter vers l'avenir et aller de l'avant.

La *terranga* avait une fleur en France, elle s'appelait Rose. Pour moi, dix ans d'enseignements de Rose ne seront jamais assez. J'aurais tellement voulu redoubler. Son mouvement perpétuel marchait à la passion et j'aimais sa contagion. Sa méthode de direction était l'humanisme et j'admirais sa conviction. Sa force a ouvert tant de portes qu'ici bas, sans elle, je ne sais parfois plus laquelle emprunter.

Fabien L. Gandon
Président de la vingtième conférence IC.
Ecrit le 30 mars 2009.

Thèmes de l'appel à contributions

Ontologies : représentations et exploitations

- conception et création d'ontologies
- modélisation orientée ontologies
- méthodes et outils de conception et de maintenance d'ontologies
- raisonnement sur et à base d'ontologies
- partage, fusion, alignement d'ontologies, etc.

Web de connaissances

- représentation de connaissances et raisonnements sur le Web ouvert et les intranets
- Web sémantique, Web 2.0 et approches Web socio-techniques pour des applications à base de connaissances
- approches et applications utilisant les technologies Web pour la gestion de connaissances
- utilisation et extension des formalismes du Web sémantique dans des systèmes à base de connaissances
- Web de données, données liées, interopérabilité, interconnexion des modèles et applications à base de connaissances
- social tagging, folksonomies,

Dimensions individuelle, collective et sociale des connaissances

- détection de communautés : communautés d'intérêt, communautés de pratique
- assistance au cycle de vie des communautés et animation
- réseautage, réseaux sociaux, réseaux d'experts, réseaux d'accointance,
- détection et gestion de compétences
- implications sociales et méthodologies pour le déploiement d'un système à base de connaissances
- solutions complètes et solutions non techniques à base de connaissances

Représentation des connaissances

- modèles, formalismes, langages formels et informels de représentation de connaissances
- terminologies, thesaurii, ontologies et lexiques
- méthodes et outils pour le cycle de vie des représentations : détection de besoin, spécification, conception et réutilisation, diffusion, utilisation, évaluation, évolution, gestion.
- échange, interopérabilité et réutilisation des modèles, standardisation,
- échanges et intégrations entre documents et représentations de connaissances

- indexation et annotation sémantiques, description de contenu textuel ou multimédia, description de services, d'applications, ou de ressources en général, à l'aide d'ontologies

Conception et génération de modèles de connaissances

- traitement automatique de la langue (TAL) pour la construction et le peuplement de modèles
- fouille de corpus textuels ou de données pour (et par) la modélisation conceptuelle
- peuplement de modèles de connaissances et d'ontologies
- outils d'aide à l'analyse et la modélisation conceptuelle

Traitements et raisonnement sur des connaissances

- recherche d'information sémantique et basée sur des ontologies
- raisonnements logiques, inférences et raisonnements à base de règles
- raisonnements non logiques, approximations, raisonnement statistiques
- modèles et raisonnements pour les connaissances spatiales et temporelles
- raisonnement par analogie, raisonnement à partir de cas

Le temps et l'espace dans la gestion de connaissances

- représentations et raisonnements situés et/ou datés
- raisonnement sur le temps et l'espace dans des modèles de connaissances
- modèles de traces et modèles à base de temps

Evolution et historique des modèles à base de connaissances

- historique, documentation et versions des représentations et des raisonnements
- évolution des modèles, des ontologies, des bases de connaissances et des mémoires organisationnelles
- maintenance évolutive, remémoration et accès aux différentes étapes d'un modèle et d'une mémoire

Conception d'interactions et interfaces avec des systèmes à base de connaissances

- interfaces d'accès et de représentation des connaissances, liens entre représentations sémantiques, sémiotiques et pragmatiques
- connaissances pour assister les interfaces et les interactions, profils utilisateurs, modèles de contexte
- visualisation et interaction avec les raisonnements dans un système à base de connaissances
- connaissances sur l'utilisateur et adaptation

Propriété, sécurité et confidentialité dans les systèmes à base de connaissances

- provenance, suivi, identification et authentification des connaissances
- respect de la confidentialité et de la vie privée dans la communication et l'exploitation de connaissances
- raisonnements sûrs, raisonnements contrôlés et limités,
- représentations anonymes de connaissances

Applications et retour d'expérience en ingénierie de connaissances

- applications à base de connaissances ; par exemple : dans le domaine de la santé, pour l'éducation et l'apprentissage (EIAH), pour la recherche d'information (RI) et la veille, pour des applications documentaires, pour le travail collectif et dans les collecticiels (CSCW), etc.
- gestion des connaissances, mémoires d'entreprises, mémoires de projets, mémoires métier
- aide à la conception, à la réutilisation
- systèmes pour la collaboration et la coopération dans les organisations et les collectifs

Ingénierie des systèmes à base de connaissances

- méthodologies de construction de SBC, cycle de vie des SBC
- interactions utilisateur à base de connaissances dans les SBC
- interactions avec les connaissances dans les SBC
- impact de l'introduction des SBC pour les utilisateurs, les organisations et les collectifs
- méthodologies d'évaluation des SBC, évaluation des modèles de connaissances dans leur contexte d'usage

Développements théoriques et interdisciplinaires de l'ingénierie des connaissances

- épistémologie de l'ingénierie des connaissances
- théorie des organisations et ingénierie des connaissances
- sciences humaines et sciences cognitives (ergonomie, psychologie, sociologie, linguistique, etc.) et ingénierie des connaissances
- génie logiciel (langages et environnements, ingénierie des modèles, ingénierie des besoins, etc.) et ingénierie des connaissances

Les indicateurs d'IC2009

62	articles soumis et effectivement considérés.
60	relecteurs invités au comité de programme.
260	relectures effectives fournies par le comité de programme.
4,2	relectures en moyenne par article.
4,3	relectures en moyenne par relecteur.
162	messages de discussion pour l'arbitrage.
26	articles acceptés et inclus dans ces actes.
42%	de taux de sélection.
36	articles refusés dont 15 articles encouragés à faire des posters non inclus dans ces actes.

Comités

Président: Fabien Gandon, INRIA, Edelweiss

Comité de Pilotage: (bureau du GRACQ)
- Nathalie Aussenac-Gilles, IRIT, CNRS, Toulouse
- Bruno Bachimont, INA, Paris et UTC, Compiègne
- Jean Charlet, AP-HP et INSERM, Paris
- Sylvie Despres, LIPN, Université Paris 13, Paris
- Marie-Christine Jaulent, SPIM, INSERM, Paris
- Gilles Kassel, MIS, Université de Picardie Jules Verne, Amiens
- Philippe Laublet, LaLIC, Université de Paris-Sorbonne, Paris
- Myriam Lewkowicz, Tech-CICO, UTT, Troyes
- Amedeo Napoli, LORIA, Nancy
- Yannick Prié, LIRIS, Lyon
- Chantal Reynaud, LRI, Université Paris 11 & INRIA Saclay, Ile-de-France
- Sylvie Szulman, LIPN, Université Paris 13, Paris
- Pierre Tchounikine, LIG, Grenoble
- Régine Teulier, CRG, Paris
- Francky Trichet, LINA, Nantes
- Manuel Zacklad, Tech-CICO, UTT, Troyes

Comité d'Organisation: (au sein de la plateforme AFIA)
- Antoine Cornuéjols, Agrotech, Paris (Plateforme AFIA)
- Sylvie Despres, LIPN, Université Paris 13, Paris (IC)
- Moncef Temani, ISI, Tunis (Plateforme AFIA)
- Ramzi Temanni LIM&Bio, Université Paris 13, Paris (Plateforme AFIA)
- Jean-Daniel Zucker, IRD, Paris (Plateforme AFIA)

Comité de Programme: constitué du comité de pilotage plus...
- Yamine Ait Ameur, LISI / ENSMA, Futuroscope
- Patrick Albert, ILOG, Paris
- Florence Amardeilh, Mondeca/MoDyCo, Paris
- Mathieu d'Aquin, KMi, Open University of Milton Keynes
- Marie-Aude Aufaure, Laboratoire MAS, Ecole Centrale, Paris
- Catherine Barry-Gréboval, MIS, Amiens
- Jean-Paul Barthès, UTC, Compiègne
- Jean-François Boujut, G-SCOP, INP, Grenoble
- Christian Brassac, Codisant, Nancy
- Bertrand Braunschweig, ANR, Paris
- Jean-Pierre Cahier, Tech-CICO, UTT, Troyes
- Sylvie Calabretto, LIRIS, Lyon
- Pierre-Antoine Champin, LIRIS, Lyon

- Olivier Corby, INRIA, Sophia Antipolis
- Françoise Darses, LIMSI, Orsay
- Patrick Drouin, OLST, Montréal
- Catherine Faron, I3S, Nice
- Frédéric Fürst, MIS, Amiens
- Faïez Gargouri, ISIM, Sfax, Tunisie
- Alain Giboin, INRIA, Sophia Antipolis
- Nathalie Girard, INRA, Toulouse
- Monique Grandbastien, LORIA, Université Henri Poincaré, Nancy
- Mounira Harzallah, LINA, Nantes
- Ollivier Haemmerlé, IRIT, Toulouse
- Danièle Hérin, LIRMM, Université Montpellier 2, Montpellier
- Nathalie Hernandez, IRIT, Toulouse
- Antoine Isaac, Vrije Universteit, Amsterdam
- Pascale Kuntz-Cosperec, LINA, Nantes
- Florence Le Ber, CEVH, Strasbourg
- Michel Leclère, LIRMM, Université Montpellier II, Montpellier
- Eric Lecolinet, TELECOM ParisTech, LTCI, Paris
- Alain Léger, Oranges Labs et LIRIS, Rennes
- Moussa Lo, Université Gaston Berger, Saint Louis du Sénégal
- Alain Mille, LIRIS, Lyon
- Gérard Sabah, LIMSI, Orsay
- Pascal Salembier, Tech CICO, UTT, Troyes
- Nathalie Souf, CERIM, Lille
- Rallou Thomopoulos, INRA, LIRMM, Montpellier
- Yannick Toussaint, INRIA, LORIA, Nancy
- Raphaël Troncy, CWI, Amsterdam
- Brigitte Trousse, INRIA, Sophia Antipolis
- Wiliam Turner, LIMSI, Orsay
- Bernard Vatant, Mondeca, Paris

Table des matières

Cycle de vie des ontologies

Peuplement d'ontologies

Annotations

Similarités et adaptations

Interfaces et interactions

Démarches, cas et pratiques

Auteurs

Qu'est-ce qu'un tag ? Entre accès et libellés, l'esquisse d'une caractérisation.

Alexandre Monnin[1]

[1] Laboratoire EXeCO, Université Paris I Panthéon - Sorbonne,
Alexandre.Monnin@malix.univ-paris1.fr

Résumé : Nous esquissons ici un essai de caractérisation des tags en les distinguant rigoureusement de toutes les expressions linguistiques dont ils sont potentiellement le support via l'analogie avec les étiquettes matérielles. Celle-ci nous permet de mettre en évidence deux types de relations : les relations d'accès, d'ordre technique, et de référence, langagières. Le recours à la sémantique formelle conduit à distinguer, de ce point de vue, mots, mots clef, descripteurs et vedettes matières. La conclusion, dans tous les cas, est l'absence de sémantique propre au tag. Enfin, dans la lignée de ce qui précède, nous soulignons l'indétermination du rapport à la ressource, tant du point de vue de l'accès que de la référence.

Mots-clés : Tag, Tagging, Folksonomies, Référence, Accès, Langage, Web Social.

1 Introduction.

Dans sa définition initiale des folksonomies, Thomas Vander Wal[1] soulignait leur complète dépendance vis-à-vis des tags, notant par-là même leur préséance comme en échos lointain au fait qu'il fallut attendre prêt de deux ans après la création de Muxway, l'ancêtre de del.icio.us, pour qu'émergeât une appellation correspondant aux résultats socialement partagés de la pratique du tagging. Pourtant, en dépit de cette primauté unanimement admise, force est de constater la relative absence de toute caractérisation précise des tags, les chercheurs ayant souvent préféré étudier les enjeux afférant aux folksonomies. Sans doute faut-il lire dans ce diagnostic l'effet d'une saturation due au vocabulaire existant, le recours incessant à la notion pour le moins confuse et protéiforme de « mot clef » ayant achevé d'obscurcir les discussions autour du tagging en y associant pêle-mêle les balises <meta> des pages HTML, les requêtes en langage naturel[2] formulées par l'entremise des moteurs de recherche ou encore les langages documentaires et leurs multiples déclinaisons lexicales : mots

[1] Cf. Vander Wal, T. (2007), Folksonomy Coinage and Definition, *vanderwal.net* : « *Folksonomy is the result of personal free tagging of information and objects (…) for one's own retrieval. The tagging is done in a social environment (*usually [Ndr : je souligne] *shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information* ». On notera que l'aspect collaboratif n'y est pas explicitement posé comme une condition nécessaire à la constitution d'une folksonomie.

[2] Nous préférons parler de « langage naturel » plus que de langue dans la mesure où la notion de langue ressortit davantage à la linguistique qu'à la philosophie (et, *a fortiori*, qu'à la philosophie du langage).

clef, nous l'avons dit, mais aussi vedettes matière ou descripteurs. Esquisser une caractérisation des tags suppose de prendre en compte le contexte technique les ayant vu naître, avant tout lié au Web et à ses technologies en constante évolution, de même que les multiples déclinaisons auxquelles cette dynamique a donné naissance. C'est à la seule condition d'accorder suffisamment d'attention à ce milieu technique qu'il sera possible de dégager, par contraste, la part proprement « symbolique »[3] des tags, pour enfin penser l'entrecroisement de ces deux dimensions.

2. Le tag : étiquette matérielle et libellé, entre accès et référence.

2.1. Les trois dimensions canoniques.

S'il ne faut pas chercher de définition canonique précise du tag, paradoxalement, du fait du rapprochement opéré de bonne heure[4] entre folksonomies et ontologies[5], longtemps perçues, de prime abord dans un rapport exclusif de pure et simple opposition, nous disposons d'ontologies susceptibles de nous éclairer quant à ses propriétés constitutives. A cet égard, un quasi consensus se dégage de la littérature dévolue à cette question. Une ontologie du tag comme celle de Richard Newman par exemple (c'est également vrai de la *TagOntology* de Thomas Gruber[6]) se propose avant tout de décrire un processus individuel de tagging en opérant une distinction entre :

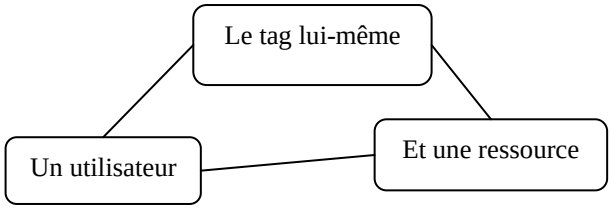


Fig. 1 – Les trois axes de la définition/réification courante du tag.

[3] Nous entendons ici *symboliques* au sens large, reprenant à notre compte la problématique ouverte par le philosophe de la technique Gilbert Hottois qui oppose symbolicité et technique, sans toutefois nier l'importance de l'activité symbolique pour les technosciences. Celle-ci se situe cependant sur un autre plan selon Hottois qui précise, avec des échos annonciateur du Web Sémantique : « Le langage des technosciences se veut purement objectif. En termes logiques, ce devrait être un langage purement extensionnel ou référentiel, dépourvu de toute étendue de sens. Un *langage qui étiquette le réel afin d'offrir des prises et de permettre l'organisation collective des relations opératoires* [Ndr : je souligne] – techniques, mathématiques – au réel. Un tel étiquetage est conventionnel et instrumental : ses mots n'apportent rien – ni sens ni supplément de maîtrise – à la maîtrise technomathématique des objets, des opérations et des processus qu'ils désignent. » Cf. Hottois (1996), *Entre symboles et technosciences*, p. 91, Champ Vallon. Aussi, et bien que nous nous penchions davantage sur les premiers, la relation symbolique ici envisagée n'exclut-elle, de notre point de vue, ni les signifiants linguistiques, ni les représentations iconiques. Nous réservons pour de futurs travaux des développements plus précis sur ce point.

[4] (NEWMAN 2005).

[5] Nous ne posons pas de coupure radicale entre ontologies philosophiques et informatiques pour de multiples raisons. Il suffira de n'en retenir ici qu'une seule : si les ontologies informatiques prétendent modéliser, et non décrire de façon définitive, la réalité, nombreux sont les philosophes qui conçoivent l'ontologie sous pareil angle déflationniste (à ceci près que leurs modèles sont langagiers et non artefactuels, et leur méthodes d'ordre logique plus qu'épistémique). Cf. (MONNIN & FELIX 2009).

[6] (GRUBER 2005).

Une telle tripartition, à laquelle on serait bien en peine de chercher des alternatives radicales dans la littérature dévolue à ces questions, pourrait cependant sembler limitée en ce qu'elle oblitère la nature duale du tag : à la fois instance *matérielle* (à l'instar de l'étiquette concrète à laquelle son nom est attaché[7]) mais également *symbolique*. Associer ces deux aspects c'est oublier que le lien symbolique usuel entre mots et choses ne nécessite aucunement de se voir implémenté d'une quelconque manière. Nul besoin d'avoir recours à des moyens d'ordre techniques pour qu'un mot atteigne son objet, aucun artefact n'y pourvoira ; autrement dit, la référence ressortit à la seule sémantique.

A l'inverse, chaque site qui emploie des folksonomies définit, selon ses besoins propres, les règles encadrant le tagging (qui a le droit de tagger ? quoi ?, comment ?, etc.), complémentant, *de facto*, la relation de *référence* par une relation associant matériellement (la dépendance de cette relation vis-à-vis d'un réseau informationnel physique en atteste) le tag à une ressource, fondée sur la notion *d'accès*[8]. Ses tenants et aboutissants sont à chercher du côté du design informationnel des interfaces et de la réalité technique des réseaux, en particulier l'architecture du Web, et non plus simplement de l'analyse du langage.

Or, l'absence de la dichotomie référence/accès au cœur de l'ontologie de Newman, ne va pas sans entraîner de sérieuses conséquences. Au premier rang desquelles, celle-ci : une telle ontologie vaut autant par ce qu'elle précise et explicite (d'où son incorporation à d'autres ontologies plus vastes à l'instar de SIOC[9]) que pour ce qu'elle offusque et qu'il nous semble impératif de restituer.

Une vision du tag tournant invariablement autour de trois axes, sans que la relativité des interfaces ne modifie cette donnée essentielle, apparaîtra pour le moins discutable du point de vue de la *description*. En revanche, une fois implémentée, elle fournira un support adéquat pour réaliser (*prescrire*) une interopérabilité entre services utilisant le tagging, implémentant, par ce fait même, une définition unifiée du tag indépendamment de toute autre considération.

2.2. Du tag au machine-tag : entre accès et libellé.

La relation d'accès diffère de la relation de référence en ceci qu'elle est indissolublement d'ordre causal, et par conséquent matériel, et relie, par une relation d'ostension, l'utilisateur à la chose taggée. L'ostension, comme n'ont cessé de le

---

[7] Un site pionnier du point de vue de l'utilisation des tags en France, Babelio.fr, traduit expressément « tag » par « étiquette ».

[8] (HAYES 2006) souligne avec force la nécessité de bien dissocier chacune de ces deux dimensions dans sa discussion du statut des URIs. Il nous semble essentiel d'intégrer ces distinctions dans la perspective qui est la nôtre, à savoir proposer une caractérisation des tags susceptible de restituer ces deux aspects ; d'une part le mot ou la suite de caractères, éventuellement doté d'une signification ; d'autre part le tag, étiquette « matérielle » ménageant un accès à la ressource encadré par des limitations très précises, imputables au système informatique en place (qu'il s'agisse du Web ou d'un logiciel comme *Photo Gallery* pour l'OS Vista de Microsoft – cf. infra). La question du rapport entre URI et ressource ayant fait l'objet de nombreuses discussions, il apparaît utile de s'en inspirer pour penser les rapports – en particulier d'accès ou de référence – entre le tag, la ressource (l'objet au sens large) et la ressource informationnelle.

[9] Semantically-Interlinked Online Communities, http://sioc-project.org/

montrer les philosophes, en particulier depuis Wittgenstein[10], se caractérise par un rapport d'indétermination intrinsèque à l'objet, ou, pour préciser les choses d'une manière plus conforme aux usages et à la réalité technologique du Web, à la ressource ainsi désignée.

Le libellé n'est rien d'autre, quant à lui, que la suite de caractères inscrite *à même le tag*, lui-même conçu à la manière d'une étiquette. En ménageant un accès à la ressource (informationnelle ou non), cette étiquette permet à l'utilisateur d'associer à celle-ci le texte[11] qu'il désire. Il devient dès lors loisible d'indexer, d'évaluer, de partager ou encore de retrouver des objets qui échappaient jusqu'alors à ces possibilités d'annotations. Précisément ce que permettait depuis longtemps, dans l'univers analogique, le traditionnel post-it : produire une surface matérielle accueillant du texte là où celle-ci faisait défaut. Illustration de ce constat, l'application *Lignes de temps*, développée au sein de l'IRI[12], autorise un accès technique à des séquences filmiques qu'il devient possible d'annoter (à l'instar, sur un versant cette fois-ci collaboratif, de la toute récente *VideoTagGame* de Yahoo![13]).

Ultime avatar de cette logique de mise à disposition de nouveaux supports *ad hoc* d'inscription des métadonnées, l'application *Gallery* de Windows Vista[14], permettant de tagger ses photos localement, propose non plus d'accoler une étiquette à une ressource numérique mais de l'y injecter directement. A la différence du post-it, souvent appelé à jouer le rôle de pense-bête mais dont la perte menace d'entraîner avec elle celle de l'objet étiqueté, *Gallery*, en conservant la trace de tous les tags présents dans le système, garantit un accès pérenne à l'ensemble des ressources taggées. La logique n'est plus celle du raccourci, susceptible de pointer dans le vide pour peu que le chemin d'accès vers la ressource visée se soit modifié, mais de *l'incorporation* au système des images, dès lors accessibles au même titre que les tags eux-mêmes ; tags et ressources formant un nouvel ensemble disposant de ses caractéristiques propres, immunisé contre la perte[15].

Rien, et l'analogie avec les étiquettes matérielles le confirme, ne contraint l'utilisateur à inscrire sur le tag une chaîne de caractère formant un mot – sans parler de sa forme lemmatisée, et ce, en dépit des contraintes syntaxiques minimales qu'imposent les différent systèmes existants[16]. Il appert de ce constat l'impossibilité d'assimiler les tags à quelque formes linguistiques précises que ce soit ; d'où une réalité incomparablement plus complexe que ne le laissait percevoir le précédent

---

[10] Wittgenstein, L., (1953), *Recherches Philosophiques*, Gallimard.

[11] Nous entendons le mot « texte » dans une acception élargie incluant tous types de signes ou de codes imaginables.

[12] http://www.iri.centrepompidou.fr/

[13] http://sandbox.yahoo.com/VideoTagGame/

[14] Gene Smith présente le point de vue des concepteurs de cette application recueilli au cours d'entretiens, cf. Smith, G. (2008), *Tagging : People-Powered Metadata for the Social Web*, 1st ed., New Riders.

[15] D'où les affinités que de telles application entretiennent vis-à-vis de la problématique dite du PIM (*Personal information management*).

[16] Del.icio.us sépare les tags par des espace ce qui oblige à former des phrases ou des expressions composées à l'aide, entre autres, de tirets ou de barres d'espacement, là où d'autre système séparent les tags par de simples virgules voire créent pour chaque tag un formulaire unique laissant toute latitude aux usagers dans le choix de leurs expressions (Zotero).

schéma à la structure ternaire où l'entité « *tag* », pour réifiée qu'elle fût en ses relations avec un *utilisateur* et une *ressource*, n'était paradoxalement pas interrogée, bien que sertie d'attributs destinés à la rendre manipulable[17]. Ce faisant, la figure unitaire du tag cède la place à une entité biface, à la fois réalité matérielle attachée à une ressource de par leur insertion à toutes deux dans un système technique donné, et relation sémiotique et langagières portée par des items de natures variables.

Avec, au surplus, et du fait de cette absence de contrainte qui constitue l'essence même du tagging, un mélange éventuel des deux : ce sont les fameux *triple tags* ou *machine-tags*, popularisés par Flickr[18]. Structuration légère des contenus produits par les utilisateurs, ils articulent trois dimensions : un espace de nom, un prédicat et une valeur associée.

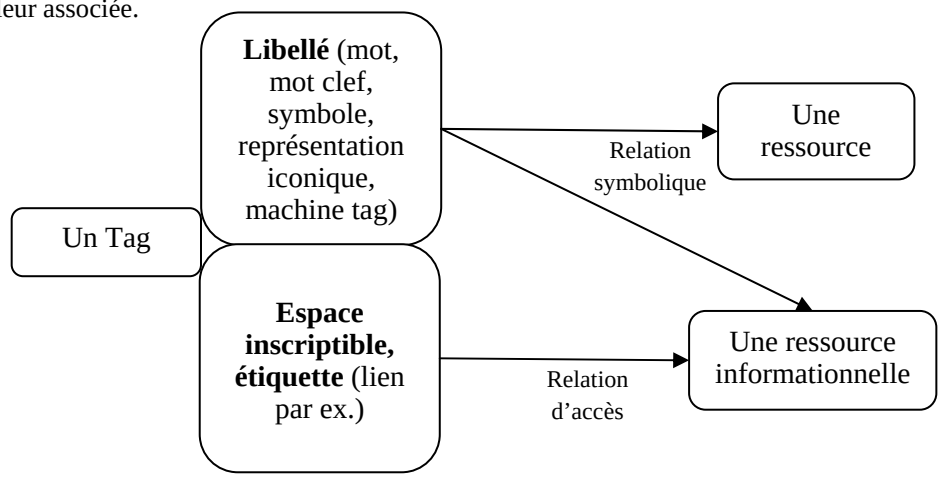


Fig. 2 – La bipartition du tag.

namespace:predicate=value[19]

[17] Un relecteur anonyme de cet article écrit : « il s'agit d'une réification de l'acte de tagging qui devient référençable et donc annotable par des relations (ex : tagging#1 est intervenu avant tagging#2) et des attributs (ex : a pour ressource, a pour tag, a pour auteur, a pour date, a pour statut, a pour langue, a pour portée, etc.) ». Nous le suivons sur ce point tout en notant avec lui que le problème de la définition du tag reste entier. Ajoutons également qu'une ontologie du tag se doit sans doute de préciser ses liens avec une ontologie de la ressource dans la lignée des débats sur les URIs auquel il a déjà été fait allusion. Cf. en particulier le travail en cours de Valentina Presutti, Harry Halpin et Aldo Gangemi sur l'ontologie de la ressource (http://ontologydesignpatterns.org/ont/web/irw.owl).

[18] On trouve aussi les « dctags », pour Dublin Core tags, cf. Johnston, P. (2006), eFoundations : dctagging, *eFoundations* (http://efoundations.typepad.com/efoundations/2006/10/dctagged.html),

Johnston, P. (2007a), dctagging revisited, *eFoundations*

(http://efoundations.typepad.com/efoundations/2007/09/dctagging-revis.html),

Johnston, P. (2007b), Flickr Machine Tags and API changes, *eFoundations*.

(http://efoundations.typepad.com/efoundations/2007/01/flickr_machine_.html).

[19] Le choix de chaque facette est laissé à la discrétion des utilisateurs. Les machine-tags employés en liaison avec des services en ligne (le premier d'entre eux fut geo.licio.us), sont interprétés de manières spécifique par ces applications et les libellés choisis assimilable à du code informatique. Les *triple tags*, dont ils s'inspirent, matérialisent une idée lancée ultérieurement sous forme de boutade, celle d'employer des tags

En scindant l'information de la sorte, et en la répartissant conséquemment au sein d'une base de données spécifiquement dédiée à chacune de ces trois facettes, il devient possible de poser des requêtes sur un espace de nom donné, un prédicat, une valeur, ou l'une quelconque des combinaisons associant ces trois aspects.

Ces tags représentent un cas limite qui illustre néanmoins la dualité accès/référence inhérente au modèle proposée dans la figure n°2. Historiquement, deux cas de figure se présentent. Dans le premier, le machine-tag, baptisé *triple tag* à l'origine, est utilisé en guise de libellé : sa syntaxe, calquée sur les langages de type XML et influencée par le développement des micro-formats et de RDF, permet d'exprimer des relations complexes à l'aide d'un seul et unique tag. Celui-ci, une fois inséré dans une URL, donne alors accès à une ressource quelconque[20].

Dans le second cas de figure, le machine-tag n'est associé à aucune URL en ce qu'il se suffit à lui-même, en vertu de sa syntaxe interprétable par les machines, pour ménager un accès à la ressource numérique par l'intermédiaire des requêtes lancées via les API des sites où furent développées des applications susceptibles de traiter ce type d'informations (c'est la cas de Flickr qui, sous l'impulsion des utilisateurs qui usaient déjà des machine-tags en guise de simples libellés, ajouta à son API de nouvelles fonctionnalités permettant de les parser, *reconduisant ainsi du même coup la dichotomie accès/référence au niveau du libellé lui-même*)[21].

Aucune règle expresse ne commandant l'inscription d'un mot ou d'un signe humainement compréhensible, rien n'interdit non plus le recours à du code informatique en guise de libellé. Mieux, une vague syntaxe que des utilisateurs reconstruisaient aisément sous forme d'énoncés du langage naturel donna naissance à des micro-formats à partir du moment où les machines furent reprogrammées pour

---

pour tagger d'autres tags, des « métatags » en somme, auxquels correspondent les facettes d'un triple tag. Cf. (MONNIN 2009) pour un rapide historique de la chose.

[20] A titre d'exemple, del.icio.us a employé cette convention pour signaler des liens pointant vers des ressources multimédia (images, vidéos), voire documentaires au sens plus traditionnel du terme (fichiers .doc ou PDF) – les valeurs des prédicats en question ne se limitaient cependant pas à de simples extensions de fichiers : les mashups réalisés à partir de mp3 étaient par exemple désignés par : « system:filetype:mp3+mashups ».

Cf. Schachter, J. (2005), casting the net wider, *delicious blog*

(http://blog.delicious.com/blog/2005/06/casting_the_net.html).

[21] Le fait que les utilisateurs de Flickr, précurseurs en la matière, qui comptaient parmi leurs tags des triple tags, durent les ajouter à nouveau au système pour que celui-ci fût en mesure de les interpréter comme des machines tags est significatif à cet égard, cf. Catt, D. (2007), Flickr Ramps up Triple Tag (Machine Tags) Support., *geobloggers.com* (http://geobloggers.com/2007/01/24/offtopic-ish-flickr-ramps-up-triple-tag-support/). Les machines-tags furent essentiellement conçus, dans un premier temps, pour lancer des recherches via une API. Comme l'écrivait Aaron Straup Cope en charge de cette question chez Flickr au moment de leur lancement, « For the moment, *machine tags* are principally an *API* "*thing*" ».

(http://www.flickr.com/groups/api/discuss/72157594497877875/). Cf. aussi sur le même sujet Keith, J. (2007), something:somethingelse=somethingspecific, *echoloquation*

(http://echoloquation.com/post/6939803/something-somethingelse-somethingspecific).

On notera toutefois qu'ils sont toujours utilisés comme des triple tags insérés dans des URLs. Pour une chronologie inverse, « *top down* », voir les tags « for:username » permettant d'envoyer un lien à l'utilisateur de son choix, que del.icio.us mit en place en juillet 2005, cf. Schachter, J. (2005a), tags for two, *delicious blog*. (http://blog.delicious.com/blog/2005/07/tags_for_two.html) et la critique de Knight, J. (2005), del.icio.us : tags for two, *del.icio.us: tags for two* (http://jk3.us/2005/07/09/delicious-tags-for-two/).

l'« interpréter » de manière idoine. Bien qu'extrêmement sommaire, de pseudo-syntaxe calquée sur des langages informatiques existants, elle acquît *ipso facto* le statut parallèle de syntaxe informatique de plein droit. Sous cet angle, les machine-tags résultent bien d'un développement logique qui tire partie de la caractéristique définitionnelle fondamentale des tags.

3. Quelle sémantique pour les libellés ?

D'ordinaire, les tags sont assimilés à des mots clefs sans autre forme de procès. Or, il faut rappeler ici les conditions d'usage précises des notions, nombreuses, issues de la bibliothéconomie, auxquelles ils ont été comparés. Sans cela, le risque serait grand d'obscurcir tout discours visant à cerner avec précision la sémantique non des tags eux-mêmes mais des libellés[22]. La pratique documentaire usuelle présente des définitions très précises, examinons-en quelques-unes.

3.1. Les descripteurs.

La norme internationale ISO 2788, déjà citée par Manuel Zacklad[23], définit le thésaurus comme le « *vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations* a priori *entre les notions (par exemple relation générique-spécifique)* ». Première conséquence immédiate de cette définition, si le but de l'indexation demeure de rendre compte du contenu d'un document, l'absence de relation référentielle confère à l'indexeur la possibilité de recourir à un nombre indéfini de descripteurs. Les descripteurs ne sauraient, dans ces conditions, et de manière univoque, correspondre à des sujets. Ceux-ci se laissent en revanche appréhender au moyen d'un faisceau de notions, toutes puisés dans le vocabulaire d'un thésaurus.

Celui-ci résulte de l'articulation d'un réseau de relations limitées entre les descripteurs qui le composent (relations de généricité et de spécificité ou d'association pour ce qui est des principales). Le choix de chacun d'entre eux présupposant que soient spécifiés les rapports qu'il entretient avec les autres, celui-ci s'effectue soigneusement en amont[24]. Un descripteur n'existe par conséquent, qu'imbriqué dans un vocabulaire contrôlé définissant avec précision la place revenant à chaque terme. Sous cet angle, la signification des descripteurs se conçoit bien davantage sous un angle inférentiel, nourri par différentes relations de sens internes au lexique, telles l'hyponymie, l'hyperonymie ou la synonymie, que référentiel.

---

[22] Sur la page de discussion consacrée à l'article « Tag (metadata) » de Wikipedia, Joshua Schacter, créateur de Muxway en 2002 et del.icio.us un an plus tard, précise les raisons qui l'ont poussé à distinguer les tags des mots clef et à forger, pour ce faire, un terme nouveau: « *While keywords are not new, I believe that tagging is a larger concept than just assigning keywords to things, however - I feel that it also includes the retrieval of the set of used terms/keywords/whatever upon view of the items. Additionally, I am reasonably sure that I named this*. »

http://en.wikipedia.org/w/index.php?title=Talk:Tag_%28metadata%29&action=edit&section=14

[23] (ZACKLAD 2007).

[24] Quant aux termes non retenus, les « non-descripteurs », les requêtes les prenant pour point de départ sont basculées sur des termes du lexique sélectionnés pour jouer le rôle de descripteurs.

Certes, on objectera qu'à chaque terme est corrélé un concept qui en fournit la dénotation. Pourtant, comme le souligne Manuel Zacklad à nouveau, « les thésaurus s'appuient sur une caractérisation des concepts qui les font au moins pour partie dépendre des langues et des mises en discours ». Cette « dépendance du concept vis-à-vis du système de la langue et la variabilité intrinsèque induite par cette dépendance » conduisent à nuancer la portée référentielle concédée aux descripteurs pour la simple et bonne raison que les deux niveaux, conceptuels et langagiers, ne se conçoivent pas l'un sans l'autre, ou dans un rapport classique de représentation supposant l'autonomie revenant en propre à des domaines hétérogènes[25]. Ils s'apparentent au contraire à l'unique face d'un anneau de Moebius, là où l'ontologie s'attache au contraire à garantir une caractérisation relativement indépendante du niveau conceptuel, outillant et opérationnalisant la rupture avec son ancrage originaire dans la langue vernaculaire.

3.2. Les vedettes matières.

Une manière de désigner une vedette-matière consiste à parler de sujets, traduction de l'anglais *subject heading*. Un sujet n'est pas un simple mot. Entité lexicale complexe extraite d'un langage documentaire artificiel, et non du langage naturel, il s'agit avant tout d'un syntagme résultant de la coordination de plusieurs descripteurs. L'écart entre langage naturel et langages artificiels de la bibliothéconomie se traduit, au plan sémantique et référentiel, par une hétérogénéité des modèles[26] servant d'appui à une appréhension de la signification des entités lexicales en jeu. Comme le note Helen Svevonius[27], traditionnellement, l'extension d'un mot fait référence à la classe des entités dénotée par ce mot ; tout l'enjeu d'une sémantique formelle, dans la lignée de ce qu'Emon Bach[28] a nommée « métaphysique du langage naturel », étant de spécifier le statut desdites entités. A l'inverse, la référence d'un syntagme tenant lieu de vedette matière ne s'entendra qu'en liaison avec un modèle divergent. Aussi un sujet, dans un contexte documentaire, fait-il expressément référence à une classe de documents portant sur un même contenu, à savoir le *thème* qui les unit[29]. Au-delà de

[25] Une manière de nuancer cette affirmation serait de souligner la mise en concurrence des termes lorsqu'il s'agit d'en sélectionner un pour *représenter* un concept. Comme le souligne Zacklad toujours, « les concepteurs d'un thésaurus vont se fixer sur une expression linguistique, le descripteur, et le considérer « toute choses étant égales par ailleurs » comme le meilleur représentant du concept visé. » Toutefois, les équivalences ainsi dégagées, l'étant sur la base de relations de synonymie ou de traduction (dans une optique respectivement intra ou inter-linguistique), vis-à-vis desquelles la philosophie a formulé de nombreux avertissements, elles relèvent avant tout du *postulat de signification*. L'exemple fameux de Quine le montre, « célibataires » et « non-mariés », bien que couramment tenus pour synonymes, ne le sont pas en vérité, ce que seul un souci pour une caractérisation précise des extensions des prédicats « célibataires » et « non-mariés » met en relief, et que recouvre, à l'inverse, la postulation d'une équivalence allant de soi entre intensions.

[26] Un « modèle », au sens de la théorie des modèle (et de sa reprise par les linguistes dans le sillage des travaux pionniers de Richard Montague).

[27] (SVEVONIUS 2000), chapitre 8.

[28] (BACH 1989).

[29] La norme AFNOR NF Z 44-070 précise ceci : « Chaque vedette-matière correspond à un seul sujet, simple ou complexe. Un même document peut avoir plusieurs sujets donnant lieu à la rédaction de plusieurs vedettes-matières ».

la difficulté notoire grevant toute tentative visant à dégager des significations lexicales dans le cadre d'une sémantique formelle, toute appréhension d'un langage quelconque prétendant lui adjoindre une interprétation articulée sur un modèle unique éclaire davantage les conditions de production et les choix ayant présidé à l'élaboration de ce langage qu'elle ne tient lieu de vérité logique. Les tags, en revanche, s'abstraient aisément de cette contrainte. Il suffit de mentionner l'existence des « *to-do* » tags, indiquant des actions à accomplir en liaison avec une ressource sans qu'ils n'en reflètent aucunement le contenu, pour que s'évanouisse toute tentative de les assimiler purement et simplement à des vedettes matières.

3.3. Les mots clef.

Il est difficile, voire impossible, de s'accorder aujourd'hui sur une définition consensuelle du mot clef, celui-ci connaissant un extraordinaire succès depuis son emploi pour signifier des requête effectuées (en plein texte) via des moteurs de recherche, jusqu'à sa caractérisation par les normes documentaires. C'est à cet usage bien établi, et à lui seul, que nous nous référerons pour l'occasion. « Mot ou groupe de mots choisi soit dans le titre ou le texte d'un document, soit dans une demande de recherche documentaire, pour en caractériser le contenu », selon la définition de la norme AFNOR NF Z 47-102, le mot clef, bien qu'issu du langage naturel, se conçoit la plupart du temps comme extrait directement d'un document analysé. Or, dans le cas précis qui nous occupe, les types de documents soumis au tagging varient dans des proportions suffisamment importantes pour qu'il ne soit tout simplement pas possible de définir les libellés uniquement en ces termes : qu'ont de commun, en effet, une photo, un événement, un plan séquence ou un enregistrement audio ? S'agissant de ce qui retient ici notre attention, leur nature de document non-textuel n'offre guère de prise à l'extraction directe de mot clef. Inversement, la force du tagging, face à ce type d'objets, est précisément de nous permettre d'ajouter du texte (entendu au sens large, cf. *supra*). A une logique d'extraction à laquelle se prêtent tout particulièrement les documents de nature textuelle, succède une autre logique, *expressive* et non seulement descriptive, que sont incapables de capter les applications proposant des « tags » générés automatiquement. Tagger c'est aussi ajouter un contenu absent d'un document (ou d'une ressource) ; en d'autres termes, lui adjoindre un contenu *extrinsèque*.

3.4. Le libellés : un espace vide et non un mot clef, un descripteur ou une vedette-matière.

Le contraste le plus saisissant du point de vue d'un indexeur professionnel au vu des normes encadrant sa pratique, fut certainement le passage d'une indexation contrôlée, *a priori*, à des formes plus libres. En l'absence d'élaboration préalable de vocabulaires contrôlés et sans préjuger des tentatives en cours pour produire des ontologies à partir de l'effort collectif de constitution d'une folksonomie, il convient de tenir les libellés des tags non pour des termes, des sujets, voire de simples mots (clefs ou non) mais tout cela à la fois – parmi une kyrielle d'autres choses encore. Au

nombre desquelles on pourra mentionner, en guise de chaînons de caractère faisant office de libellés : les machines-tags, les URL, les émoticônes et autres représentations iconiques, les signes à la signification plus ou moins idiosyncrasique, etc. Rien n'interdit en effet de doter un simple signe tel @, par exemple, d'une signification idiosyncrasique (idiolecte) affirmée en dehors de tout système socialement constitué de la langue (sociolecte). Quant aux machine-tags, ils constituent un cas limite où du code informatique, venant nourrir des applications en ligne pour la constitution de nouveaux services, fournit une information compréhensible (au moins dans certains cas) par des êtres humains, du fait d'une syntaxe ressortissant aux langages artificiels tout en s'avérant pour partie copiée sur le langage naturel.

Si l'on s'accorde à souligner l'extrême liberté avec laquelle il devient possible de choisir ses libellés, liberté expliquant en grande partie le succès du tagging, reste alors à en tirer l'unique conclusion qui s'impose. Contrairement aux vedettes matières ou aux descripteurs dont la sémantique est attachée, d'une manière contrainte, soit à un modèle spécifique, soit à un lexique intégralement ordonné par des relations de sens en vue d'éliminer toute ambigüité, les libellés inscrits sur les tags sont susceptibles d'accueillir des entités contrastées, linguistiques ou non, interdisant *de facto* une intelligence globale de la sémantique sous-jacente à leur usage. En d'autres termes, le *libellé* d'un tag est un *espace vide inscriptible* susceptible d'accueillir les entités symboliques voire informatiques de son choix[30]. *En tant que tel, il est dès lors dénué de toute sémantique fixe* (une telle question ne se posant tout simplement pas au niveau du tag lui-même, compris sur son versant matérielle et technique).

4. Indétermination de la ressource, indétermination de l'objet.

Rashmi Sinha[31], dans son analyse des fondements cognitifs du tagging, suggère d'y voir un assouplissement des procédures de catégorisation, fondées jusqu'alors sur l'utilisation des répertoires et jugées contraignante. Ces derniers exigeraient en effet un investissement cognitif très important pour déterminer où ranger une ressource. Obstacle qui disparaîtrait grâce au tagging, celui-ci éliminant cette barrière en vertu de sa supposée simplicité. Tout le monde n'en acquière pas pour autant les connaissances et le savoir-faire d'un indexeur professionnel. En revanche, l'indexation, par ce biais, est rendue plus accessible : elle autorise l'erreur et la redondance aussi bien que les hapax.

Une telle vision n'est pas sans exercer une certaine séduction qui tendrait à l'accréditer en retour. Seulement, en dépit de ses attraits, le tagging nous semble néanmoins comporter une difficulté d'un tout autre ordre. Nous l'avons dit, les tags instancient des relations d'accès et de référence. A des entités réelles ? A des ressources numériques ? Loin d'être évidente, la réponse renvoie en partie à la

[30] La limite dépend bien sûr des symboles, polices, etc. disponibles dans un système technique donné.

[31] Rashmi Sinha est la fondatrice du site de partage de présentations *Slideshare* et l'auteur d'une analyse souvent cité dans la littérature consacrée aux tags sur les ressorts cognitifs du tagging. Cf. Sinha, R. (2005), A cognitive analysis of tagging, *Rashmi's blog*, http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/.

question du déréférencement des URIs. Les différents modes de visées des tags vers leurs objets suggèrent en effet, corrélativement, une multitude d'entités visées.

Les entités présentes sur le web quant à elles sont largement assimilables à ce que le philosophe Daniel Dennett nomme des « artefacts intentionnels »[32]. Le commerce que nous entretenons avec ces derniers est nécessairement affecté par une caractéristique qui leur échoit : celle d'être *à propos de*. Un texte peut être à propos de quelque chose de même qu'une image ou un enregistrement. Aussi, lorsqu'un utilisateur taggue une image, affirme-t-il quelque chose à propos d'une entité présente ou absente du Web (l'image vs ce qu'elle représente) et se donne-t-il la possibilité d'accéder à nouveau à une ressource numérique en ligne (la photo déposée sur le site à une adresse précise).

Cette relation d'accès elle-même est indéterminée[33] à l'instar de la description linguistique de l'objet, le passage d'une URI à une URL résultant d'une négociation entre agents. Une fois générée une page[34] (par exemple), reste à savoir ce que l'on décrit grâce aux libellés des tags : la page elle-même, notre rapport à la page, une partie de celle-ci ? Et laquelle ? Le billet d'un blog ou le commentaire particulièrement intéressant qui lui fait suite ? Le lien hypertexte lui-même ? Une action à accomplir liée à cette page ?, etc.

Le site *Upcoming.org* permet par exemple de créer une page dédiée à un événement et lui associe un identifiant unique. Or, alors qu'une requête HTTP ne livrera accès qu'à la page disponible sur le site et jamais à l'événement lui-même (qui, la plupart du temps, du fait de son caractère temporel passager, n'est pas encore advenu ou, au contraire, est déjà révolu), un tag, par l'intermédiaire de son libellé, en vertu de la relation symbolique qu'il établit, saura s'affranchir de ces limitations et référer à un événement passé, présent ou à venir. Mieux, seul l'utilisateur pourra établir avec certitude à quoi renvoient lesdits libellés de chacun de ses tags, un ensemble de tags pouvant servir à décrire des entités très différentes : une page, un lien, une partie d'un texte, une action, un jugement, une relation, une personne, un lieu, un événement, etc. Une telle indétermination cadre tout à fait avec l'absence de sémantique propre à la partie libellé des tags. Elle n'en exige pas moins un attachement scrupuleux au contexte, seul à même de rendre compte de l'utilisation singulière d'un tag car, si des relations sont *ipso facto* établies, elles n'en sont pas explicitées pour autant.

Conclusion.

Nous avons esquissé ici une caractérisation inédite des tags en les différenciant rigoureusement de toutes les chaînes de caractères dont ils sont potentiellement le support, et ce, quelque soit leur statut : mots, mots clef, descripteurs, vedettes matières, etc. En insistant sur la variabilité des inscriptions (libellés) qu'un tag

---

[32] Dennett, Daniel C (1990), The Interpretation of Texts, People and Other Artifacts.

[33] On parle, s'agissant du phénomène de la *deixis*, de « *deferred ostension* », pour distinguer l'index du référent, ce qui est montré et ce à quoi il est fait référence. La monstration, de ce point de vue, se situe à mi-chemin entre l'accès technicisé et la référence linguistique : une zone où le langage s'incorpore les gestes.

[34] Qu'est-ce qu'une page au demeurant, la difficulté qu'il y a à s'accorder sur une réponse à cette question explique l'indétermination à laquelle nous faisions allusion au sujet de la relation d'accès elle-même.

(étiquette) est susceptible de recevoir, nous avons cherché à montrer que les premières n'étaient pas pourvues d'une sémantique fixe. Aussi n'est-il pas possible de partir du postulat qu'un tag réfère toujours à un concept en vertu de sa sémantique dénotationnelle. Un tel axiome sert évidemment le rapprochement entre ontologies et folksonomies en garantissant une commune appartenance à un même registre sémantique. On notera cependant qu'en laissant de côté nombre de libellés potentiels (des machines tags aux représentations iconiques), c'est une partie importante de l'intérêt des tags qui se trouve par-là même évacuée. Qu'il soit nécessaire d'opérationnaliser le rapprochement entre ontologies et folksonomie par de nécessaires concessions est une chose que l'on ne saurait contester[35]. L'objection ne peut demeurer au plan théorique. Elle appelle par conséquent une réflexion sur l'apport des technologies du Web sémantique pour réintégrer tous les usages non-dénotationnels, expressifs, du langage. Mieux, outre l'approche guidée par un réalisme opérationnalisé dont les URIs sont le meilleur exemple, la réflexion menée autour des tags en appelle une autre, d'ordre sémiotique cette fois.

Références (sélection).

BACH E. W. (1989), *Informal Lectures on Formal Semantics*, State University of New York Press.

GRUBER, T. (2005), TagOntology - a way to agree on the semantics of tagging data, http://tomgruber.org/

GRUBER, T. (2007), Ontology of Folksonomy: A Mash-up of Apples and Oranges, *Int'l Journal on Semantic Web & Information Systems*, 3(2).

HAYES, P. (2006), In Defense of Ambiguity, Edinburgh, Scotland.

LIMPENS F., GANDON F. BUFFA M. (2008), Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées : un état de l'art., *IC2008, 19èmes Journées Francophones d'Ingénierie des Connaissances*, Nancy, France.

MONNIN A. (2009), From Game Neverending to Flickr. Tagging systems as ludic systems and their consequences. In: *Proceedings of the WebSci'09: Society On-Line*, 18-20 March 2009, Athens, Greece. (In Press)

MONNIN A. & FELIX E. (2009), Essai de comparaison des ontologies informatiques et philosophiques : entre être et artefacts, *XVIèmes Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels de Rochebrune,* Megève, France.

NEWMAN R. (2005), Tag Ontology, *holygoat*. http://www.holygoat.co.uk/owl/redwood/0.1/tags/

NEWMAN R. (2005b), Tags, *holygoat.co.uk*. http://www.holygoat.co.uk/blog/entry/2005-03-23-2

PASSANT A. (2007), Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs, *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM)*, Boulder, Colorado, USA.

PASSANT A. & LAUBLET P. (2008), Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data, *Proceedings of LDOW 2008*, Beijing.

SVENONIUS E. (2000), *The Intellectual Foundation of Information Organization*, MIT Press (MA).

ZACKLAD M. (2007), Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI), http://archivesic.ccsd.cnrs.fr/sic_00202440/en/

[35] (PASSANT & LAUBLET 2008) soulignent cet aspect en notant les limitations que leur application impose.

Indexation de photos sociales par propagation sur une hiérarchie de concepts

Michel Crampes[1], Jeremy de Oliveira-Kumar[2], Sylvie Ranwez[1], Jean Villerd[1]

[1] LGI2P, Ecole des Mines d'Alès
Site EERIE, Parc scientifique G. Besse, F – 30 035 Nîmes, France
{Michel.Crampes, Sylvie.Ranwez, Jean Villerd}@ema.fr
[2] School of Computer Science and Engineering, UNSW,
Sydney NSW 2052, Australia
jdok706@cse.unsw.edu.au

Résumé : Nous nommons 'photos sociales' les photos qui sont prises lors d'événements familiaux ou de soirées entre amis et qui représentent des individus ou des groupes d'individus. Leur indexation consiste à repérer l'événement et les personnes présentes sur les photos. Dans cet article nous présentons une méthode et des outils pour faciliter cette tâche.
De nouvelles photos sont indexées à partir de photos déjà indexées selon un procédé de 'propagation' qui se compose d'un 'glisser-déposer' suivi d'une fusion et d'une affectation des contenus. Il convient au préalable d'organiser sur l'écran les photos déjà indexées selon une disposition qui facilite l'identification des personnages. Dans ce but nous faisons appel aux techniques d'Analyse Formelle de Concepts et nous proposons un algorithme de construction incrémentale d'un Diagramme de Hasse pour à la fois faciliter le repérage, intégrer les photos nouvellement indexées dans le processus d'indexation et maintenir la représentation mentale de l'utilisateur.

Mots-clés : Ingénierie des connaissances, indexation, Analyse Formelle de Concepts, Diagramme de Hasse incrémental.

1 Introduction

Les appareils photographiques numériques et les téléphones portables dotés de capacités de prise de vues sont très utilisés pour conserver la mémoire d'évènements sociaux tels que les mariages, les soirées entre amis, les anniversaires, etc. Le rappel de ces photos sociales et leur partage nécessite souvent une phase préalable d'indexation, qui consiste à repérer l'événement (lieu, date, nature, etc.) et les personnes visibles sur les photos. Ces informations peuvent évidemment être saisies manuellement, comme dans 'Facebook', mais ce procédé est vite fastidieux quand il faut indexer des centaines de photos. En fait il existe peu d'outils qui assistent réellement un utilisateur dans cette tâche. Nous proposons ici une approche originale pour indexer les photos sociales en mettant en œuvre trois stratégies.

(i) le contenu de certaines photos déjà indexées sert à l'indexation de nouvelles photos selon un procédé de 'propagation' ;

(ii) pour retrouver certaines photos afin de propager leur contenu, nous faisons appel aux techniques de classification et de représentation de l'Analyse Formelle de Concepts (Ganter & Wille, 1999), et en particulier aux Diagrammes de Hasse ;

(iii) les photos nouvellement indexées sont automatiquement intégrées dans un Diagramme de Hasse de manière incrémentale tout en maintenant la carte mentale de l'utilisateur.

Toutes ces fonctionnalités ont été testées avec des étudiants de premier cycle. Selon ceux-ci et au vu d'observations que nous présentons à la fin, nos méthodes et nos outils se révèlent plus attractifs et plus performants que les solutions proposées actuellement dans la littérature et dans des applications comme Facebook, ou Flicker. Bien que limitée pour l'instant aux photos sociales, la méthode pourra être envisagée par la suite pour d'autres types de photos avec sans doute quelques adaptations.

2 Etat de l'art

Face à l'enjeu qu'elles présentent, de nombreuses solutions sont explorées pour indexer les photos numériques, et en particulier les photos sociales.

- Les techniques d'analyse d'images ont fait de grands progrès avec certaines applications de reconnaissance de visages déjà arrivées au stade commercial comme dans l'application iPhoto d'Apple. Mais les photos doivent avoir été prises dans un environnement très contrôlé (éclairage, orientation, etc.), ce qui est rarement le cas des photos sociales telles que nous les avons précédemment définies (voir par exemple un état de l'art en la matière dans (NSTC report, 2006). Notre méthode pourrait venir en complément de la reconnaissance de visages pour les photos prises dans de mauvaises conditions.
- L'alternative la plus évidente à l'indexation automatique est la saisie au clavier des noms des personnes présentes sur les photos. Mais cette méthode présente deux inconvénients majeurs. Tout d'abord l'utilisateur doit décider des mots à saisir, ce qui, même quand on considère le cas limité des photos sociales, pose le problème du choix des mots et de leur orthographe. Par ailleurs, ce procédé est long et fastidieux. Beaucoup d'utilisateurs se découragent et les photos restent non classées, ce qui ne facilite pas les échanges personnalisés.

Pour éviter les saisies répétées et pour mieux contrôler le vocabulaire, il est possible de faire appel à une liste incrémentale des noms des personnes. Cette approche facilite effectivement le travail de l'indexeur, mais affecter des noms à partir d'une liste sur des photos reste encore une tâche fastidieuse comme nous avons pu l'observer lors des tests présentés dans la conclusion.

- Certains chercheurs proposent d'exploiter des index observables lors de la prise de photos (localisation géodésique, date) accompagnées d'informations sociales pour suggérer des lieux et des personnages (Monaghan & O'Sullivan, 2007). Cette approche pourrait aussi compléter notre méthode, cette dernière permettant de passer de la suggestion à l'indexation précise de chaque photo dans sa dimension humaine.
- L'indexation partagée ("social tagging") constitue une réponse astucieuse à la pénibilité de l'indexation. Les index créés par des utilisateurs peuvent être utilisés par

d'autres. Mais le problème du choix des mots et le caractère fastidieux de l'indexation mot à mot de centaines de photos en limite la portée pour le cas des photos sociales.

- A notre connaissance, l'indexation par propagation qui consiste à indexer de nouveaux éléments à partir d'éléments anciens n'a été proposée que dans le domaine de la musique (Crampes et al., 2007). Cependant, la méthode ne fait pas appel à un support formel comme un Diagramme de Hasse représentant une sous-hiérarchie de Galois.
- Certaines applications récentes utilisent une Hiérarchie de Galois pour naviguer dans une collection photo comme dans (Eklund et al., 2006), (Ferré, 2007) et (Loisant et al., 2003). Mais les auteurs n'utilisent pas un Diagramme de Hasse pour indexer les photos et ne proposent pas une technique d'intégration des photos nouvellement indexées de manière incrémentale dans la sous-hiérarchie de Galois.

3 Indexer des photos sociales par propagation

Afin d'introduire notre méthode, nous prenons l'exemple réel d'une personne qui veut indexer un ensemble de photos avec notre application expérimentale. Marie, notre personnage, est allée à un anniversaire et a pris environ 100 photos qu'elle veut indexer avec les noms des personnes afin de les partager de manière personnalisée.

Elle a la possibilité de disposer toutes ses photos sur l'écran, que nous appelons la "planche photo" ("photo-board"). Dans un premier temps, elle peut caractériser l'événement avec son lieu, sa date, sa nature, etc. Toutes ces données sont rangées sous le même événement qui apparaît à l'écran. Quand celui-ci est sélectionné, toutes les photos qui seront indexées hériteront de ses caractéristiques.

La tâche suivante qui consiste à saisir les personnes présentes sur les photos est la plus longue et la plus fastidieuse. Une photo peut montrer un, deux ou de nombreux personnages dans différentes situations lors de l'événement. Notre méthode va permettre de faciliter cette saisie et de la rendre plus sûre et plus complète en évitant la saturation de l'utilisateur. Elle peut être mise en œuvre après qu'un certain nombre de photos concernant l'événement ont été préalablement indexées ou bien qu'il existe déjà un jeu de photos indexées venant d'un événement similaire avec un sous-ensemble commun de personnes. Cette phase d'initialisation sera présentée plus loin. Pour l'instant nous considérons qu'il existe un ensemble de photos déjà indexées et une liste alphabétique des personnes présentes sur ces photos. La planche photo est présentée sur la Figure 1 avec les photos à indexer à droite et la liste des personnes à gauche.

Pour indexer de nouvelles photos, Marie va utiliser un principe de 'propagation'. Nous appelons 'indexeurs' les photos déjà indexées qui propageront leur contenu, et nous utilisons le néologisme "indexables" pour désigner les photos à indexer sur lesquelles seront propagés les contenus des indexeurs.

Dans la Figure 2 l'utilisatrice a sélectionné un indexable encadré de bleu. Elle peut visuellement identifier les personnages sur la photo en utilisant éventuellement une lentille "fish eye" grossissante (Furnas, 1986). Marie peut alors amener l'indexable près d'un ou plusieurs indexeurs qui, selon elle, contiennent les personnages de la

photo. En relâchant et en cliquant sur le bouton de la souris, elle peut sélectionner plusieurs indexeurs qui affichent alors leur contenu et sont entourés d'un disque bleu. Dans la Figure 2 l'indexable comporte trois personnages. L'indexeur au dessus a déjà été sélectionné et l'indexeur au dessous qui vient d'être sélectionné affiche la liste de ses personnages. Un double clic sur la souris déclenche la propagation qui opère de la manière suivante.

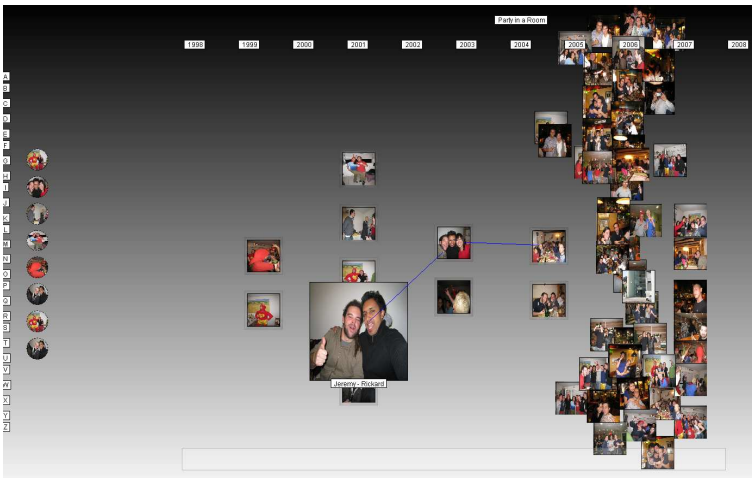


***Fig. 1** – La planche photo*

Formellement nous désignons par N l'ensemble des personnages qui sont visibles sur l'indexable, par S_i l'ensemble des personnages qui sont présents dans l'indexeur i (par exemple dans l'indexeur 2 sur la Figure 2, S_2 = {Jeremy, Maria}). Le contenu de l'indexable après propagation est égal à l'union des contenus des indexeurs sélectionnés : $N = \cup S_i$

Par exemple, sur la figure 2,

$N = S_1 \cup S_2$

Comme S_1 = {Jeremy, Willey} et S_2 = {Jeremy, Maria}, {Jeremy, Maria, Willey} est propagé vers N.

Dans certains cas des personnages visibles sur l'indexable ne figurent pas dans le contenu de tous les indexeurs disponibles. Ils peuvent alors être créés à la main et intégrés dans la liste des personnages, et en même temps au contenu de l'indexable.

A l'inverse il peut aussi y avoir des cas où il y a des intrus dans les indexeurs sélectionnés, c'est-à-dire des personnages en plus de ceux qui figurent sur l'indexable. Si on propage l'union des contenus des indexeurs sélectionnés, l'indexable risque de se retrouver avec des personnages en trop. Or ceux-ci sont nécessairement dans la liste des personnages puisqu'il a fallu les créer pour qu'ils figurent comme contenu d'un ou plusieurs indexeurs. Ils peuvent être retirés du contenu de l'indexable en cliquant sur leur image dans la liste des personnages. L'expérience a montré lors des tests que ce procédé de soustraction est généralement peu nécessaire, mais peut s'avérer utile.

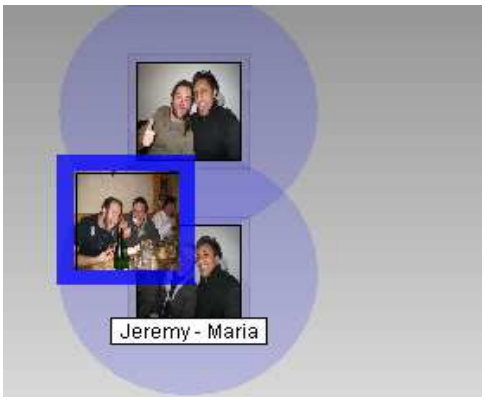


Fig. 2 – *Indexation d'une photo avec deux indexeurs*

4 Organisation des photos en un Diagramme de Hasse

4.1. Le treillis de concepts des photos indexées

La mise en œuvre de la propagation présentée ci-dessus suppose que les indexeurs soient organisés sur la planche photo de manière à rendre leur contenu clair et facile d'accès. A cette fin nous utilisons les techniques de l'Analyse Formelle de Concepts (AFC) et nous nous proposons de les organiser selon un Diagramme de Hasse.

Un Diagramme de Hasse est une représentation pratique d'un treillis de concepts (ou treillis de Galois). Dans l'AFC, un ensemble d'objets dotés de propriétés (ou attributs) peut être organisé en un treillis de concepts ; un concept contient l'ensemble des objets qui possèdent un même sous-ensemble de propriétés, en retenant le plus grand des sous-ensembles de propriétés communes entre les objets. Dans notre cas, nous considérons les photos comme des objets et les personnages sur les photos comme les propriétés de ces objets. Le processus d'organisation débute avec la construction d'un contexte formel, ou plus simplement contexte, qui est une table avec les objets disposés en lignes et les propriétés en colonnes. Chaque case est marquée (par exemple avec la valeur 1) si l'objet en ligne possède la propriété en colonne ; elle n'est pas marquée (évaluée à 0) dans le cas contraire.

Formellement, un contexte est un triplet (G, M, I) où G est un ensemble d'objets, M un ensemble de propriétés, et I une relation binaire entre les objets et les propriétés, i.e. $I \subseteq G \times M$.

La table 1 présente un contexte formel très simple construit à partir de notre exemple de photos déjà indexées, où G contient 6 photos et M contient quatre propriétés, c'est-à-dire quatre personnages potentiellement présents sur les photos.

Propriétés / Objets	Maria	Willey	Jeremy	Peter
Photo P1	1	0	0	1
Photo P2	0	0	1	1
Photo P3	1	1	1	1
Photo P4	1	1	0	0
Photo P5	0	0	0	1
Photo P6	0	0	0	1

Table 1. Un contexte avec photos (objets) et personnages (propriétés)

La construction du treillis de concepts se poursuit avec la recherche des concepts. Un concept est défini par une paire de sous-ensembles : un sous-ensemble d'objets, appelé l'extension du concept, et un sous-ensemble de propriétés appelé l'intension, qui est le sous-ensemble maximal des propriétés que les objets du concept partagent.

Pour le contexte de la table 1, les concepts sont les nœuds du graphe présenté dans la Figure 3, comme par exemple le concept ({P2, P3}, {Jeremy, Peter}) qui a pour extension {P2, P3} et pour intention {Jeremy, Peter}.

Pour un ensemble d'objets $O \subseteq G$ et un ensemble de propriétés $A \subseteq M$, on définit l'ensemble des propriétés communes aux objets de O :

$f : 2^G \rightarrow 2^M$, $f(O) = \{a \in A \mid \forall o \in O, (o, a) \in I\}$

et l'ensemble des objets qui ont toutes leur propriétés dans A :

$g : 2^M \rightarrow 2^G$, $g(A) = \{o \in O \mid \forall a \in A, (o, a) \in I\}$.

La paire (f, g) est une correspondance de Galois entre $(2^G, \subseteq)$ et $(2^M, \subseteq)$.

Un concept formel du contexte (G, M, I) est une paire (O, A) avec $O \subseteq G$, $A \subseteq M$ et $A = f(O)$ et $O = g(A)$.

Après l'identification des concepts, le but suivant est de construire le treillis dont les éléments sont les concepts. Nous définissons un ordre partiel dans l'ensemble des concepts, tel que toute paire de concepts a un supremum (le plus petit majorant commun aux deux concepts), et un infimum (le plus grand minorant commun). Formellement, soit L l'ensemble des concepts de (G, M, I) et soit $\leq_L$, l'ordre partiel est défini par :

$(O_1, A_1) \leq_L (O_2, A_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow O_2 \subseteq O_1$.

La paire $(L, \leq_L)$ est appelée le treillis des concepts de (G, M, I).

La Figure 3 montre le Diagramme de Hasse du treillis de concepts des photos de l'exemple. C'est un graphe dont les nœuds sont les concepts, ordonnés de haut en bas, selon leur ordre dans le treillis. Chaque concept montre son extension et son intension entre accolades (les éléments en gras seront explicités dans la suite). Les arêtes du graphe relient les concepts qui sont dans une relation d'ordre directe, sans concept intermédiaire. L'extension d'un concept est un groupe de photos Pi qui partagent les mêmes personnages ; ces derniers forment l'intension du concept. On peut voir qu'une photo peut apparaître dans plusieurs concepts. Il en est de même pour une propriété.

Pour simplifier les étiquettes des concepts, il est possible de ne mentionner que les extensions réduites (en gras sur la Figure 3). Une extension réduite d'un concept (O, A) est l'ensemble des objets qui appartiennent à O et n'appartiennent pas à un concept inférieur, c'est-à-dire l'ensemble des objets qui n'ont pas d'autres propriétés que celles appartenant à A.

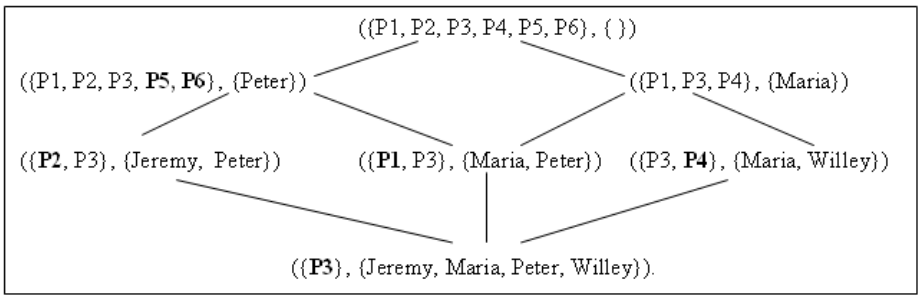


***Fig. 3** – Treillis de concepts associé au contexte de la Table 1*

Un concept dont l'extension réduite est non vide est appelé concept-objet. Pour chaque objet (photo), il existe un unique concept-objet qui constitue le concept le plus spécifique contenant l'objet. Dans la mesure où nous souhaitons qu'une photo ne soit associée qu'à une unique description, la plus spécifique, nous réduisons le treillis aux seuls concept-objets.

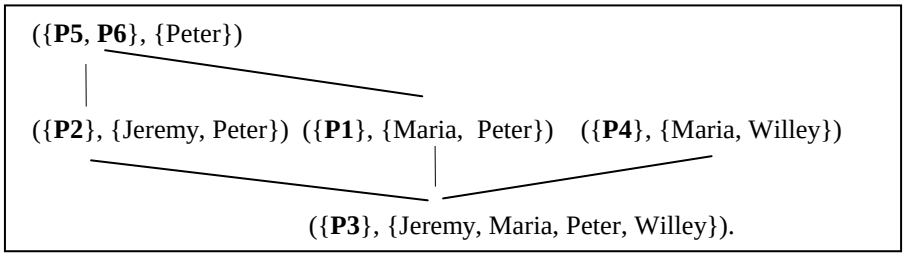


***Fig. 4** – La Sous Hiérarchie de Galois associée au contexte de la Table 1*

Ce principe d'élimination des nœuds dont les extensions réduites sont vides a été proposé dans (Godin et al., 1995) sous le nom PCL/X. On peut le voir appliqué à notre exemple dans la Figure 4. Nous obtenons une Sous Hiérarchie de Galois selon la terminologie employée dans (Godin & Chau, 1999) restreinte aux concepts objets. De manière plus explicite nous l'appellerons une Sous-Hiérarchie de Galois Objet (Object Galois Sub-Hierarchy – OGSH). La visualisation est plus claire puisque dans notre cas on se focalise sur les objets, à savoir les photos.

4.2. Construction incrémentale du Diagramme de Hasse

De nombreux algorithmes ont été proposés pour construire un treillis de Galois et son Diagramme de Hasse à partir d'un contexte. On trouvera un état de l'art succinct dans (Valtchev et al., 2000). Plus précisément (Arévalo et al., 2007) comparent trois applications principales qui ne considèrent que des OGSH, ce qui est aussi notre cas. Ces algorithmes calculent en premier lieu la topologie du graphe, et ensuite génèrent le Diagramme de Hasse, ou au mieux construisent de manière incrémentale le Diagramme de Hasse à chaque pas de calcul de la topologie du graphe. Cette approche présente un inconvénient majeur pour notre cas.

Lors de l'indexation d'une photo à l'aide des indexeurs organisés en Diagramme de Hasse, l'indexable, après propagation, doit pouvoir être intégré dans le diagramme qui doit être en conséquence mis à jour dynamiquement. Comme le diagramme est bâti à partir d'ensembles de photos représentant les mêmes groupes de personnes – ces ensembles forment les concepts – il est essentiel que l'utilisatrice trouve facilement les groupes de personnes pour indexer de nouvelles photos. En conséquence à chaque fois qu'une nouvelle photo vient d'être indexée elle doit être intégrée dans le diagramme avec un minimum de remaniements. Les anciens concepts doivent ne pas bouger, ou au pire bouger mais sans changer de voisinage. L'objectif d'une telle stratégie est le maintien de la carte mentale de l'utilisatrice, l'exigence que l'utilisatrice ne soit pas désorientée par des mouvements inattendus (Misue et al., 1995). Les algorithmes de visualisation de Diagrammes de Hasse qui calculent en premier le treillis ne sont pas applicables ici car le diagramme, à l'image de la sous-hiérarchie, n'est pas connu à l'avance puisqu'il se construit au fur et à mesure qu'il sert à indexer de nouvelles photos. Tous les autres algorithmes qui construisent le Diagramme de Hasse de manière incrémentale sont intéressants s'ils conservent les concepts à leur place, ou du moins ne les déplacent que légèrement en particulier en préservant au maximum leurs voisins sur le plan. Mais les algorithmes proposés dans la littérature ont d'autres priorités : la vitesse et le respect de certaines règles d'esthétisme parmi lesquelles la limitation du nombre de croisements d'arêtes (Battista et al., 1999). Bien que ces objectifs soient justifiés, et même s'ils restent pertinents dans notre cas, ils nous conduisent à un dilemme. La construction du graphe est incrémentale et on ne peut donc pas prédire les concepts et les arêtes qui vont apparaître. Comme par ailleurs on veut maintenir au maximum la carte mentale de l'utilisatrice, on ne peut pas réorganiser les concepts sur le plan en permanence afin de limiter le nombre de croisements d'arêtes. Le maintien de la carte mentale est notre première priorité, même si elle implique de ne pas gérer les croisements d'arêtes. De plus nous verrons plus loin que, dans notre cas particulier, la contrainte esthétique sur les arêtes est moins exigeante. Il nous faut donc proposer un nouvel algorithme incrémental centré sur le maintien de la carte mentale.

L'algorithme que nous présentons met en œuvre deux techniques. La première fait appel à un jeu de forces, une simulation d'un jeu de ressorts bien connu dans la communauté 'dessin de graphes' (Eades, 1984). Un algorithme de ce type est par exemple appliqué pour dessiner des treillis dans (Freese, 2004) (Hannan & Pogel, 2006). Cependant dans ce dernier cas le treillis est calculé avant de mettre en œuvre

l'algorithme d'affichage, ce qui va à l'encontre d'une stratégie incrémentale avec maintien de la carte mentale. A l'inverse, la seconde approche originale que nous présentons ici consiste à faire émerger le Diagramme de Hasse en premier à l'aide du jeu de forces, et ensuite d'identifier l'apparition des concepts. L'OGSH est donc le résultat de l'organisation du Diagramme de Hasse, et non le résultat d'un calcul préalable.

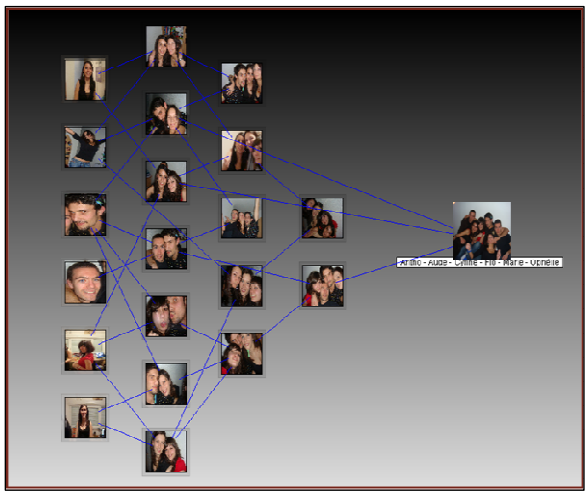


Fig. 5 – *Un Diagramme de Hasse auto-organisé de photos*

L'algorithme est le suivant. Comme dans (Freese, 2004), une nouvelle propriété est affectée à chaque photo lors de son indexation. Elle représente son rang. C'est un entier dont la valeur est égale au nombre de propriétés de la photo, c'est-à-dire au nombre de personnages présents. Une force horizontale est appliquée à toutes les photos indexées avec une intensité proportionnelle à leur rang. Les photos se déplacent alors et se positionnent sur une ligne horizontale selon leur rang. Les photos qui ont le même nombre de personnages forment des colonnes. Nous les fixons sur l'axe des abscisses.

Une nouvelle force est appliquée entre tous les objets uniquement par colonne et selon l'axe vertical. La force entre deux atomes sur une même colonne est calculée en utilisant la distance de Hamming. C'est un entier dont la valeur est proportionnelle à la différence des propriétés entre deux objets. Suite à l'application de cette force, toutes les photos qui représentent les mêmes personnages forment des tas correspondant exactement aux concepts objets, et tous les concepts sont séparés verticalement sur chaque colonne. L'intension du concept est l'ensemble des individus présents sur les photos qui forment le tas. Une troisième force est finalement appliquée pour égaliser la distance entre concepts.

A chaque fois que les forces sont appliquées sur une photo nouvellement indexée, elle va soit rejoindre un concept existant qui contient le même groupe d'individus, soit créer un nouveau concept sur une colonne contenant le même nombre d'individus.

Pour se faire une place, elle repousse verticalement les concepts déjà présents sur la colonne. La nouvelle photo peut éventuellement créer une nouvelle colonne si elle possède un nombre de personnes différent de ceux des colonnes déjà en place. Ces dernières glissent à leur nouvelle position sans modifier l'organisation des concepts qui les composent.

Quand une nouvelle photo crée un nouveau concept, des liens sont rajoutés avec les concepts immédiatement inférieurs et supérieurs qui contiennent des personnages en commun. Ce calcul est pour l'instant en cours d'optimisation. La Figure 5 montre un Diagramme de Hasse créé lors des tests qui sont décrits dans la conclusion ci-dessous. Pour des raisons pratiques, le diagramme est organisé hiérarchiquement de gauche à droite.

5 Utilisation du Diagramme de Hasse pour l'indexation

Il est maintenant possible d'utiliser le Diagramme de Hasse pour indexer de nouvelles photos à l'aide d'indexeurs bien organisés. Visible au dessus de chaque concept on trouve soit la photo la plus récemment indexée soit une photo choisie par l'utilisatrice parce qu'elle la trouve la plus représentative du groupe de personnes correspondant à l'intension. Sur la Fig. 1 les nouvelles photos à indexer sont entassées à droite, le Diagramme de Hasse des indexeurs est au centre ; une lentille "fish eye" est appliquée selon les principes décrits dans (Furnas, 1986) et la liste des personnages est située verticalement à gauche. Les liens entre les concepts en relation apparaissent localement et uniquement lorsqu'on fouille le graphe à l'aide de la souris. Ceci justifie que la minimisation du nombre de croisements d'arêtes, bien qu'intéressante, est une contrainte moins importante que le maintien de la carte mentale.

D'autres outils sont disponibles comme la possibilité de regrouper les indexables qui montrent un même groupe d'individus et de les indexer en une seule fois.

6 Tests

Deux séries de tests ont été menées avec un protocole bien établi portant sur des données objectives et subjectives. Le but était de mesurer les performances de notre application dénommée PhotoMap pour l'indexation de photos sociales face à Facebook et Flicker. Les populations concernées étaient des étudiants recrutés sur la base du volontariat. Ils avaient en moyenne 21 ans, et étaient aux trois quarts des hommes. Il est à noter que cette population correspond à un profil d'utilisateurs type pour l'application. Le premier test a permis de mettre au point la méthode d'évaluation avec 4 testeurs. Le second test avec 17 volontaires a permis de mesurer certaines performances et de recueillir des critiques. La plupart des testeurs connaissaient déjà ou avaient utilisé Facebook, quelques uns avaient utilisé Flicker, et évidemment aucun ne connaissaient Photomap. Il est intéressant de noter que

PhotoMap partait avec un léger handicap sur ce point parce que sa manipulation était ignorée de tous. Chaque test sur un logiciel commençait par une formation avec cinq photos à manipuler. Les testeurs devaient ensuite indexer seuls 40 photos, toujours les mêmes, prises lors de deux soirées auxquelles ils avaient participé. L'opération était répétée par chaque testeur sur les trois logiciels dans un ordre tiré au hasard.
Résultats. Deux indicateurs étaient particulièrement observés. 1) Le temps pour trouver une photo particulière après l'indexation d'un premier noyau de photos : il était jusqu'à deux fois inférieur en moyenne avec PhotoMap (environ 11 secondes) qu'avec les concurrents (environ 20 secondes). 2) Le temps moyen d'indexation observé pour les 15 dernières photos : les résultats donnaient en moyenne 3'10'' pour Facebook, 2'30'' pour Flicker et 2'35'' pour PhotoMap. Mais pour PhotoMap, le plus pénalisant était la création de nouveaux personnages alors que, inversement, le principe de propagation était très sollicité et vécu comme un facteur très productif.
Au plan qualitatif, la majorité des testeurs ont classé PhotoMap en première position seul ou à égalité avec un autre de ses deux concurrents sur tous les critères proposés : "simplicité", "qualité de l'indexation produite", "fonctions d'assistance", et surtout "caractère ludique". Le résultat le plus inattendu a été d'observer que 7 des 17 participants ont spontanément abandonné en cours de route par lassitude l'indexation des 40 photos avec Facebook et/ou Flicker. Par contre tous ont été jusqu'au bout avec PhotoMap, appréciant la révélation progressive des groupes humains dans le Diagramme de Hasse. Contrairement à une crainte préalable, la navigation dans le diagramme n'a jamais été qualifiée de complexe. La construction incrémentale et animée répond semble-t-il à l'objectif que l'on s'était fixé de maintien de la carte mentale de l'utilisateur. Seule l'utilisation des liaisons entre concepts a posé problème, les liens ayant été peu compris, peu utilisés, et causant même parfois de la gêne, non par leur croisement, mais par leur simple présence.
Sur les suites à donner, tous les testeurs étaient positifs, souhaitant disposer d'un tel environnement seul ou avec Facebook. Le point fort de la méthode est que l'effort d'indexation s'accompagne d'un retour d'information par la révélation de l'organisation des photos dans un Diagramme de Hasse.

7 Conclusion

Le Diagramme de Hasse et sa construction incrémentale semblent donc être de bons outils tant d'indexation que d'organisation de photos sociales, au moins en nombre limité (40 photos). La suite de nos travaux portera sur la montée en charge et sur l'assistance visuelle. En effet en nous basant sur les fondements théoriques bien établis de l'Analyse Formelle de Concepts nous pouvons avancer l'hypothèse que ce mode d'organisation est le plus à même de gérer la complexité d'une montée en charge avec plusieurs centaines de photos. Par contre, en termes de visualisation, il conviendra de mettre en œuvre des outils de zoom et de filtrage puissants et simples à manipuler sur l'ensemble du diagramme. La suite des travaux portera aussi sur la construction d'albums personnalisés qui pourront être partagés et collectivement enrichis sur un Diagramme de Hasse également partagé.

Références

ARÉVALO G., BERRY A., HUCHARD M., PERROT G., SIGAYRET A. (2007). Performances of Galois Sub-hierarchy-building Algorithms. *ICFCA*. p 166-180.

BATTISTA G., EADES P., TAMASSIA R., TOLLIS I. (1999). Graph drawing. Algorithms for the visualisation of graphs. *Prentice Hall*.

CRAMPES M., VILLERD J., EMERY A., RANWEZ S.. (2007). Automatic Playlist Composition in a Dynamic Music Landscape. *In Proceedings of the International Workshop On Semantically Aware Document Processing And Indexing, SADPI'07. ACM New York, vol. 259*, p. 15–20.

EADES P. (1984). A heuristic for graph drawing. *In Proc. of the 13th Manitoba Conference on Numerical Mathematics and Computing. Utilitas Mathematica, vol. 2.* p. 149-160.

EKLUND P., DUCROU J., WILSON T. (2006). An Intelligent User Interface for Browsing and Search MPEG-7 Images using Concept Lattices. *In Proc of the 4th International Conference on Concept Lattices and Their Applications, LNAI, Springer-Verlag*.

FERRÉ S. (2007). CAMELIS: Organizing and Browsing a Personal Photo Collection with a Logical Information System. *In Proc. of the 5th International. Conference on. Concept Lattices and Their Applications.* p. 112-123.

FREESE R. (2004). Automated Lattice Drawing. *Lecture Notes in Artificial Intelligence, 2961, Springer*. p. 112-127. Berlin.

FURNAS G.W. (1986). Generalized Fisheye Views. *In Human Factors in Computing Systems (CHI'86) Mantei & Orbeton Eds, ACM.* p.16-23.

GANTER B. & WILLE R. (1999). Formal Concept Analysis: Mathematical Foundations. *Springer*.

GODIN R., MINEAU G., MISSAOUI R. (1995). Incremental structuring of knowledge bases. *In Proceedings of the International Knowledge Retrieval, Use, and Storage for Efficiency Symposium (KRUSE'95).* p. 179-198. Santa Cruz

GODIN R. & CHAU T. (1999). Comparaison d'algorithmes de construction de hiérarchies de classes. *L'objet 5(3/4).*

HANNAN T. & POGEL A. (2006). Spring-Based Lattice Drawing Highlighting Conceptual Similarity. *In Proc of ICFCA'06*. p. 264-279.

LOISANT E., SAINT-PAUL R., MARTINEZ J., RASCHIA G., MOUADDIB N. (2003). Browsing Clusters of Similar Images. *In Proc. of BDA '03*. Lyon.

MISUE K., EADES P., LAI W., SUGIYAMA K. (1995). Layout Adjustment and the Mental Map. *In Journal of Visual Languages and Computing, Vol. 6.* p. 183–210.

MONAGHAN F. & O'SULLIVAN D. (2007). Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation. *In 2nd International Conference on Semantics and Digital Media Technologies.* P. 5-7. Genova

NSTC REPORT, (2006). Face Recognition. *NSTC, Committee on Technology, Committee on Homeland and National Security, Subcommittee on Biometrics.*

VALTCHEV P., GROSSER D., ROUME C., ROUANE HACENE M. (2003). Galicia: an open platform for lattices. *In Using Conceptual Structures: Contributions to the 11th Intl. Conference on Conceptual Structures ICCS'03*.

VALTCHEV P., MISSAOUI R., LEBRUN P. (2000). A Fast Algorithm for Building the Hasse Diagram of a Galois Lattice. *In Proceedings of the Colloque LaCIM*. Montréal.

Outil de gestion des connaissances d'une Interconnexion de Communautés de Pratique

Élise Garrot-Lavoué, Sébastien George et Patrick Prévôt

Université de Lyon, INSA Lyon, Laboratoire LIESP, F-69621 Villeurbanne, France,
{Elise.Garrot, Sebastien.George, Patrick.Prevot}@insa-lyon.fr

Résumé : Nous présentons dans cet article un modèle d'Interconnexion de Communautés de Pratique (ICP) ayant pour but de mettre en relation différentes Communautés de Pratique (CoPs) s'intéressant à une même activité générale. L'objectif est d'assurer la capitalisation de leurs connaissances de façon contextuelle. Ce modèle a été implémenté pour aboutir à la plate-forme TE-Cap qui a pour domaine d'application des activités de tutorat pédagogique. En particulier, nous proposons un outil d'indexation et de recherche des connaissances de l'ICP, outil offrant une combinaison de la structuration des taxonomies et de l'effet communautaire des folksonomies. Afin de valider ce travail, nous exposons les principaux résultats d'une expérimentation conduite en conditions réelles.

Mots-clés : Représentation des connaissances, Interconnexion de Communautés de Pratique, Web 2.0, Indexation des ressources, Contextualisation.

1 Introduction

Des Communautés de Pratique (CoPs) émergent lorsque des personnes échangent de façon informelle, s'entraident afin de résoudre des problèmes et développer leurs compétences et expertise. Des personnes peuvent appartenir à des CoPs au niveau local de leur entreprise ou institution. À ce niveau, les membres des CoPs échangent énormément en face-à-face pour résoudre des problèmes souvent très contextuels. Ces échanges sont peu instrumentés et les connaissances ainsi créées ne sont pas ou peu capitalisées. Les technologies Web ont permis le développement de CoPs en ligne regroupant des personnes n'appartenant pas aux mêmes entreprises ou institutions mais exerçant une même activité. À ce niveau plus général, les CoPs en ligne sont instrumentées (e.g. forums, blogs, wikis) mais les échanges sont pas ou peu structurés et non contextualisés. Les connaissances ainsi créées sont donc difficilement réutilisables, aucun contexte ne leur étant associé. Nos travaux ont pour buts (1) de faire la liaison entre ces deux types de CoPs en favorisant la mise en relation de personnes appartenant à des CoPs centrées autour d'une même activité générale et distribuées géographiquement et de (2) capitaliser toutes les connaissances produites en les contextualisant pour les rendre accessibles et réutilisables par tous les membres dans leur contexte de travail. Nous avons également pour objectif de favoriser le passage des connaissances d'une CoP à l'autre, pour éventuellement mener à la création de nouvelles connaissances.

Dans cet article, nous commençons par définir le concept de Communauté de Pratique (CoP) et plus spécifiquement de CoP en ligne. Nous nous appuyons sur le principe du Web 2.0 pour soutenir la construction de connaissances par les membres des CoPs. Nous proposons ensuite un modèle d'Interconnexion de CoPs (ICP), dans lequel les membres sont les noeuds des échanges de connaissances entre CoPs. Nous montrons la faisabilité de ce modèle par la réalisation de la plate-forme TE-Cap, destinée à mettre en relation et gérer les connaissances d'un ensemble de CoPs de tuteurs et formateurs de différentes institutions, pays et disciplines qui veulent échanger en ligne. Cette plate-forme offre un outil spécifique de gestion des connaissances d'une Interconnexion de CoPs reposant sur une classification évolutive. Nous validons nos travaux, d'une part, en montrant que cet outil répond à des besoins auxquels ne répondent pas les modes de représentation des connaissances existants et, d'autre part, en présentant une expérimentation qui a été menée pendant 5 mois en conditions réelles. Les travaux présentés seront illustrés tout au long de l'article par l'exemple du tutorat, que nous définissons comme l'accompagnement pédagogique des étudiants durant leur apprentissage.

2 État de l'art : Communautés de pratique et Web 2.0

2.1 Les Communautés de Pratique (CoPs)

Une Communauté de Pratique (CoP) est un type particulier de communauté qui rassemble des membres de manière informelle, pour une durée non déterminée, du fait qu'ils ont des pratiques, des centres d'intérêts et des buts communs (e.g. partager des idées et des expériences, construire des outils communs, développer des relations entre pairs) (Wenger, 1998). Les membres échangent et s'entraident pour développer leurs compétences et expertise afin de résoudre des problèmes. Ils développent une identité communautaire autour des connaissances partagées et des pratiques communes établies. Nous distinguons plusieurs types de CoPs : les CoPs locales auxquelles les personnes appartiennent au sein de leur organisation et les CoPs en ligne composées de personnes qui interagissent sur un espace en ligne (e.g. blog, wiki) (Koh & Kim, 2004).

Dans leur pratique quotidienne, les membres d'une CoP locale se donnent des conseils ou résolvent des problèmes concrets très précis liés à leur contexte de travail. Si l'on prend l'exemple des tuteurs au sein d'une même formation pédagogique, les échanges peuvent concerner la mise en place de scénarios, des impressions sur les apprenants ou encore l'évaluation de la formation qu'ils encadrent. Dans ce cas, les connaissances produites sont très contextuelles et sont rarement capitalisées du fait que les échanges ont surtout lieu oralement, autour d'un café ou dans les couloirs. Au sein des CoPs en ligne, les sujets abordés sont généraux à l'activité concernée, par exemple des problèmes rencontrés ou des conseils formulés, déconnectés du contexte dans lequel ils sont apparus, du fait que les personnes proviennent d'organisations différentes. Dans le cas des tuteurs, les échanges peuvent concerner le suivi des étudiants, les attitudes à adopter ou encore l'accompagnement de l'apprenant vers

l'autonomie. De plus, les outils utilisés ne permettent pas la contextualisation des échanges et la construction d'une connaissance partagée. Les listes de diffusion par exemple servent surtout à s'approprier de la connaissance sans avoir à en produire (Caviale, 2008). Les blogs sont principalement utilisés pour partager des histoires, des expériences et des opinions et minoritairement pour se mettre en relation avec d'autres personnes (Pashnyak & Dennen, 2007). Les systèmes tels que blogs, listes de diffusion, email et chat permettent seulement des discussions sans construction de sens concrets, les forums apportant un degré légèrement plus haut d'émergence explicite, grâce à la représentation spatiale en fils de discussions qui font ressortir les relations entre les messages.

Ainsi, des personnes exerçant une même activité, par exemple le tutorat, peuvent avoir des pratiques similaires sans être nécessairement au courant, principalement du fait qu'ils n'appartiennent pas à la même entreprise ou institution. S'ils n'appartiennent pas à des CoPs en ligne, ils ne vont pas interagir et échanger sur leurs pratiques et vont développer leurs propres pratiques, chacun réinventant ce qui se fait certainement déjà ailleurs. Dans le cas contraire, ils vont échanger et s'entraider mais les connaissances créées perdent leur sens puisqu'elles deviennent indépendantes de tout contexte. Pour répondre à ce problème, nous proposons dans la partie 3 un modèle d'Interconnexion de CoPs soutenant la mise en relation de CoPs centrées autour d'une même activité. Nous montrons dans la partie suivante une approche de la gestion des connaissances des CoPs fondée sur le principe du Web 2.0.

2.2 Le Web 2.0

Le concept de Web 2.0 (O'Reilly, 2005) fait référence à une évolution des usages des technologies du Web plutôt qu'à de nouvelles technologies, l'utilisation du Web s'orientant de plus en plus vers l'interaction entre les utilisateurs et la création de réseaux sociaux. En particulier, sont qualifiées de Web 2.0 les applications permettant aux internautes d'interagir à la fois avec le contenu des pages mais aussi entre eux. L'un des principes généraux de conception pour des plates-formes Web est de tirer partie de l'intelligence collective. C'est par exemple en laissant des commentaires sur des messages postés ou en liant des ressources (messages, blogs, wikis) par des liens Web, que vit et se développe un soutien à une communauté en ligne (Dennen & Pashnyak, 2007). La discussion et l'engagement dans la communauté peuvent être vus comme un agrégat de commentaires entre des personnes interreliées.

Partant de ce principe, nous adhérons à une vision de la construction informelle des connaissances et des liens entre ces connaissances par les membres des CoPs en ligne eux-mêmes. L'objectif est de concevoir une plate-forme se nourrissant de la participation des utilisateurs qui apportent leur propre contenu et de l'activité collective ressortant, d'une part, sous forme de liens hypertextes entre les contenus Web ou de commentaires et d'autre part, sous forme de mots-clefs qui sont associées aux contenus par les utilisateurs pour en décrire le contexte. La formalisation du contexte d'un message par son auteur entraîne un processus de réflexivité sur la pratique ou l'expérience dont il témoigne ou à laquelle il répond, processus participant au développement professionnel (Barak, 2006). Nous considérons la

formalisation que les personnes effectuent du contexte des ressources produites comme des méta-connaissances.

3 Le modèle d'Interconnexion de CoPs (ICP)

3.1 Le modèle général

Dans cette partie, nous proposons un modèle d'Interconnexion de Communautés de Pratique (cf. Fig. 1) ayant pour but de mettre en relation des CoPs aussi bien locales (i.e. pour les tuteurs : leurs propres écoles, départements d'enseignement ou institution) que générales (CoPs en ligne) de personnes pratiquant une même activité (i.e. le tutorat). Ce modèle est basé sur l'hypothèse que considérer un ensemble de personnes pratiquant une même activité générale comme appartenant à des CoPs interconnectées peut favoriser leur mise en relation, leur participation et la création de connaissances. Nous proposons de voir ce groupe d'acteurs non pas comme une seule entité délimitée par un domaine d'activité mais comme « *un ensemble de CoPs soutenues par une plate-forme Web, reliées entre elles par les membres qui les composent, ceux-ci étant les nœuds de pratiques interconnectées* ». Nous précisons qu'une CoP n'est pas définie par un métier mais par des pratiques communes.

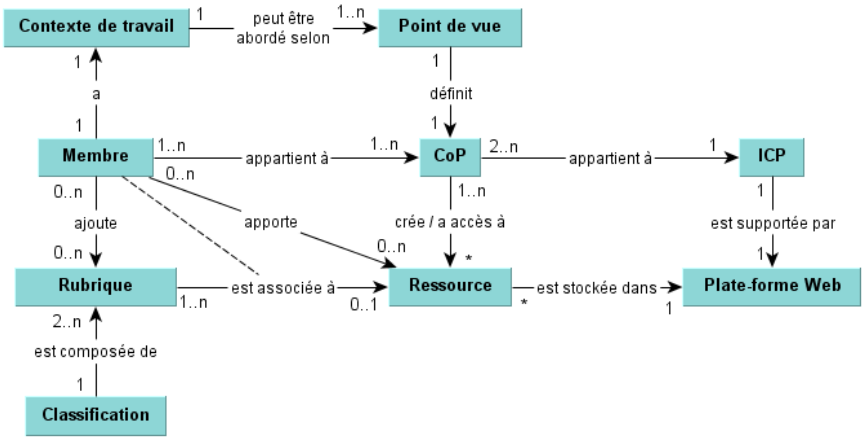


Fig. 1 – Modèle général d'une Interconnexion de Communautés de Pratique (ICP)

À un niveau individuel, l'activité d'un acteur peut être abordée selon de multiples points de vue dépendant de son contexte de travail. Les CoPs auxquelles il appartient sont définies par son contexte de travail, une CoP correspondant au niveau élémentaire de pratique d'un acteur. Au niveau général, une ICP est composée de toutes les CoPs élémentaires définies par tous les acteurs qui participent sur la plate-forme Web. Nous pouvons la voir comme une seule communauté d'acteurs pratiquant une même activité, rassemblés sur la même plate-forme et qui peut être abordée selon de multiples points de vue et par de multiples entrées. Les ressources de l'ICP sont

stockées dans la base de données selon une classification hiérarchique basée sur un modèle des pratiques des acteurs (cf. §3.2). Ces ressources contextualisées sont les connaissances partagées par l'ensemble des CoPs soutenues par la plate-forme.

Par exemple (cf. Fig. 2), le tuteur 1, travaillant au Département Génie Industriel (GI) à l'INSA de Lyon en France et qui encadre des projets collectifs en maintenance industrielle peut appartenir à 5 CoPs : des tuteurs qui encadrent des activités collectives, des tuteurs en maintenance industrielle, des tuteurs qui encadrent des activités de type projet, des tuteurs du département GI et des tuteurs de l'INSA de Lyon. Le tuteur 2 appartenant à une autre institution, par exemple la TéléUniversité du Québec (Téluq) au Canada, peut appartenir à plusieurs CoPs, certaines étant les mêmes que celles auxquelles appartient le tuteur 1. Ces deux tuteurs, de différents pays, vont être mis en relation du fait que leur contexte de travail peut être abordé selon des points de vue similaires, ce qui implique qu'ils appartiennent à des mêmes CoPs. Le tuteur 3 va être mis en relation avec les tuteurs 1 et 2 du fait qu'il appartient au même établissement et à la même formation que le tuteur 1 et qu'il encadre le même type d'activité que le tuteur 2. Cet exemple illustre donc le fait que les tuteurs sont les nœuds de l'Interconnexion de CoPs.

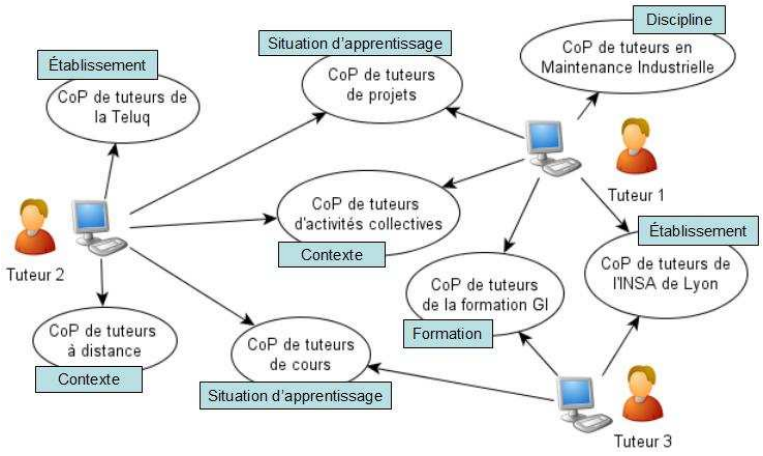


Fig. 2 – Les tuteurs en tant que nœuds de l'Interconnexion de CoPs

Dans cet exemple, l'activité des tuteurs peut être abordée selon plusieurs points de vue : le contexte de l'activité (collective, à distance), la situation d'apprentissage (projet, cours), la discipline (maintenance industrielle), la formation (Génie Industriel) et l'établissement (INSA Lyon, Téluq). Ces points de vue sont des catégories de CoPs et nous proposons dans la section suivante une approche pour définir un modèle des pratiques des acteurs, qui implique de déterminer toutes les catégories de CoPs et CoPs correspondant à une activité donnée.

3.2 Gestion et diffusion des connaissances d'une ICP

Les connaissances d'une ICP sont classées selon une classification hiérarchique reposant sur un modèle des pratiques des acteurs. Pour les tuteurs, nous avons élaboré

un modèle de leurs pratiques réparti en au plus 4 niveaux, le premier niveau correspondant aux facteurs principaux qui différencient les pratiques des acteurs (e.g. l'institution, la formation, la discipline, l'activité) (Garrot, 2008). Ces facteurs correspondent aux catégories principales de CoPs. Chaque catégorie est ensuite divisée en sous-catégories et ainsi de suite. Les nœuds terminaux correspondent aux CoPs. Cette taxonomie du tutorat a été développée de façon itérative, à partir d'entretiens avec 6 tuteurs (premier cycle de conception) et des résultats obtenus lors de l'expérimentation d'un premier prototype (deuxième cycle de conception). Elle n'a pas à être exhaustive car c'est seulement une base amenée à évoluer par modifications ou additions par les membres de l'ICP eux-mêmes.

À la création d'une ressource (message, document, lien Web), l'auteur doit l'associer à une ou des CoPs (en choisissant une rubrique du niveau le plus bas de la classification). Quand ils trouvent une ressource (résultat d'une requête), les membres de l'ICP peuvent également associer de nouvelles rubriques à cette ressource pour la diffuser à de nouvelles CoPs. Ils peuvent l'associer soit à une CoP soit à une catégorie de CoPs (rubriques des niveaux plus élevés de la classification) pour diffuser la ressource à toutes les CoPs filles. En effet, les CoPs filles héritent de toutes les ressources d'une catégorie de CoPs. Ainsi, la participation des membres de l'ICP ne consiste pas seulement à créer de nouvelles ressources mais aussi à créer des liens entre ces ressources. La pertinence de ces liens est estimée par les membres eux-mêmes qui considèrent qu'une ressource peut être utile ou intéressante pour une CoP. L'ajout d'une ressource à une CoP peut mener à un débat sur cette ressource et éventuellement à la création de nouvelles ressources pour cette CoP. Des événements rapportés dans un contexte précis peuvent mener à un partage d'expériences, pouvant être utilisées comme base pour générer des règles ou des recommandations qui deviennent des connaissances globale de l'ICP.

4 Proposition d'un outil de gestion des connaissances

Nous avons développé la plate-forme TE-Cap selon une approche co-adaptative par un processus itératif incluant 3 cycles de conception, chaque cycle reposant sur le développement d'une maquette ou prototype, sur son évaluation par les utilisateurs au moyen d'entretiens ou expérimentations et sur l'interprétation des résultats de l'activité des utilisateurs (Garrot *et al.*, 2008). Cette démarche a eu pour but de faire émerger les besoins des utilisateurs, de les amener à les expliciter et de faire évoluer les spécifications de la plate-forme afin de répondre au plus près à ces besoins. Nous nous sommes tout particulièrement attachés à développer un outil d'indexation et de recherche des connaissances pour une ICP que nous présentons.

4.1 Gestion du profil utilisateur

L'outil d'indexation et de recherche des connaissances repose sur le profil de l'utilisateur pour une personnalisation des rubriques qui lui sont proposées. L'utilisateur définit son profil en remplissant des champs correspondant à des

catégories de CoPs de la classification hiérarchique. Les valeurs données aux champs définissent des CoPs et impliquent l'appartenance de l'utilisateur à ces CoPs. Le profil est constitué de 3 caractéristiques principales : le profil identitaire, le contexte de travail et les centres d'intérêt secondaires. Le contexte de travail concerne toutes les CoPs directement liées au contexte de travail de l'utilisateur, alors que les centres d'intérêts secondaires concernent toutes les CoPs qui ne sont pas directement liées à son contexte de travail mais qui peuvent l'intéresser (donne accès à des ressources susceptibles de l'intéresser et au profil de personnes partageant des pratiques et expériences similaires).

Pour permettre une utilisation de l'outil par les membres d'une CoP dans leur pratique quotidienne, celui-ci leur offre un accès rapide aux ressources pertinentes pour eux, ceci par deux moyens (cf. Fig. 3). Premièrement, un lien entre l'interface de recherche et le profil utilisateur permet de ne montrer à l'utilisateur que les rubriques de la classification qui le concernent et qui l'intéressent en fonction de son profil. Il n'a ainsi accès qu'aux ressources des CoPs auxquelles il a déclaré appartenir et ne peut créer des ressources que pour celles-ci. Deuxièmement, l'utilisateur a la possibilité, selon son but en se connectant à la plate-forme, d'appliquer un filtre pour ne montrer sur l'interface de classification que les rubriques liées à son contexte de travail ou à ses thèmes d'intérêt secondaires. Dans sa pratique quotidienne, il est pertinent de proposer en premier à l'utilisateur uniquement les rubriques qui concernent directement son contexte de travail. S'il ne trouve pas l'information qu'il recherche, il peut élargir la recherche aux autres thèmes d'intérêt liés à son activité.

4.2 Outil de recherche et d'indexation des connaissances

Les interfaces de recherche et d'indexation des connaissances (messages et profils de membres) reposent sur la classification des ressources de l'ICP (cf. Fig. 3). Une section dynamique (au centre de l'écran) est composée de trois onglets permettant une navigation aisée et rapide entre la classification et les résultats d'une requête. L'onglet « Recherche » permet de naviguer dans la classification et de sélectionner des rubriques pour une requête. Ces rubriques sont représentées sous la forme de bulles, ce qui apporte convivialité et attractivité à l'interface. L'utilisateur peut naviguer dans la classification en double-cliquant sur une bulle, ce qui l'éclate en bulles représentant les sous-rubriques. Les rubriques de dernier niveau (correspondant aux CoPs) sont représentées sous la forme d'un champ à sélection multiple. Les utilisateurs peuvent retourner aux niveaux supérieurs à l'aide du chemin de navigation. La plate-forme propose la même interface pour rechercher des ressources et des profils de membres, tout en les distinguant par deux onglets, pour que l'utilisateur puisse, à chaque requête, consulter des profils et « découvrir » des personnes qui ont des pratiques similaires ou qui offrent une expertise.

La partie à la droite de l'écran donne la possibilité de stocker les rubriques choisies pour effectuer des requêtes (par un simple « *drag and drop* » depuis la partie centrale), permettant ainsi à l'utilisateur d'affiner ou élargir sa requête en fonction des résultats obtenus (en sélectionnant ou désélectionnant les rubriques). Les rubriques dans cette colonne sont toujours visibles quand l'utilisateur navigue dans les onglets

de la partie dynamique, ainsi que d'une requête à l'autre. L'utilisateur peut déplacer les bulles dans la colonne de recherche pour modifier l'ordre des rubriques selon sa préférence et également supprimer une rubrique en déplaçant la bulle en dehors de la colonne. Le principe de sélection des rubriques dans cette colonne peut être comparé à celui des paniers sur les sites Web commerciaux. Cette interaction homme-machine originale a été choisie pour favoriser la navigation dans la classification et pour simplifier la sélection des rubriques de recherche.

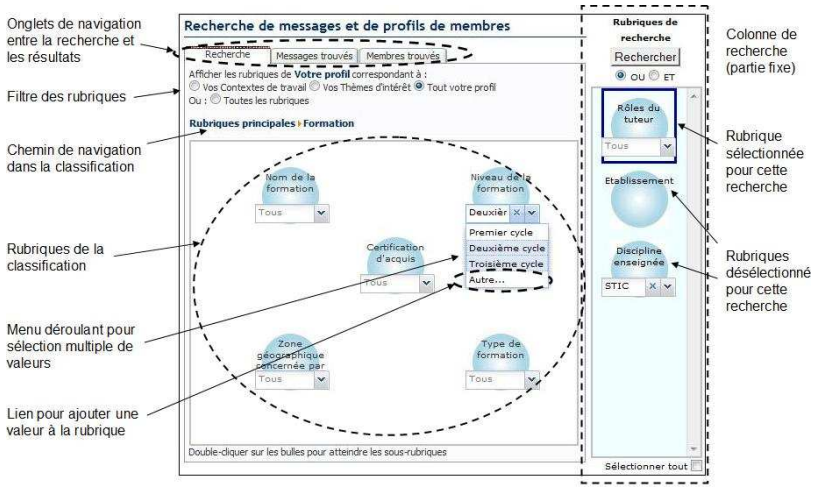


Fig. 3 – Interface de recherche des messages et profils de membres

L'indexation d'un message initiatif d'une discussion se fait selon le principe suivant : l'utilisateur associe des rubriques au message en même temps qu'il l'écrit en déplaçant des bulles de la partie centrale (rubriques de la classification) à la colonne de classement. Ce principe a pour but de l'amener à avoir une réflexion sur l'expérience ou le témoignage qu'il écrit et de lui permettre de clarifier le contexte du message en même temps que se précise la pensée qu'il exprime. Pour faciliter cette action, une interface sous la forme d'onglets assure une navigation facile entre l'écriture du message et son classement. Tout utilisateur peut associer de nouvelles rubriques à une discussion, une régulation étant assurée par l'auteur qui a le droit de supprimer les rubriques qu'ils ne considèrent pas pertinentes pour la discussion.

4.3 Évolution de la classification

Les utilisateurs peuvent faire évoluer la classification par leur participation sur la plate-forme, pour aboutir à une classification utilisant un vocabulaire au plus près de leurs pratiques. Pour cela, l'interface donne à tout moment la possibilité d'ajouter de nouvelles rubriques à la classification, que ce soit en remplissant le profil, en classant une ressource, en recherchant une ressource ou en visualisant une ressource. Les rubriques utilisées sont sauvegardées, permettant de supprimer celles jugées inutiles. Les rubriques non utilisées sont ensuite supprimées, signifiant qu'elles ne sont pas adaptées au domaine de pratique des acteurs ou pas situées au bon niveau de la

classification. Cette évolution des rubriques est nécessaire pour que la classification effectuée *a priori* suive l'évolution des usages et des pratiques des acteurs.

5 Validation

5.1 Un intermédiaire entre taxonomies et folksonomies

Nous discutons dans cette partie de l'apport de l'outil présenté, reposant sur un mode de représentation des connaissances structuré et évolutif, par rapport aux outils et modèles de représentation des connaissances existants. D'une part, il existe des systèmes d'organisation des connaissances reposant sur des représentations conceptuelles plus ou moins structurées, telles que les taxonomies et ontologies. Les taxonomies, représentation hiérarchique de la connaissance sous la forme de catégories, proposent une manipulation assez simple et intuitive mais amènent à différentes interprétations suivant les individus. L'ontologie est « la spécification d'une conceptualisation d'un domaine de connaissance » (Gruber, 1995). Les ontologies sont employées pour raisonner à propos des objets d'un domaine concerné, à l'aide d'outils informatiques dédiés, et servent souvent à l'indexation automatique des contenus avec des méthodes du Web Sémantique (Berners-Lee *et al*., 2001). Ces modèles ont pour avantage de représenter des connaissances de manière structurée. Mais ils sont généralement construits par des experts du domaine concerné : ils nécessitent un consensus qui peut être coûteux et long à atteindre (Brooks *et al*., 2006) et ne sont pas forcément utilisables et compréhensibles par ceux qui ne les ont pas construits, surtout pour les utilisateurs novices dans le domaine. Le caractère consensuel de ces modèles donne le moyen à la communauté qui les a construits de se comprendre pour produire et diffuser des connaissances mais il impose un sens *a priori* aux connaissances produites. C'est pourquoi nous nous interrogeons sur la possibilité de création et de négociation de sens par une CoP en évolution et avançons qu'un modèle de représentation des connaissances n'est pas une fin en soi et doit être capable d'évoluer au cours du temps.

D'autre part, il existe les folksonomies, sorte de catégorisation collaborative des contenus Web utilisant des mots clés (ou « tag ») librement choisis (Limpens *et al*., 2008). Ce concept est considéré comme faisant partie intégrante du Web 2.0 et est exploité par de nombreux sites Web pour le partage de ressources, comme des photos, des images vidéo, des pages Web, des signets, etc. Des systèmes supports à des CoPs utilisent les tags, par exemple pour annoter les contenus d'un wiki (El Ghali *et al.*, 2007) ou construire des profils utilisateurs (Diederich & Iofciu, 2006). L'intérêt des folksonomies est lié à l'effet communautaire : pour une ressource donnée, sa classification est l'union des classifications de cette ressource par les différents contributeurs. L'avantage est de permettre à beaucoup d'utilisateurs d'associer des tags à une ressource et d'enrichir sa description. Mais ce système de « tagging » trouve sa limite dans son manque de structure (Guy & Tonkin, 2006) que certains travaux tendent à compenser en donnant la possibilité aux utilisateurs de déterminer des relations entre les tags (Huynh-Kim-Bang & Dané, 2008), amenant ainsi une

structuration progressive des tags. Mais cette approche n'est appliquée qu'à l'indexation de documents et n'a pas encore été validée par expérimentation en usage réel. Dans le cadre d'une CoP, nous considérons nécessaire d'apporter une structure aux utilisateurs pour indexer et rechercher les connaissances. Les systèmes de tags fonctionnent bien pour des communautés d'intérêt où les utilisateurs vont naviguer dans l'application sans but précis. Mais ces systèmes ne sont pas vraiment adaptés à une CoP où les utilisateurs recherchent des ressources pertinentes liées à leur contexte de travail et doivent pouvoir les retrouver très rapidement pour s'en servir dans leur pratique quotidienne.

L'outil que nous avons développé offre une combinaison de la structuration des taxonomies et de l'effet communautaire des folksonomies. En effet, il apporte une structuration des connaissances, point essentiel dans le cadre d'une CoP, tout en supportant une nécessaire évolutivité avec indexation des contenus par les utilisateurs eux-mêmes. L'interface offre convivialité, attractivité et simplicité d'utilisation, tel que défini par les principes du Web 2.0 pour un environnement communautaire. La nécessité de remplir le profil peut être contraignante pour les utilisateurs mais elle est compensée par l'efficacité de recherche offerte (rapidité et pertinence). En comparaison avec les folksonomies, l'outil permet une recherche à partir de plusieurs rubriques de différents niveaux de la classification, une personnalisation avec, d'une part, une présentation des rubriques renseignées dans le profil uniquement et, d'autre part, un filtrage des rubriques par contexte de travail ou thèmes d'intérêt secondaires.

5.2 Expérimentation

Nous avons mené une expérimentation en conditions réelles, du 25 février au 5 juillet 2008 en situation réelle. L'adresse Web de la plate-forme TE-Cap a été diffusée à 3 CoPs d'enseignants et tuteurs, à 7 campus en ligne et aux utilisateurs d'un premier prototype de TE-Cap (Garrot *et al.,* 2008), espérant qu'ils incitent les autres utilisateurs à participer. Les flux de discussions créés lors de la première expérimentation ont été conservés pour servir de base à de nouvelles discussions. Pour aider à la compréhension du fonctionnement de l'interface, nous avions mis en ligne des vidéos de démonstration. Cette expérimentation avait pour but de valider la plate-forme TE-Cap comme soutien à l'Interconnexion de CoPs de tuteurs. Nous avions défini des indicateurs pour évaluer la sociabilité, les niveaux de partage et de création de connaissances et l'utilisabilité et utilité de la plate-forme (Garrot, 2008). Les résultats provenaient de trois types de données : des traces d'utilisation, des réponses à un questionnaire en fin d'expérimentation et des tests d'utilisabilité.

42 personnes de 9 pays francophones se sont inscrites sur TE-Cap. Nous ne présentons ici que les principaux résultats concernant l'outil d'indexation et de recherche des connaissances. Tout d'abord, les réponses au questionnaire montrent que notre objectif de mettre en relation des CoPs locales et générales répond à un besoin existant puisque les tuteurs recherchent des informations aussi bien au niveau local de leur formation qu'à un niveau plus général de leur activité (leurs rôles, les outils pédagogiques, les apprenants). Cependant, assez peu de messages ont été écrits (15) bien que 27 utilisateurs aient lu des discussions. Ceci s'explique par le fait que,

d'après les réponses au questionnaire, les utilisateurs se sont inscrits autant par curiosité pour un nouvel outil que pour réellement participer. De plus, la participation à une communauté sera toujours moins prioritaire qu'enseigner ou tutorer. Un résultat positif est le nombre assez important de rubriques ajoutées à la classification (45 par 19 utilisateurs), ce qui implique une évolution significative de la classification et ainsi une appropriation par les utilisateurs. Enfin, des tests d'utilisabilité, effectués auprès de 3 tuteurs suivant un scénario, ont mis en avant le fait que les interfaces d'indexation et de recherche de TE-Cap sont très faciles à utiliser et efficaces. Mais l'utilisation de ces interfaces nécessite une étape d'apprentissage, ce qui est normal pour une interface innovante qui propose de nouvelles fonctionnalités. De plus, 23 utilisateurs n'ont pas rempli ou utilisé leur profil, ce qui nous laisse supposer qu'ils n'en n'ont pas vu l'intérêt ou n'ont pas pris le temps (nécessite 5 à 10 minutes, comme l'ont montré les tests d'utilisabilité). La cause mise en avant est qu'ils n'ont pas compris le lien entre le profil et la classification proposée et il serait nécessaire de mieux expliquer ce lien pour qu'ils en voient l'intérêt. L'aide apportée par les vidéos n'est pas suffisante ou pas adaptée et une amélioration pourrait être l'addition d'une aide contextuelle ou d'un compagnon logiciel.

De plus amples résultats ne pourront être obtenus que par une utilisation par un grand nombre de personnes et sur le long terme. Ce n'est que dans ces conditions que la plate-forme et les outils proposés révèleront leur potentiel.

6 Conclusion et perspectives

Dans cet article, nous avons défini un modèle d'Interconnexion de CoPs pour assurer la mutualisation et la diffusion des connaissances de CoPs locales s'intéressant à une même activité générale, dans notre cas le tutorat. Nous avons validé l'implémentation de ce modèle par le développement de la plate-forme TE-Cap. De plus, nous avons conduit une expérimentation en conditions réelles pendant plusieurs mois avec des tuteurs de différentes disciplines et pays. Les résultats ont mis en avant la facilité d'utilisation et l'utilité de l'outil de recherche et indexation des connaissances, bien que toutes les possibilités offertes n'aient pas été utilisées.

Plusieurs perspectives sont ouvertes par nos travaux. La première est de créer, en plus des relations de hiérarchisation, des associations sémantiques entre rubriques (discipline, activité,…). Ces associations peuvent être faites soit par les utilisateurs, soit par le système lui-même, au moyen par exemple d'une recherche des occurrences de termes désignant des rubriques et souvent utilisés ensemble dans un même message. Le système peut proposer des relations sémantiques que les utilisateurs peuvent valider, modifier, supprimer ou compléter. Une perspective à plus long terme est le développement d'un outil de recherche de connaissances indépendant de toute plate-forme, afin de mettre en relation différentes CoPs en ligne qui existent déjà et de mutualiser les connaissances contenues sur chacune de leur plate-forme. L'objectif sera de concevoir l'outil dans une approche orientée service. L'atteinte de cet objectif serait un aboutissement car il permettrait de concrétiser le concept d'ICP pour des CoPs n'utilisant pas les mêmes plates-formes.

Références

Barak, M. Instructional principles for fostering learning with ICT: teachers' perspectives as learners and instructors. *Education and Information Technologies*, 2006, vol. 11, n°2, p. 121–135.

Bateman, S, Brooks, C. & McCalla, G. Collaborative Tagging Approaches for Ontological Metadata in Adaptive ELearning Systems. *Fourth International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL 2006)*. Dublin, Ireland, 2006, p. 3-12.

Berners-Lee, T., Hendler, J. & Lassila, O. The semantic web. *Scientific American Magazine*, 2001, vol. 279, n°5. Disponible sur : http://www.sciam.com/article.cfm?id=the-semantic-web (consulté le 23.01.2009).

Caviale, O. Analyse d'une liste de discussion d'enseignants. Un reflet des normes personnelles ou institutionnelles ? *Journées Communication et Apprentissage Instrumentés en Réseau (JOCAIR)*. Amiens, 2008, p. 137-148.

Dennen, V.P. & Pashnyak, T. Finding Community in the Comments: the Role of Reader and Blogger Responses in a Weblog Community of Practice. *IADIS Web Communities Conference*. Salamanca, Spain, 2007, p. 11-17.

Diederich, J. & Iofciu, T. Finding Communities of Practice from User Profiles Based On Folksonomies. *1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06)*. Crete, Greece, 2006, p. 288-297.

El Ghali, A., Tifous, A., Buffa, M., Giboin, A. & Dieng-Kuntz, R. Using a Semantic Wiki in Communities of Practice. *2nd International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice*. Crete, Greece, 2007, p. 22-31.

Garrot, E. *Plate-forme support à l'Interconnexion de Communautés de Pratique (ICP). Application au tutorat avec TE-Cap*. INSA de Lyon, 2008, 309 p.

Garrot, E., George, S. & Prévôt, P. Supporting a Virtual Community of Tutors in Experience Capitalizing. *International Journal of Web Based Communities*, 2008, vol. 5, n°3, in press.

Gruber, T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*, 1995, vol. 43, n°5-6, p. 907-928.

Guy, M. & Tonkin, E. Folksonomies: Tidying up Tags? *D-Lib Magazine*, 2006, vol. 12, n°1. Disponible sur : http://www.dlib.org/dlib/january06/guy/01guy.html (consulté le 23.01.2009).

Huynh-Kim-Bang, B. & Dané, E. Social bookmarking et tags structurés. *19èmes Journées Francophones d'Ingénierie des Connaissances*. Nancy, France, 2008, p. 111-122.

Koh, J. & Kim, Y. Knowledge sharing in virtual communities: an e-business perspective. *Expert Systems with Applications*, 2004, vol. 26, n°2, p. 155–166.

Limpens, F., Gandon, F. & Buffa, M. Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées : un état de l'art. *19èmes Journées Francophones d'Ingénierie des Connaissances*. Nancy, France, 2008

O'Reilly, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *O'Reilly Media*, 2005. Disponible sur : http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (consulté le 23.01.2009).

Pashnyak, T.G. & Dennen, V.P. What and Why do Classroom Teachers Blog? *IADIS Web Based Communities Conference*. Salamanca, Spain, 2007, p. 172-178.

Wenger, E. *Communities of practice: Learning, meaning, and identity*. Cambridge : Cambridge University Press, 1998, p. 336.

Sémantique des folksonomies: structuration collaborative et assistée

Freddy Limpens[1], Fabien Gandon[1], Michel Buffa[2]

[1] Edelweiss, INRIA Sophia-Antipolis, France, 2004 route des lucioles - BP 93, FR-06902 Sophia-Antipolis Cedex
{freddy.limpens, fabien.gandon}@sophia.inria.fr

[2] KEWI, Laboratoire I3S, Université de Nice, France
buffa@unice.fr

Résumé : L'essor du tagging et des folksonomies pour l'organisation des ressources partagées au sein du Web social et collaboratif constitue une opportunité pour l'acquisition des connaissances par ceux-là même qui les manipulent. Cependant l'absence de liens sémantiques entre les tags, ou la variabilité d'écriture de certains tags appauvrissent les potentiels de navigation et de recherche d'information. Pour remédier à ces limitations, nous proposons d'exploiter l'interaction entre les utilisateurs et les systèmes à base de folksonomies pour valider ou invalider des traitements automatiques effectués sur les tags. Ces opérations se basent sur notre modèle pour l'assistance à la structuration des folksonomies qui autorise des vues conflictuelles portant sur les liens entre les tags, tout en permettant aux concepteurs des systèmes d'exploiter la diversité de ces descriptions sémantiques afin d'offrir des fonctionnalités de navigation enrichies.
Mots-clés : Folksonomies, Ontologies, Partage de Connaissances

1 Introduction

Le *social tagging* s'est récemment imposé dans le paysage du web social et collaboratif (Web 2.0) comme support à l'organisation de ressources partagées en permettant aux utilisateurs de catégoriser ces ressources, simplement en leur associant des mots clefs, appelés tags. Les folksonomies constituent le résultat de la collecte de tags ainsi créés, c'est à dire associés à des ressources par des utilisateurs. L'exploitation des folksonomies pour la recherche d'informations et de ressources pose néanmoins quelques problèmes. La variabilité d'écriture entre certains tags équivalents (comme "électricité" et "electricite"), ou l'absence de liens sémantiques entre les tags sont pénalisantes lors d'une recherche de ressources par tags. De nombreux travaux de recherche, recensés dans une précédente contribution (Limpens *et al.*, 2008), tentent de dépasser les limitations des folksonomies en les rapprochant de représentations sémantiquement struc-

turées. Ainsi, certains tentent de constituer des "ontologie légères" [1] à partir de folksonomies (Mika, 2005), ou d'assister les cycles de vies des ontologies en les nourrissant des notions extraites des folksonomies (Passant, 2007). D'autres approches proposent d'impliquer les usagers directement dans la construction d'ontologies légères basées sur les formalismes du Web Sémantique (Braun *et al.*, 2007), ou sur le modèle HyperTopic du Web Socio-Sémantique (Cahier *et al.*, 2007).

Notre contribution se concentre sur les systèmes à base de folksonomies outillant le partage de connaissances au sein de groupes de personnes appartenant à des réseaux d'intérêts communs, et regroupés autour de l'usage d'une même plateforme. Le type de système que nous envisageons permet à ses utilisateurs de contribuer au partage, au commentaire, à l'indexation, et à l'élaboration de documents de natures diverses (photos, bookmarks, pages de wiki, etc.). En tant que *ressource* le bookmark a un statut particulier, car il constitue à la fois un support pour documenter les traces de lectures, mais également, et comme par effet secondaire dès qu'il est partagé et associé à des tags, une opportunité pour l'indexation et la (multi)catégorisation collaborative. A ce titre, nous pensons que les pratiques de bookmarking social du Web 2.0 peuvent être adaptées à l'échelle plus restreinte des organisations et des communautés d'intérêts. Le projet ANR ISICIL auquel nous participons s'intéresse notamment à une adaptation de ces outils et usages pour les appliquées à la veille technologique et économique.

Dans cet article nous proposons des méthodes pour constituer des ontologies "légères" qui peuvent être exploitées, par exemple, pour suggérer des termes "sémantiquement" proches lors d'une recherche de documents par tags, ou encore pour enrichir les résultats d'une requête par les variantes orthographiques d'un même tag, ou *spelling variant*, comme "écologie" et "ecologie". Pour atteindre cet objectif, nous proposons d'allier des traitements automatiques sur les folksonomies et l'expertise des utilisateurs en leur proposant, d'une part, de valider ou invalider les résultats de ces traitements, et d'autre part, de suggérer certaines propriétés sémantiques entre les tags à travers des fonctionnalités simples et ergonomiques de l'interface. Ce système est basé sur notre modèle qui, plutôt que de répercuter et représenter explicitement les traitements automatiques des tags, permet, dans un premier temps, de recueillir les résultats de ces traitements ainsi que les opérations de validation des utilisateurs. L'exploitation de ces résultats est alors repoussée aux étapes ultérieures, par exemple lors du tri ou du filtrage des réponses à une requête.

Notre article est organisé de la manière suivante. Dans la section 2, nous présentons les principales méthodes proposées pour établir des liens sémantiques entre les tags d'une folksonomie, avant de détailler notre méthode pour l'enrichissement sémantique des folksonomies. Nous présentons ensuite dans la section 3 l'implantation de cette méthode dans un système de gestion et d'exploration de bookmarks tagués et partagés. Puis, en section 4, nous positionnerons et discuterons les apports de notre approche, avant de conclure en section 5.

1. Gandon (2008) définit les ontologies légère comme étant des "ontologies qui ne comportent typiquement pas ou peu de définitions formelles et qui se focalisent souvent sur la représentation de hiérarchies de types ne nécessitant pas des langages très expressifs (ex : RDFS)"

2 Présentation de notre approche

2.1 Scénario d'application

Ce travail trouve un terrain d'application dans le cadre d'un partenariat avec l'Ademe (Agence pour l'Environnement et la Maîtrise de l'Energie) où nous cherchons à évaluer la validité des usages liés aux outils du Web 2.0 dans le contexte d'une organisation professionnelle. L'un des scénarios d'application envisagés ici est l'assistance aux pratiques de veille, et plus particulièrement la recherche d'information et le partage de ressources au sein d'un groupe d'experts. Nous souhaitons ainsi promouvoir l'usage du *bookmarking* et du *tagging social* des ressources internes et partagées, ainsi qu'une intégration, dans les tâches quotidiennes des usagers, de l'organisation sémantique des folksonomies.

2.2 Traitements sur les folksonomies

Une des limitations couramment reconnue aux folksonomies (Mathes, 2004) est la variabilité d'écriture des tags supposés équivalents comme "écologie" et "ecologie". Une solution possible pour traiter ce problème consiste à mesurer la distance d'édition entre les tags (par exemple de type distance de Levenshtein (1966)), et au delà d'un certain seuil, de considérer ces tags équivalents. Specia & Motta (2007) ont appliqué cette méthode sur un extrait de la folksonomie de delicious.com, et exploité des bases de connaissances externes (Wordnet) et quelques règles simples pour sélectionner le libellé du tag le plus représentatif de ses variantes orthographiques.

Un autre type de traitement des folksonomies consiste à mesurer la distance de "similarité" entre tags en se basant sur les liens entre les tags, les ressources, et les utilisateurs (Mika, 2005). (Cattuto *et al.*, 2008) distinguent sur ce point différents types de mesures de similarité : les mesures basées sur une fréquence de cooccurrence "simple" de deux tags sur une même ressource, ou les mesures distributionnelles, qui prennent en compte trois types de contextes d'association des tags. Chaque contexte correspond à un espace vectoriel prélevé dans l'espace vectoriel global de la folksonomie. Les mesures distributionnelles prennent en compte l'association des tags : (1) *via* leur usage par un même utilisateur (contexte utilisateur-tag), ou (2) *via* leur usage pour une même ressource (contexte ressource-tag), ou (3) *via* leur associations communes avec d'autres tags (contexte tag-tag).

Afin de caractériser en termes sémantiques ces différentes mesures de similarités entre tags, Cattuto *et al.* proposent d'exploiter la structure hiérarchique de Wordnet (Fellbaum, 1998) pour les tags dont le libellé est présent dans cette base lexicale. L'issue de cette expérience montre que les tags associés *via* des mesures de cooccurrences simples tendent à entretenir des relations de subsomption, alors que les tags associés *via* une mesure distributionnelle de similarité dans le contexte tag-tag tendent à se situer au même niveau hiérarchique, soit partageant le même parent, soit le même grand-parent. Cattuto *et al.* expliquent que l'association des tags *via* leur cooccurrence sur une même ressource renvoie à leur utilisation simultanée dans le même acte de tagging où l'usager a tendance à couvrir différents niveaux de généralité. Par exemple, les tags "java" et "programming", ou encore "tobuy" et "shopping" sont fréquemment utilisés

simultanément, et on peut supposer que, du point de vue du "tagueur", ces tags ont des niveaux différents de généralité. Le lien mesuré par la mesure distributionnelle dans le contexte tag-tag associe des tags ayant des schémas de cooccurrence similaires mais qui ne sont que peu ou pas utilisés simultanément. Ce cas de figure correspond par exemple aux tags "tobuy" et "whishlist" qui ne sont pas utilisés simultanément mais plutôt conjointement avec le tag "shopping".

La principale limite à l'utilisation de Wordnet comme base de connaissance est que cette ressource termino-ontologique inclue peu de termes spécifiques à un domaine, alors qu'ils sont fréquents dans les folksonomies. Des ressources plus spécifiques à un domaine pourrait donc permettre d'élargir la portée de la validation sémantique des liens de similarités entre certains tags. Cependant la rareté de telles ressources, et la limite de leur couverture d'un domaine repousse toujours plus loin le problème. L'expertise des utilisateurs d'un système semble en définitive la plus adaptée, mais aussi la plus complexe à exploiter si on cherche autant que possible, afin d'éviter toute surcharge cognitive, à limiter l'effort de contribution nécessaire à la formalisation de cette expertise.

2.3 Réification des assertions sémantiques sur les tags

L'objectif de notre modèle est de permettre la description des relations sémantiques qui peuvent exister entre des tags, tout en prenant en compte le caractère discutable des assertions portant sur ces relations sémantiques, et ceci autant lorsqu'elles sont le fruit d'un processus automatique que de l'action d'un utilisateur. Ainsi, chaque proposition, validation, ou invalidation de relation sémantique devient un événement dont le système garde une trace. A cette fin nous proposons un schéma RDF/s qui décrit les notions d'assertions et de relations sémantiques en spécifiant les liens qu'elles entretiennent avec d'autres notions issues notamment du modèle RDF de réification des assertions [2].

Dans notre modèle (voir figure 1), une assertion portant sur la relation sémantique entre deux tags d'une folksonomie est représentée par une classe RDF/s (`TagSemantic-Statement`) reliée (par la propriété `hasSemanticRelation`) à une autre classe décrivant la relation sémantique en question (`SemanticRelation` et ses sous-types). De plus, un utilisateur (`sioc:User` [3], qui peut être aussi un agent automatique) agit sur une assertion sémantique qu'il peut avoir proposée (`hasProposed`), approuvée (`hasApproved`) ou rejetée (`hasRejected`) ; une assertion sémantique hérite des propriétés de la classe `rdf:Statement` et a donc un sujet (`tag_subject` sous-type de `rdf:subject`) et un objet (`tag_object` sous-type de `rdf:object`). La notion de relation sémantique permet de spécifier, a minima, qu'il existe une relation sémantique entre deux tags, relation qui est spécifiée par ses sous-types dont les significations sont inspirées des propriétés de l'ontologie SKOS [4] : "plus particulier" (`Narrower`) ; "plus général" (`Broader`) ; "sémantiquement relié" (`Related`), qui peut être précisée par différents types de mesures de similarités ; et enfin "variations

2. voir http ://www.w3.org/TR/rdf-mt/#Reif
3. voir http ://rdfs.org/sioc/spec/
4. Simple Knowledge Organisation System, http ://www.w3.org/2004/02/skos/

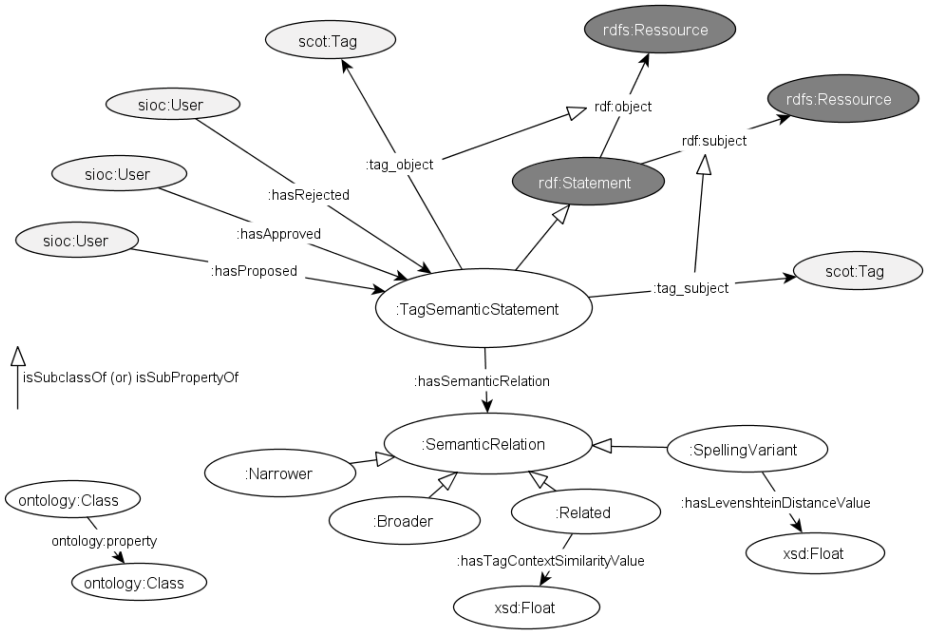


FIGURE 1 – Réification de la notion de relation sémantique

orthographiques" (`SpellingVariant`), qui peut par exemple être spécifiée par la distance de Levenshtein entre les deux tags.

Notre modèle inclue des ontologies déjà existantes comme SIOC (Bojars *et al.*, 2008), ou SCOT (Kim *et al.*, 2007). Le modèle SIOC permet de dissocier la notion de personne de celle d'utilisateur, et finalement d'usage propre, car une personne peut avoir autant d'instances de `sioc:User` que de comptes dans divers services. De plus, le rattachement de chaque instance `sioc:User` à un profil FOAF (Brickley & Miller, 2004) permet à toute personne de réunir et de gérer en un seul point d'accès toute son activité en ligne. L'ontologie SCOT réifie également la notion de tagging (en se rattachant à la Tag-Ontology de Newman *et al.* (2005)) en liant tout évènement de tagging à un compte `foaf:OnlineAccount`, lui même caractérisé par un type précis de plateforme web (blog, forum, etc.). Ainsi, la formalisation de l'acte de tagging nous semble conserver la nature duale du tag, soulignée par Monnin (2009) [5], qui est à la fois chaîne de charactères interprétable (ici capturé comme label de l'instance de la classe `scot:Tag`), mais aussi "étiquette matérielle" associée à la ressource taguée dans le cadre d'un usage précis et fortement conditionné.

5. que nous rejoignons par ailleurs à propos de l'accent qui doit être mis sur le "design informationnel" des plateformes d'échanges de connaissances, en nécessaire complément aux analyses plus spécifiquement sémantiques.

tag1	tag2	Distance de Levenshtein
informatique	information	0.75
geographie	geographique	0.83
déchets	déchet	0.85
industrie	industriel	0.9
développementdurable	développement-durable	0.95

TABLE 1 – Distance de Levenshtein pour certains couples de tags

3 Implantation et résultats

3.1 Détecter les variations orthographiques

En suivant l'exemple de Specia & Motta (2007), nous avons utilisé la méthode de Levenshtein [6] pour mesurer la distance d'édition entre deux chaînes de caractères, ceci dans le but de détecter les variations orthographiques de tags supposés équivalents. Le tableau 1 montre les valeurs de la distance de Levenhstein pour une série de tags extraits d'un échantillon des bookmarks d'utilisateurs de delicious.com ayant utilisé au moins deux fois le tag "ademe" [7] (ou ses variantes orthographiques). A la lecture de ce tableau nous voyons qu'il est délicat de trouver une valeur permettant de dire dans tous les cas que deux tags sont équivalents. Une manière de remédier à ces limitations serait d'employer un dictionnaire, ainsi que certaines règles heuristiques pour valider l'équivalence de deux tags dont la mesure de distance d'édition pour passer de l'un à l'autre passe un certain seuil, ou encore de combiner différentes mesures d'éditions. L'idée étant, dans le cadre de cet article, d'illustrer nos idées relatives à la validation de traitements automatiques par les utilisateurs, et dans l'attente du développement de ces améliorations, nous avons choisis pour notre implantation une valeur seuil de la distance de Levenshtein entre deux tags égale à 0,83.

3.2 Détecter les tags "thématiquement proches"

Nous proposons dans cette partie une méthode qui permet de suggérer des tags "thématiquement proches". Nous nous appuyons dans ce sens sur les résultats de l'étude de Cattuto *et al.* (2008) qui suggèrent dans ce cas l'utilisation d'une mesure distributionnelle de similarité basée sur le contexte tag-tag, par contraste avec les autres mesures distributionnelles ou les mesures basées sur la simple cooccurrence qui ont tendance à refléter des liens de types hiérarchiques. Cette mesure consiste tout d'abord, pour deux tags t_1 et t_2, à calculer leurs vecteurs associés v_1 et v_2, où v_{ik} correspond à la valeur de cooccurrence des tags t_i et t_k qui est augmentée d'une unité à chaque fois que les tags t_i et t_k sont employés pour le même bookmark. La mesure de similarité entre t_1

6. telle qu'implantée par http ://www.dcs.shef.ac.uk/˜sam/simmetrics.html

7. "ademe" correspond à l'anagramme de Agence De l'Environnement et de la Maîtrise de l'Energie. Notre échantillon se compose des 6054 bookmarks postés par 16 utilisateurs, ayant associé, globalement, 5153 tags distincts à 5969 URL distinctes.

voiture	auto (0.81), automobile (0.83), co2 (0.85), pollution (0.83)
développement	durable (0.88), ecologie (0.8)
construction	habitat (0.95), isolation (0.92), pdf (0.77)
solaire	photovoltaïque (0.74)
réglementation	logement (0.79), thermique (0.82)

TABLE 2 – Pour un tag donné, tags ayant une valeur de similarité dans le contexte tag-tag supérieure à 0.7

et t_2 correspond quant à elle au cosinus de l'angle entre les vecteurs v_1 et v_2, soit : $\cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|_2.\|v_2\|_2}$.

Le tableau 2 nous montre une sélection de tags ayant une valeur de similarité (contexte tag-tag) supérieure à $0,7$. Pour ces mesures, nous avons prélevé une partie de notre jeu de données en ne conservant que les tags associés aux 100 bookmarks ayant été tagués avec le tag "ademe" (ou une de ses variantes orthographiques) [8] . Nous pouvons observer que les liens inférés reflètent bien les relations thématiques liées au domaine de l'environnement, hormis pour le tag "pdf" qui est associé au tag "construction". Ceci peut s'expliquer par le fait que le tag "pdf" ait pu être associé souvent aux autres tags liés au tag "construction", simplement car les documents taggués était au format .pdf.

3.3 Intégration dans un système de gestions de *bookmarks*

Le système que nous proposons pour illustrer notre propos est un système de navigation au sein d'une base de *bookmarks* extraits de delicious.com (le jeu de données utilisé ici est le même que celui décrit à la section 3.1). Dans notre modèle, nous avons formalisé la notion de bookmark à l'aide de la classe `Bookmark` qui est une sous-classe de la classe `sioc:Item`, faisant du bookmark un document au même titre qu'un billet de blog (`sioc:Post`). De plus, les propriétés `scot:tagOf` et `scot:hasTag` relient le bookmark à un ou plusieurs tags, et la propriété `sioc:about` le relie à une ressource (en tant que `rdf:Ressource`), ce qui rend compte de l'indexation faite *via* les bookmarks et permet de retrouver les tags associés à une ressource à l'aide d'une simple requête SPARQL [9].

Notre système s'appuie également sur notre modèle de réification des relations sémantiques et se compose d'agents automatiques effectuant en tâche de fond des traitements sur les folksonomies, et d'une interface d'exploration de la base de bookmarks. L'imprécision et le caractère discutable des traitements automatiques décrits ci-dessus (cf. sections 3.1 et 3.2) rendent délicate leur application systématique. Nous proposons donc de donner la possibilité aux utilisateurs du système de contribuer à la validation ou l'invalidation des relations sémantiques automatiquement suggérées entre les tags qu'ils manipulent. La figure 2 montre, lors de la recherche de bookmarks par tag, un exemple de fonctionnalité sémantique suggérée par l'interface qui propose une liste

8. cet extrait du jeu de données se compose donc des 100 bookmarks des 75 utilisateurs ayant associé, globalement, 221 tags distincts à 107 URL distinctes.

9. SPARQL Query Language for RDF : http ://www.w3.org/TR/rdf-sparql-query/

de ressources associées à un tag et ses variations orthographiques (*spelling variant*) calculés grâce à la distance de Levenshtein (la valeur seuil utilisée est 0,83). La fonctionnalité suggérée en question consiste à retirer un des termes de la liste des termes équivalents en cliquant sur la croix rouge encerclée située à côté de chaque terme.

Le recours à la fonctionnalité sémantique proposée reste optionnelle, car l'utilisateur est libre d'interpréter les résultats et peut tout à fait conserver les termes suggérés ("industrial" et "industriel" pour "industrie"). Si toutefois l'utilisateur désire retirer l'un de ces termes, notre modèle laisse le choix aux concepteurs du système d'appliquer cette assertion sémantique (le tag "industrial" n'est pas équivalent à "industrie") pour toutes les requêtes futures d'autres utilisateurs, ou d'un certain groupe d'utilisateurs seulement, ou simplement de ce même utilisateur (comportement que ce dernier attendra certainement en toute logique).

Cet exemple montre la capacité de notre modèle à supporter les actions contradictoires ou conflictuelles des utilisateurs. En effet, lorsqu'un utilisateur choisit de retirer le terme "industrial", le système générera une annotation rendant compte de cette action. Cette annotation s'ajoutera à celle rendant compte du lien d'équivalence entre "industrie" et "industrial", sans l'annuler pour autant. Ainsi notre modèle permet que l'action d'un utilisateur n'annule pas systématiquement celle d'un autre lorsque ces deux actions sont conflictuelles. La décision finale revient aux concepteurs du système qui peuvent choisir différentes "politiques" d'applications des actions de validation par les utilisateurs. Plusieurs solutions sont en effet possibles pour gérer les situations conflictuelles : il est possible de (1) rendre visible ces divergences en les organisant en points de vue qui sont explicitement montrés à l'utilisateur, ou (2) d'appliquer ces divergences différemment selon l'appartenance des utilisateurs à des sous-groupes d'intérêt identifiés par ailleurs au sein de l'organisation considérée, ou encore (3) de proposer un système de vote au sein des groupes d'utilisateurs pour sélectionner l'assertion sémantique à conserver.

4 Positionnement et discussion

Dans le cadre de la recherche sur le Web social et sémantique, plusieurs applications concrètes ont implanté des fonctionnalités sémantiques pour organiser des contenus partagés. Les concepteurs de Revyu.com (Heath & Motta, 2007) proposent d'exploiter les formalismes du Web sémantique afin de faciliter l'interopérabilité entre les plateformes de partages de contenus et d'éviter les redondances inutiles. Passant & Laublet (2008) proposent un modèle (MOAT) et des outils qui permettent d'associer les différents sens d'un tag à des documents contenant la définition visée, ou à des concepts d'ontologies du Web Sémantique. CartoDD (Cahier *et al.*, 2007) est l'exemple d'un autre type d'approche basée sur les formalismes du Web Socio-Sémantique (Zacklad *et al.*, 2007), et qui propose d'effectuer une cartographie de contenus à l'aide de cartes de thèmes multi points de vue construites collaborativement par les utilisateurs.

Notre approche quant à elle se démarque de celle de Passant & Laublet (2008) en décrivant, dans un premier temps, le sens des tags grâce à des relations sémantiques entre les tags ("plus général" ou "plus particulier") sans pour autant s'interdire, par la suite et de manière indépendante, de relier les concepts ainsi qualifiés à des concepts

FIGURE 2 – Exemple de fonctionnalité sémantique suggérée par l'interface pour retirer un tag non-équivalent

d'ontologies plus formelles lorsque cela est pertinent pour nos usagers. Notre but est de construire en premier lieu des ontologies légères qui s'apparenteraient aux thésaurus tels que modélisés par le schéma SKOS, et qu'il est toujours possible ensuite de rapprocher d'autres ontologies, formelles ou non, soit en adaptant des techniques d'alignement (Euzenat & Shvaiko, 2007), soit en les mettant en perspectives sous des points de vues différents et explicités dans l'interface, à la manière de (Cahier *et al.*, 2007) qui ont intégré le thésaurus GEMET [10] comme un des points de vue de l'ontologie sémiotique mise en œuvre dans le système CartoDD [11].

Nous cherchons donc à augmenter les systèmes invitant les usagers à contribuer directement à l'élaboration de vocabulaires partagés en insérant des fonctionnalités d'organisation sémantique dans les interfaces de recherche et de navigation. Ces fonctionnalités consistent à permettre aux utilisateurs de valider ou corriger des suggestions automatiques de termes pertinents pour une recherche d'informations. Elles doivent également rester les moins intrusives possibles afin de ne pas perturber les autres tâches des usagers. Les solutions en cours d'élaboration présentées dans cet article peuvent être vues comme des fonctionnalités complémentaires d'autres outils collaboratifs explicitement dédiés à l'enrichissement sémantique de folksonomies comme par exemple celui développé dans SweetWiki (Buffa *et al.*, 2008).

Notre modèle cherche également à prendre en compte les différents points de vue, à la manière de Cahier *et al.* (2007). En effet, en réifiant la notion de relation sémantique, notre modèle permet de faire de chaque assertion portant sur la sémantique des tags un événement au même titre que le tagging. Même les relations sémantiques entre les tags qui seraient contradictoire (le tag "co2" est plus précis que le tag "polluant" mais également plus précis que le tag "ressource-photosynthèse" par exemple) peuvent être

10. http ://www.eionet.europa.eu/gemet/index_html ?langcode=fr
11. http ://tech-web-n2.utt.fr/dd/ ?mod=navigation

recueillies et permettre ainsi de mettre en avant les différents points de vues portant sur une même notion, et de répercuter ces distinctions dans l'ontologie.

L'enrichissement sémantique des folksonomies a également été abordé par Mika (2005) qui propose d'analyser la structure de graphe liée aux folksonomies (*via* les associations ressources/tag/utilisateurs) pour en déduire des liens sémantiques entre les tags. D'autres approches établissent des correspondances entre les tags et des éléments d'ontologies disponibles en ligne sur le Web Sémantique, et étendent les requêtes effectuées sur une folksonomie avec ces éléments d'ontologies (Angeletou *et al.*, 2008). Si nous exploitons le même type de traitements sur les folksonomies que Mika et Angeletou *et al.*, nous cherchons cependant à tirer partie de l'expertise des utilisateurs (à la manière de Tanasescu & Streibel (2007) qui proposent de taguer les tags, où comme Braun *et al.* (2007) qui proposent d'intégrer les approches collaboratives du Web 2.0 dans les processus d'élaboration d'ontologies), ceci afin d'améliorer l'adéquation de ces traitements avec les usages. Par ailleurs, notre approche peut être mise en regard avec celles visant à construire des ontologies de domaines à partir d'une extraction terminologique menée au sein d'un corpus de documents (Aussenac-Gilles *et al.*, 2000). D'un point de vue méthodologique, les tags que nous cherchons à lier sémantiquement à d'autres tags peuvent être vus comme des "candidats-tags" (par analogie aux "candidats-termes"), c'est à dire des syntagmes soumis à la validation d'experts du domaine, avant de devenir éventuellement des concepts ou des relations sémantiques d'une ontologie.

5 Conclusion

Notre approche consiste à intégrer les données folksonomiques dans un processus de construction collaborative de représentations des connaissances, et ceci dans le but de fournir des services et des fonctionnalités plus avancées aux systèmes à base de folksonomies. Nous proposons à cet égard d'exploiter des traitements automatiques tout en permettant aux utilisateurs de les valider ou de les invalider. Les deux types de fonctionnalités sémantiques que nous proposons dans cet article sont la reconnaissance de variations orthographiques des tags équivalents et la recherche de tags thématiquement proches [12]. Afin de valider ces inférences automatiques, nous avons montré un exemple de fonctionnalité suggérée par l'interface invitant l'utilisateur à retirer un tag de la liste des tags automatiquement inclus dans la recherche.

Nous avons proposé également un modèle de formalisation des traitements automatiques et des actions de validations par les utilisateurs qui supporte les situations conflictuelles. Nous suggérons ainsi de capturer les assertions, éventuellement divergentes, portant sur la sémantique des tags (et résultantes des traitements automatiques ou de l'action d'utilisateurs), puis de repousser leur traitement au moment de l'exploitation de ces résultats, en fonction des choix des concepteurs du système. Ces derniers peuvent ainsi choisir de montrer les résultats des traitements sémantiques en fonction de l'appartenance à un sous-groupe d'usagers, ou bien en fonction d'une valeur seuil de l'occurrence d'une assertion.

12. La première fonctionnalité a été implantée dans notre système de gestion de bookmarks partagés, et la seconde, permettant de suggérer des tags thématiquement proches ("related"), est en cours d'implantation

Nos futurs travaux incluent, outre les tests de terrain avec l'Ademe, la recherche d'autres types de traitements des tags permettant de proposer d'autres fonctionnalités comme la caractérisation plus fine des relations sémantiques entre les tags (relation "plus précis" ou "plus particulier"). D'autres fonctionnalités présentent un intérêt dans la perspective d'outiller l'organisation collaborative des folksonomies, comme la reconnaissance de divergence ou de convergence entre les utilisateurs pour la catégorisation de ressources similaires. Une des applications serait l'assistance à la constitution de groupes d'intérêts qui pourraient, dans le cadre de notre modèle, être utilisés pour personnaliser les interfaces en fonction des annotations sémantiques recueillies lors de l'usage du système. A cet égard, de multiples modalités d'applications des connaissances formalisées ainsi recueillies sont possibles et feront l'objet de recherches futures.

Notre étude s'inscrit également dans la recherche de méthodes pour outiller de manière "dynamique" l'élaboration d'ontologies légères et partagées. A cette fin, nous cherchons à développer des outils permettant aux usagers de saisir au cours de leurs tâches quotidiennes la dimension partagée de leur usage de certains termes. Dans un premier temps, nous avons tenté d'intégrer des fonctionnalités sémantiques aux tâches de recherche d'informations, et dans le cas de notre illustration, plus particulièrement à la recherche au sein d'une base de bookmarks tagués. Nous souhaitons donc étendre notre recherche à l'analyse des usages et des tâches effectuées par les membres d'une communauté ou d'un réseau, dans le but d'identifier d'autres tâches susceptibles d'être autant d'occasions pour l'organisation des connaissances partagées. Une meilleure connaissance des usages permettra également d'accroître les possibilités de personnalisation et de manipulation des résultats donnés par le système, ainsi que son utilisabilité *via* une plus grande transparence des raisonnements appliqués pour obtenir un résultat.

Remerciements. Nous remercions l'ANR pour le financement du projet ISICIL ANR-08-CORD-011 qui a permis la production de ces résultats.

Références

ANGELETOU S., SABOU M. & MOTTA E. (2008). Semantically enriching folksonomies with flor. In *CISWeb Workshop at Europ. Semantic Web Conf.*

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Corpus analysis for conceptual modelling. In *EKAW - Workshop on Ontologies and Texts*.

BOJARS U., PASSANT A., CYGANIAK R. & BRESLIN J. (2008). Weaving SIOC into the Web of Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*.

BRAUN S., SCHMIDT A., WALTER A., NAGYPÁL G. & ZACHARIAS V. (2007). Ontology maturing : a collaborative web 2.0 approach to ontology engineering. In *CKC*, volume 273 of *CEUR Workshop Proceedings* : CEUR-WS.org.

BRICKLEY D. & MILLER L. (2004). *FOAF Vocabulary Specification*. Namespace Document 2 Sept 2004, FOAF Project. http ://xmlns.com/foaf/0.1/.

BUFFA M., GANDON F., ERETEO G., SANDER P. & FARON C. (2008). SweetWiki : A semantic Wiki. *J. Web Sem.*, **6**(1), 84–97.

CAHIER J.-P., ZAHER L. & ZACKLAD M. (2007). Information seeking in a "socio-semantic web" application. In *ICPW07 : Proceedings of the 2nd international conference on Pragmatic web*, p. 91–95, New York, NY, USA : ACM.

CATTUTO C., BENZ D., HOTHO A. & STUMME G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. *7th International Semantic Web Conference*.

EUZENAT J. & SHVAIKO P. (2007). *Ontology Matching*. Berlin, Heidelberg : Springer.

C. FELLBAUM, Ed. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London : The MIT Press.

GANDON F. (2008). *Graphes RDF et leur Manipulation pour la Gestion de Connaissances*. Habilitation Ã diriger des recherches, University of Nice - Sophia Antipolis.

HEATH T. & MOTTA E. (2007). Revyu.com : a Reviewing and Rating Site for the Web of Data. In *ISWC/ASWC*, volume 4825 of *LNCS*, p. 895–902 : Springer.

KIM H.-L., YANG S.-K., SONG S.-J., BRESLIN J. G. & KIM H.-G. (2007). Tag Mediated Society with SCOT Ontology. In *Semantic Web Challenge, ISWC*.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, **10**(8), 707–710.

LIMPENS F., GANDON F. & BUFFA M. (2008). Rapprocher les ontologies et les folksonomies : un Etat de l'art. *IC*.

MATHES A. (2004). *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Rapport interne, GSLIS, Univ. Illinois Urbana-Champaign.

MIKA P. (2005). Ontologies are Us : a Unified Model of Social Networks and Semantics. In *ISWC*, volume 3729 of *LNCS*, p. 522–536 : Springer.

MONNIN A. (2009). Qu'est ce qu'un tag ? entre accès et libellés, l'esquisse d'une caractérisation. In *Ingénierie des Connaissances, Hammamet, Tunisie*.

NEWMAN R., AYERS D. & RUSSELL S. (2005). Tag Ontology Design. http ://www.holygoat.co.uk/owl/redwood/0.1/tags/.

PASSANT A. (2007). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In *International Conference on Weblogs and Social Media*.

PASSANT A. & LAUBLET P. (2008). Meaning of a tag : A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*.

SPECIA L. & MOTTA E. (2007). Integrating folksonomies with the semantic web. *4th European Semantic Web Conference*.

TANASESCU V. & STREIBEL O. (2007). ExtremeTagging : Emergent Semantics through the Tagging of Tags. In *ESOE at ISWC*.

ZACKLAD M., BÉNEL A., CAHIER J., ZAHER L., LEJEUNE C. & ZHOU C. (2007). Hypertopic : une Métasémiotique et un Protocole pour le Web Socio-Sémantique. In *IC*, p. 217–228 : Cépaduès. ISBN 978-2-85428-790-9.

Une démarche de conception de services d'information et de communication dédiés aux communautés d'aidants

Matthieu Tixier, Myriam Lewkowicz

[1] Laboratoire Tech-CICO (ICD, FRE CNRS 2848– Université de Technologie de Troyes)
12 rue Marie Curie - BP 2060 - 10010 TROYES Cedex
{prénom.nom}@utt.fr

Résumé[1] : Avec l'adoption massive des Technologies de l'Information et de la Communication (TIC) et la dilution des liens sociaux que connait notre société, Internet s'est imposé comme un nouveau lieu de soutien social. Ainsi, on voit des personnes connaissant des situations difficiles comme la maladie s'apporter du soutien en termes de connaissances, de réconfort, voire d'aide tangible au travers de services en ligne comme les forums de discussions. Notre projet vise à concevoir une plate-forme dédiée au soutien social. Pour cela, nous proposons une démarche de conception interdisciplinaire basée sur l'idée de *traduction*. Cette démarche s'appuie sur l'observation de pratiques réelles de soutien social entre aidants familiaux de patients pris en charge par un réseau de santé. Notre plate-forme vise à développer une pratique de soutien social en ligne chez ces aidants en complément de la pratique en face à face et à évaluer l'impact de l'introduction de cette nouvelle pratique sur la communauté.

Mots-clés : soutien social, communauté d'aidants, scripts, conception interdisciplinaire.

1 Introduction

Le soutien social est généralement défini comme l'échange de messages verbaux et non-verbaux, qui transmettent des émotions ou de l'information, afin de réduire l'incertitude ou le stress d'une personne (Barnes & Duck, 1994). Il est traditionnellement apporté par les proches, amis, famille. Or, le mode de vie actuel tend à accroître l'éloignement géographique et la dilution des liens sociaux. Il y a donc naturellement la recherche d'alternatives pour obtenir du soutien social. Celui-ci peut être apporté par des experts (assistantes sociales, psychologues, …), des aidants professionnels, ou par des pairs (au sein de groupes de parole, d'espaces de discussion en ligne). Dans ce cas, l'idée est d'échanger des expériences, du vécu, et on observe la constitution de communautés basées sur le partage d'expérience. Ce phénomène est d'autant plus présent dans le cas de situations de détresse comme la maladie. On

[1] Ce travail de recherche est conduit avec le soutien du Conseil Général de l'Aube.

observe ainsi des patients ou des aidants de patients à la recherche de soutien informationnel, émotionnel ou tangible (Thoits, 1986).

Nous nous intéressons aux aidants non-professionnels de patients atteints de pathologies lourdes. Rien ne prépare les aidants à supporter ce rôle lorsque des circonstances dramatiques viennent frapper leurs proches. Au-delà des informations médicales complexes à intégrer, ils sont démunis tant en terme d'information disponible, de savoir-faire pour interagir avec le malade, qu'en terme d'écoute des difficultés qu'ils éprouvent dans leur rôle d'aidant. Et ce d'autant plus qu'ils sont accaparés par l'accompagnement de leur proche malade et n'ont que très peu de temps à consacrer à cette recherche d'information, à une prise de recul sur leurs pratiques, et plus généralement aux interactions sociales les éloignant de leur proche malade. Un outil informatique disponible 24/24 et 7/7 à domicile trouve ici toute son utilité.

Notre projet de recherche se situe à deux niveaux : d'une part nous souhaitons concevoir un outil pour le soutien social au sein d'une communauté d'aidants que nous avons observée, et d'autre part nous menons une réflexion sur la démarche de conception en elle-même. Le soutien social est une activité multidimensionnelle qui pose tant la question des connaissances que les aidants développent dans l'accompagnement du malade au quotidien et des difficultés auxquelles ils font face au quotidien, que celle de l'émotion ou du social (association, réseaux de santé, amis, famille). A ce titre, nous nous inscrivons dans une démarche d'ingénierie pluridisciplinaire afin de concevoir un outil efficace qui prenne en compte ces nombreuses dimensions, dont celle des connaissances, qui est centrale (traitements médicaux, évolution de la maladie, interaction avec le patient). C'est d'autant plus le cas dans le cadre du soutien social en ligne, où plusieurs auteurs comme (Gustafson *et al.* 2002, Xie 2008) mentionnent la place centrale de la recherche d'information et de contenus de qualité, en d'autres termes de l'acquisition de connaissances par les utilisateurs, au delà du besoin de réconfort. Les informations et connaissances se trouvent la plupart du temps mêlées dans un même message avec les marques de réconfort (Gaglio & Atifi 2008) ; la seule compassion ne suffit pas à apporter un soutien efficace.

Nous présentons dans cet article une approche que nous qualifions de *traduction*, qui participe à une démarche d'ingénierie pour les communautés, car elle vise à expliciter le processus de passage entre des besoins exprimés ou observés, et la définition de fonctionnalités du système permettant d'y répondre. Nous tentons en cela d'ouvrir la « boîte noire » de l'intuition du concepteur. L'originalité de notre approche tient essentiellement en deux points : (1) elle est située car elle est ancrée dans l'observation de pratiques existantes, et (2) elle propose de traduire des descriptions de situations prototypiques en fonctionnalités d'une plate forme de soutien social.

Afin de mener à bien ce projet, nous travaillons avec un réseau de santé focalisé sur les troubles de la mémoire, et en particulier la maladie d'Alzheimer. Ce réseau tente de mettre en place des communautés d'aidants pour développer des espaces de rencontre et ainsi favoriser les contacts entre ces aidants. Nous adoptons ici une perspective interactionniste au sens où ce sont les interactions régulières entre membres qui fondent la communauté au delà des caractères ou traits sociodémographiques qu'ils partagent.

Cet article est organisé comme suit : après une description du terrain que nous avons étudié, nous présentons les résultats de notre analyse de plates-formes existantes dédiées au soutien social en ligne. Nous présentons alors notre démarche de conception, illustrée par des exemples de traduction en fonctionnalités de situations de communication au sein des communautés observées. Nous concluons en indiquant les prochaines étapes de notre projet de recherche.

2 Communauté d'aidants familiaux et réseau de santé

Le réseau pôle mémoire (RPM) est un réseau de santé concernant les troubles de la mémoire et notamment la maladie d'Alzheimer. Le réseau a à la fois une mission de coordination des professionnels des domaines médicaux, psychologique et social qui interviennent dans la prise en charge des pathologies complexes des patients, et une mission de production de soin de par ses actions en terme de diagnostic neuropsychologique (bilan mémoire) et d'accompagnement et suivi psychologique des patients et de leurs accompagnants. Le RPM propose également des groupes de paroles mensuels aux aidants conjoints et enfants de patients afin que ceux-ci puissent partager leur situation entre pairs. Les formations sur la maladie, dispensées par les différents professionnels de santé (orthophoniste, neurologue, kinésithérapeute, etc.), psychologues et assistants sociaux intervenant dans la prise en charge du patient, organisées par le RPM sont également une bonne occasion pour les aidants de rencontrer et d'échanger sur leur vécu.

Ces aidants sont très investis dans l'accompagnement de leur proche malade, bien que la situation varie selon qu'ils soient enfant ou conjoint du patient, l'aide qu'ils apportent prend une part importante sur leur propre vie et laisse peu d'espace pour s'évader de la maladie, voire même simplement pour s'occuper d'eux mêmes. Parmi les facteurs de saturation identifiés des aidants de patients atteints de la maladie d'Alzheimer, on note le fait d'avoir à s'occuper du patient plus de 2 heures par jour, le manque de soutien à domicile (aide ménagère…), d'avoir assumé le rôle d'aidant depuis plus de 50 mois et d'avoir dû renoncer à un soin (consultation ou hospitalisation) par manque de temps pour soi (ce qui est le cas d'un aidant sur cinq). (PIXEL 2000).

Plus que le besoin de soutien financier, qui est pour l'heure la réponse essentielle des institutions, les aidants expriment un besoin de soutien moral (IFOP 2008). Ils ressentent un manque d'écoute à leur problème auquel les professionnels de santé ne sont pas forcément en mesure de répondre. Ainsi, pouvoir échanger avec d'autres aidants traversant ou ayant traversés les mêmes difficultés semble un bon moyen de combler ce besoin d'information et de réconfort. Les groupes de paroles sont des dispositifs qui permettent à des personnes connaissant des situations difficiles de se rencontrer et de partager leur expérience. Cependant, cette solution reste assez contraignante tant au niveau des horaires, des lieux de rencontres que des freins psychologiques existants liés au fait de s'exprimer en public (Salem *et al* 1997.). Ainsi, Internet, en abolissant les contraintes temporelle et géographique, est devenu au fil des années également un lieu de rencontre pour les patients et aidants de patients

souhaitant partager leurs expériences et leurs connaissances comme le montre les succès rencontrés par des sites comme Doctissimo (plus de 7 millions de visiteurs uniques par mois, Nielsen/NetRatings, novembre 2008).

Le RPM en tant que structure organisée est à l'initiative de la création de deux groupes d'aidants, correspondant aux deux groupes de paroles que le réseau organise. L'un est destiné aux aidants enfants des patients et l'autre aux conjoints. Chaque groupe comprend une vingtaine de personnes qui participent plus ou moins régulièrement aux groupes. Les groupes de paroles sont organisés le premier vendredi de chaque mois, de 15 à 17h pour les aidants de conjoint, et de 19 à 21h pour les aidants enfants. Cette séparation répond à la perspective thérapeutique choisie par la psychologue coordinatrice du réseau, et généralement admise (Pillemer *et al.* 2002), les deux situations d'accompagnement étant différentes (partage du domicile, âge, activité professionnelle, relation au malade).

La « communauté des aidants » du RPM, actuellement séparée entre ces deux groupes de paroles, est encore en pleine émergence. Elle n'existe pas indépendamment de la structure institutionnelle qui est à l'initiative du cadre de leurs interactions. Nous pensons qu'étendre la pratique de soutien social existant chez les aidants au travers d'un outil en ligne permettra de développer le capital social de la « communauté » afin qu'en retour une communauté plus dense et active vienne animer cette pratique indépendamment de l'institution.

3 Solutions existantes dédiées au soutien social en ligne

Le soutien social ouvre un champ de recherche particulièrement intéressant sur les communautés en ligne et les conditions de ces pratiques de soutien social en ligne. Plusieurs travaux essentiellement nord-américains ont été conduits sur le sujet, comme ceux de Maloney-Krichnar et Preece (2005). Toutefois, peu se sont intéressés à développer des services de communication et d'information spécifiques et innovants pour instrumenter cette activité complexe (Lewkowicz *et al.*, 2008). Ainsi, nous avons cherché à constituer un panel représentatif afin d'avoir une vue globale sur les solutions existantes, d'une part au niveau de l'organisation des contenus sur ces plateformes, et d'autres part quant aux fonctionnalités proposées aux utilisateurs.

La notion de fonctionnalité en tant que catégorie d'analyse pose un certain nombre de problèmes qui appellent à apporter quelques précisions. D'une part, une fonctionnalité a un caractère intentionnel (Kroes, 2005) en ce qu'elle sert une finalité de l'utilisateur (i.e. envoyer un message). D'autre part, une fonctionnalité peut être elle-même un élément d'une fonctionnalité plus complexe. Par exemple, la fonctionnalité « rédiger un commentaire » est un composant identifié d'un blog, cependant qu'un blog peut lui même apparaître comme un composant fonctionnel d'un CMS (Content Management System). Par ailleurs, la fonctionnalité de commentaire n'est pas indissociable du blog puisqu'elle peut également être trouvée dans de nombreux autres outils comme une plate-forme de réservation de voyages en ligne ou un site de réseau social. On peut même pousser le raisonnement plus loin en s'intéressant seulement au champ permettant de rédiger le commentaire, qui est en soi

un composant fonctionnel utilisé dans la plupart des applications disposant d'une interface graphique. La notion de fonctionnalité pose donc également la question de la granularité de l'analyse.

Le point de vue que nous avons adopté dans notre analyse des plates-formes est celui du concepteur, dans le sens où l'on se restreindra ici à envisager les fonctionnalités dans leur caractère conventionnel et leur usage suggéré par les concepteurs des dispositifs, point de vue que nous estimons être en mesure de comprendre en tant que chercheurs en informatique et utilisateurs expérimentés des TIC. Afin de répondre à la question de la granularité de l'analyse, nous nous sommes attachés à utiliser systématiquement les mêmes items pour les fonctionnalités présentes sur plusieurs plateformes, la variabilité des items présents sur ces systèmes étant finalement la base de nos résultats.

Treize sites ont été retenus en essayant de donner une place égale aux sites utilisant des solutions classiques de type forum de discussion (5), aux outils de la génération Web2.0 tendant à intégrer des dimensions des plateformes de réseaux sociaux (5) et aux outils issues du monde de la recherche : HutchWorld (Cheng *et al.* 2000), CHESS (Gustafson *et al.* 2002), et Krebsgmeinschaft (Leimeister *et al.* 2005). Chaque fois qu'il a été possible (certains projets de recherche n'ayant pu être accédés qu'au travers des publications dont ils ont fait l'objet), un compte utilisateur a été créé sur la plate-forme afin d'accéder à l'ensemble des fonctionnalités et à la présentation proposée aux utilisateurs membres. Le recueil des données a été conduit au cours du deuxième semestre 2008 et celles-ci sont à jour du mois de septembre, certaines plates-formes connaissant des changements et évolutions rapides.

Nous avons conduit cette étude suivant une grille d'analyse en deux axes afin d'assurer une collecte systématique des items pour chaque site. Notre grille s'intéresse d'une part à l'organisation des contenus et des fonctionnalités sur la plate-forme, et d'autre part aux dispositifs de support, direct ou indirect, à la communication et aux interactions. Nous précisons nos axes dans ce qui suit :

Organisation : s'intéresse à la structuration du site, aux différents espaces accessibles à l'utilisateur. Ces espaces sont clairement visibles en tant que tels sur l'interface de l'application, toutefois il ne s'agit pas pour nous de recopier l'arborescence du site mais plutôt de mettre en évidence les traits de structure et les lieux pertinents pour notre analyse. Cette catégorie d'analyse s'intéresse autant que possible aux dimensions logique, temporelle (navigation, parcours utilisateur suscité) et spatiale.

Dispositifs de communication et d'interactions : décrit les dispositifs d'échanges entres les utilisateurs. Nous nous intéressons notamment au caractère multidimensionnel des dispositifs de communication médiatisée (Xie, 2008) au travers des dimensions synchrones/asynchrones, public/privée, adressée ou non.

Chacun des sites a été étudié en suivant cette grille et nous avons ensuite cherché à rapprocher les items identifiés afin de mettre en évidence des composants fonctionnels communs à plusieurs sites, ainsi que des modes généraux d'organisation des contenus et fonctionnalités. Les données complètes de l'étude peuvent être consultées sur http://www.orkidees.com/missWiki

Nous présentons ici les principaux résultats de notre analyse.

En ce qui concerne l'organisation des contenus et fonctionnalités des sites, cinq types d'organisation sont mis en évidence. Plusieurs types d'organisation peuvent être proposés sur un même site.

Thésaurus : Les contenus sont organisés de façon systématique autour d'un ou plusieurs thésaurus de mot-clés comme des listes de noms de pathologie ou de traitements.[2]

Thématique : Des thématiques générales organisent les contenus sans forcément chercher à être systématique ou exhaustif à la manière de sujets généraux de discussion : les allergies à l'école par exemple (Forums).

Temporelle/Narrative : Le site s'appuie sur un découpage narratif de la situation d'intérêt. Par exemple les contenus sont organisés selon les différentes étapes prototypiques d'un divorce sur Divorce360[3] : deciding, beginning, process, ongoing, moving on.

Contenus « profanes » / « certifiés » : Une séparation est marquée entres les contenus et fonctionnalités donnant accès à des contenus certifiés par une autorité (médicales par exemple) et les contenus produits par les utilisateurs. (Krebsgemeinschaft, PatientLikeMe)

Spatiale : L'organisation de la plate-forme repose sur une métaphore spatiale en référence à des lieux plus ou moins réel (Hutchworld).

Le second axe de notre étude visait à identifier les fonctionnalités existantes dans les plates-formes de soutien social en ligne considérées. Nous avons distingué 27 composants fonctionnels. Comme nous l'avons mentionné plus tôt, certaines fonctionnalités peuvent être présentes dans des composants qui desservent un but plus spécifique. Certains composants peuvent être qualifiés d'« élémentaires » et d'autres de « complexes », au sens où ils se présentent comme une combinaison de fonctionnalités élémentaires.

Les innovations que nous avons trouvées dans les plateformes issues du monde de la recherche ou de la génération du web2.0 ne relèvent pas vraiment de l'intégration de dispositifs techniques inédits. L'utilisation de la 3D ou de contenus vidéo par exemple reste marginale dans nos observations et ce sont toujours les classiques champs textes et cases à cocher qui sont utilisés. Cela fait plus d'une décennie que nous pouvons proposer des contenus graphiques de qualité et des technologies emblématiques comme l'AJAX apportent essentiellement à la fluidité de transfert des requêtes et de rafraîchissement des contenus. Les innovations qui nous ont interpellés sont d'un autre ordre. En effet, nous identifions une tendance générale où il s'agit, en combinant plusieurs fonctionnalités, de proposer des situations de communication plus spécifiques que celles que l'on trouve dans des dispositifs traditionnels comme les forums de discussions. Ainsi, on n'échange plus seulement des messages mais on *pose et répond à des questions*, on échange des *conseils* ou on *partage des expériences*. Notre analyse nous montre que des concepteurs ont cherché à prendre comme point de départ des situations concrètes et existant dans le vécu de tout un chacun comme le partage d'expérience, afin d'en donner une traduction, en combinant

[2] Parmi les sites étudiés : IMedix (http://www.imedix.com), DailyStrength (www.dailystrength.com), PatientsLikeMe (http://www.patientslikeme.com)

[3] Divorce360 (http://www.divorce360.com)

des fonctionnalités générales comme un dispositif de communication asynchrone, et la possibilité de réagir à l'aide d'un système de commentaires. C'est à la question de cette *traduction* que nous nous intéressons afin d'en dégager des principes pour la conception.

4 Objectifs

La mise à disposition d'un outil de communication sur Internet disponible 24 heures sur 24 et 7 jours sur 7 apparaît comme particulièrement pertinente pour les aidants familiaux. En effet, les aidants ne disposent pas ou de très peu de temps à consacrer à d'autres activités que la prise en charge du patient. Un tel dispositif permettrait d'atteindre des aidants ne bénéficiant pas des groupes de parole, soit parce qu'ils n'en ont pas la possibilité, soit parce que ce dispositif ne leur convient pas (Salem *et al.*, 1997). Par ailleurs, il constituerait un prolongement pour les aidants qui participent déjà aux groupes de parole. Notre système sera ainsi un complément aux dispositifs de soutien existants et non un substitut. Cela permettrait également aux aidants de ne pas laisser s'accumuler des questions ou des soucis en attendant le prochain groupe de parole ou la prochaine visite auprès d'un professionnel de santé.

Ces avantages sont perçus par les aidants que nous avons rencontrés, qui expriment dans leur grande majorité leur souhait de disposer de tels moyens de communication. Cependant, comme nous le montrent des études sociodémographiques et notre enquête sur le terrain, la plupart des aidants familiaux ont des compétences faibles en informatique et une habitude peu fréquente de l'usage d'internet ; seuls 7% d'entre eux déclarent utiliser Internet pour les aider dans leur rôle d'aidant alors que 60% y ont accès (IFOP, 2008).

Notre objectif est donc de concevoir un outil qui serait le plus intuitif possible pour ces aidants. Notre hypothèse est que cet outil sera le plus intuitif possible s'il est conçu en gardant à l'esprit les pratiques de soutien social. Notre proposition consiste donc (1) à intégrer les utilisateurs dans la conception de l'outil, mais non pas en les interviewant sur leurs attentes, mais en analysant leurs pratiques actuelles en face à face et, (2) à rester le plus fidèles possible à ces pratiques en proposant un moyen de « traduire » les situations et conventions de communication en usage au sein de ces aidants.

5 Cadre de conception

Un enjeu pour le développement d'outils de coopération et communication en ligne est de permettre la réalisation de pratiques qui existent d'ores et déjà hors ligne, comme le soutien social, avant l'introduction de tels outils. Le soutien social dans sa réalisation en ligne n'est qu'un cas particulier de l'activité de soutien social.

Ces pratiques manifestent à différents niveaux certaines régularités dans leur déroulement auxquels nous nous intéressons pour dégager des scripts d'interactions (Schank & Abelson 1977) plus généraux qui guident leur actualisation. Ainsi, nous avons pu, au travers des groupes de paroles et de nos rencontres avec les aidants,

observer sur le terrain des régularités dans les pratiques, au-delà de leur caractère situé, dans l'organisation des tours de table, dans l'initiation de la séance par la coordinatrice du réseau, dans la façon dont les membres se présentent les uns aux autres, s'expriment, posent des questions. Le caractère régulier et typique de ces situations nous donne à penser que celles-ci sont particulièrement pertinentes dans l'activité de soutien social et qu'il est intéressant de les *traduire* pour la conception d'un outil en ligne dédié. Pour nous, la problématique de conception de tels outils viserait à stimuler l'actualisation en ligne de ces situations de communication existant dans l'activité concrète, et donc familières aux utilisateurs.

Les résultats de notre analyse des plateformes existantes de soutien social en ligne nous semblent illustrer cette idée. Ainsi, un exemple de cadre stimulant l'actualisation d'une situation de communication familière aux utilisateurs serait la fonctionnalité de Questions/Réponses (Q&R). En effet le « jeu » de questions/réponses est une situation qui, même si elle n'est pas formellement courante, a une référence intuitive dans l'expérience de tout un chacun. Ainsi, la situation de Q&R, qui est une des situations typiques du soutien social, gagnerait à être présentée réellement comme une fonctionnalité de Q&R plutôt qu'à l'aide d'un dispositif général d'échange de messages comme un forum. Nous pensons qu'ainsi, l'utilisateur aura une appréhension plus intuitive de la situation de communication qui est proposée, car il pourra s'appuyer sur le schéma général de la situation, son script, pour guider le déroulement de ses interactions avec le système et les autres utilisateurs. Les utilisateurs sont bien sûr tout à fait capables d'utiliser un dispositif relativement abstrait comme un forum de discussion. Mais nous soulignons toutefois que celui-ci n'a pas de traduction évidente dans l'expérience de beaucoup de gens. En effet, le fait de s'adresser à un public relativement diffus, « à la cantonade », ne correspond à aucune des situations en face à face décrites dans le modèle du cadre participatif de Goffman (1981). Ceci explique selon nous en partie les difficultés de communication rencontrées par les utilisateurs des forums (Lewkowicz & Marcoccia 2004).

En faisant référence à des situations de communication existantes hors-ligne, les utilisateurs peuvent, au contraire, se reposer intuitivement sur les scripts qu'ils y attachent pour organiser le cours de l'interaction. La médiatisation modifie certes quelque peu la situation et invite les utilisateurs à renégocier en partie le déroulement de l'interaction, notamment lorsque l'outil apporte de nouvelles possibilités inexistantes dans la situation de référence, mais la charge reste plus faible.

Le problème délicat, une fois les scripts identifiés, reste de concevoir des fonctionnalités qui permettent et stimulent l'actualisation de ces scripts. Nous cherchons donc à donner une *traduction* à ces situations et aux scripts qui y sont attachés. Nous pensons que cette *traduction* passe à un premier niveau par l'identification des mots, signes et symboles qui sont utilisés en situation afin de s'appuyer sur ceux-ci, en les reproduisant sur l'interface, afin que la situation soit reconnue par les utilisateurs. Nous rejoignons en cela l'analyse de Norman (1999) dans le cadre de la conception d'interface homme-machine. Nous pensons que les signes et symboles affichés à l'écran font référence à des conventions partagées, par exemple entre concepteurs et utilisateurs et que c'est en cela qu'elles stimulent l'interaction. La notion de convention nous semble particulièrement intéressante afin

que les situations auxquelles nous faisons référence soient reconnues par les utilisateurs.

6 Illustrations

Une situation typique que nous avons pu observer lors de groupes de paroles, et qui illustre bien la place des conventions, est celle où les aidants se présentent lorsqu'un nouveau membre participe au groupe de parole. Une première prise sémiotique attachée à la convention est tout simplement la façon dont les participants se réfèrent et s'accordent sur le nom de la situation : ils « se présentent », ce qui est un aspect important pour toute situation conventionnelle par ailleurs. On trouve d'autres prises sémiotiques dans la façon de se présenter des membres du groupe de paroles qui est révélatrice d'un certain script en partie ordonné, mais surtout qui mobilise des mots et expressions particulières pour partager les informations qu'ils estiment pertinentes pour le nouveau participant.

Ils ne parlent pas du malade ou du patient dont ils ont la charge comme nous l'avons fait tout au long de cet article, ils parlent de leur « époux », de leur « père » ou de leur « mère » « dont la *pathologie*/i.e. la maladie d'Alzheimer a été diagnostiquée il y a X mois/années ». Ils évoquent si leur proche est « placé à *une certaine institution* » le cas échéant. Selon s'il s'agit d'aidant conjoint ou enfant, ils vont préciser la façon dont ils s'occupent de leur parent, elle ne sera d'ailleurs pas forcément précisée par un aidant conjoint puisque par défaut il vit 24 heure sur 24 avec la personne malade

La question du choix des informations que les aidants communiquent lorsqu'ils se présentent est bien entendu importante et relève du script de la situation. Mais elle fait finalement partie d'une question plus large qui est celle de comment l'on se présente conventionnellement dans le groupe de parole. Et nous pensons que pour le développement d'une fonctionnalité comme une page profil utilisateur, qui correspond à cette situation, il est important, d'une part de faire correspondre les champs aux étapes du script, mais également, et nous souhaitons ici attirer l'attention sur cet aspect, d'utiliser les même mots et expressions afin de guider l'utilisateur dans l'actualisation de cette situation de présentation de soi à la communauté. Ainsi, il serait proposé à l'utilisateur de « se présenter », d'indiquer que « *ma femme* est atteinte de *la maladie d'Alzheimer*. *Elle* a été diagnostiquée il y a *N mois/année* » (les informations susceptibles de choix de la part de l'utilisateur ou d'inférence du système sont indiquées en italique).

Un second niveau de cette traduction au delà du fait que la situation conventionnelle de communication soit reconnue, est de faire en sorte que le dispositif permette son bon déroulement au travers de ses fonctionnalités, comme nous allons l'illustrer maintenant.

Un autre exemple de situation typique dans les groupes de paroles que nous avons pu observer est l'initiation de « tour de table » autour d'un sujet sur lequel chacun est invité à partager son expérience, son vécu ou son opinion. Par ailleurs, ces situations de tours de tables constituent un moment privilégié permettant aux aidants de partager des conseils et des connaissances pratiques pour gérer leurs difficultés au quotidien. Tous les participants ont une idée, un script du déroulement de cette situation

relativement formalisé dans le groupe, et qui focalise des attentes auxquelles nous nous intéressons pour illustrer la spécification de fonctionnalités sur la base de scripts. Ainsi :

La coordinatrice du réseau initie la situation en proposant une question ou un thème sur lequel réagir (i.e. les structures d'accueil de jour) et se réfère à la situation « Est-ce que chacun peut raconter son expérience des centres d'accueil de jour ? ». La coordinatrice s'assure de l'accord et de l'attention du groupe. Elle désigne ensuite la personne qui sera la première à s'exprimer sur le sujet en se tournant vers elle ou en s'adressant à elle explicitement : « Mme *nom* qu'est ce que vous pensez de… / comment cela se passe avec votre mari quand il est à *nom de l'institution* ? ».

L'aidant s'exprime sur le sujet.

Les autres membres du groupe peuvent réagir.

La coordinatrice peut relancer la personne pour qu'elle donne plus de détails ou recadrer le discours en cas de digression ou en provoquer une elle-même si un point demandant des précisions est abordé par l'aidant.

Une fois que l'aidant a pu dire tout ce qu'il avait à dire, il passe la parole en confirmant auprès du prochain aidant qui s'exprimera (se tourner vers lui et montrer son attente) ou auprès de la coordinatrice (simple jeu de regard) qu'il n'a plus rien à ajouter.

Les échanges se poursuivent jusqu'à ce que chacun ait pu s'exprimer ou que le tour de table prenne fin faute de temps.

La coordinatrice peut éventuellement faire une synthèse pour clore le tour de table.

Dans le cadre d'une situation de tour de table les utilisateurs sont en attentes d'être invités à s'exprimer sur un sujet où chacun d'eux sera amené à apporter sa contribution et à pouvoir recevoir les commentaires et appréciations des autres utilisateurs. On peut sur la base de ce script proposer une première spécification d'une fonctionnalité d'initiation de tour de table pour une plate-forme de soutien social en ligne. Ainsi un utilisateur pourrait initier le tour par un message précisant le sujet qui serait adressé à un certain groupe d'utilisateurs. Les contributions de chacun seraient rassemblées sous une même rubrique et chacun des participants pourrait réagir à l'aide d'un champ à la manière des commentaires sur un billet de blog. Bien entendu, il ne s'agit pas de reproduire exactement le script mais plutôt d'en proposer un analogue compte tenu des contraintes et nouvelles possibilités offertes par la médiatisation. En capitalisant les contributions des aidants au travers de cette fonctionnalité de tours de tables, ils conserveront le contexte des conseils et savoir-faire qu'ils partagent, lequel est indispensable à leur mobilisation en tant que connaissances (Charlet, 2005).

Pour la clarté du discours, nous avons illustré séparément chacun des deux guides (conventions et scripts) que nous proposons d'utiliser pour traduire des situations typiques en fonctionnalités. Mais nous insistons sur le fait que les deux dimensions sont indissociables afin de parvenir à une spécification complète qui permette de concevoir une fonctionnalité dont la situation de référence soit reconnue au travers de signes conventionnels par les utilisateurs, et qui satisfasse leurs attentes en regard du script qu'ils ont en tête.

7 Conclusion et perspectives

Dans cet article, nous avons décrit la situation difficile d'aidants de patients atteints de pathologies lourdes, et le rôle que les TIC pourraient jouer pour les aider à assumer ce fardeau. Notre analyse des outils existants pour le soutien social nous a permis d'identifier le besoin, pour les concepteurs de tels dispositifs, d'être ancrés dans des situations réelles. Nous avons alors présenté notre démarche de conception, basée sur la traduction de situations de communication prototypiques. Nous faisons l'hypothèse que cette démarche nous permettra de concevoir un système qui sera le plus proche possible des pratiques et attentes des utilisateurs, et qui sera d'usage intuitif. Ce projet interdisciplinaire combinera plusieurs dimensions d'analyse qui viendront enrichir notre démarche de conception autour de l'idée de traduction. Tout d'abord, nous enrichirons notre description des situations à l'aide d'entretiens avec les aidants. Ces entretiens sont en cours de réalisation par un collègue sociologue. Par ailleurs, nous complèterons nos résultats sur les scripts à l'appui d'analyses d'échanges de soutien social observés sur des forums sur lesquels travaillent nos collègues linguistes. Lorsque cette analyse fine des situations de soutien social sera terminée, nous serons en mesure de développer la plate forme de soutien social en ligne pour les aidants. Cette plate forme sera présentée aux aidants familiaux soutenus par le RPM, mais elle sera ouverte et donc disponible sur Internet, pour des aidants en dehors du réseau de santé. Enfin, nous évaluerons cette plate forme, en premier lieu en vérifiant son usage par les aidants que nous suivons, et ainsi l'efficience empirique de notre démarche (Bachimont, 2004). Par la suite, nous mobiliserons la notion de capital social afin d'avoir une vision synthétique de l'évolution de la communauté et du rôle de la plate forme dans ce phénomène. Nous nous intéresserons en particulier à la question des TIC comme éventuel catalyseurs de communautés.

Références

BACHIMONT B. (2004). Pourquoi n'y a-t-il pas d'expérience en ingénierie des connaissances ? In Actes de la conférence « Ingénierie des connaissances (IC2004) », N. Matta (ed), Lyon. Presses Universitaires de Grenoble.

BARNES, M.K., & DUCK, S. (1994). Everyday communicative contexts for social support. In B. R. Burleson, T. L. Albrecht & I. G. Sarason (Eds.), Communication of social support: Messages, interactions, relationships and community. Thousand Oaks: Sage. p. 175-194.

CHARLET J. (2005). L'ingénierie des connaissances, une science de gestion ? In Teulier R. et Lorino P., coordinateurs, Entre la connaissance et l'organisation, l'activité collective, chapitre 11. La découverte. Actes du colloque de Cerisy « Activité, connaissance, organisation ».

CHENG L., STONE L., FARNHAM S., CLARK A.M. & ZANER M. (2000). HutchWorld: Lessons Learned-A Collaborative Project: Fred Hutchsinson Cancer Research Center & Microsoft Research, Proceedings of Virtual Worlds Conference, 2000, p. 1-12.

GAGLIO GERALD & ATIFI HASSAN (2008). L'entraide en mots : le cas d'un forum de discussion des «marocains d'ailleurs», Congrès AISLF, Istanbul, 8-11 juillet 2008.

GOFFMAN, E. (1981). Forms of talk. Oxford, Basil Blackwell.

GUSTAFSON D. H., HAWKINS R. P., BOBERG E. W., MCTAVISH F., OWENS B., WISE M., BERHE H. & PINGREE S. (2002). CHESS: 10 years of research and development in consumer health informatics for broad populations, including the underserved," International Journal of Medical Informatics, vol. 65, 2002, p. 169-177.

IFOP (2008). Etude national "Connaître les aidants and leurs attentes". http://www.aveclesaidants.fr/index.php?rub=alaune&ssrub=enbref&lid=522#contenu

KROES P. (2006). Coherence of structural and functional descriptions of technical artefacts, Studies In History and Philosophy of Science Part A, Volume 37, Issue 1, The dual nature of technical artefacts, 2006, p. 137-151.

LEIMEISTER J. M. & KRCMAR H. (2005). Acceptance and Utility of a Systematically Designed Virtual Community for Cancer Patients. Proceedings of the Second Communities and Technologies Conference, Milano, p. 129-149.

LEWKOWICZ M., MARCOCCIA M., ATIFI H., BÉNEL A., GAGLIO G., GAUDUCHEAU N. & TIXIER M. (2008). Online Social Support: Benefits of an Interdisciplinary Approach for Studying and Designing Cooperative Computer-Mediated Solutions. Proceedings of the 8th Conference on the Design of Cooperative Systems p 99-110.

LEWKOWICZ, M. & MARCOCCIA, M. (2004). The Participative Framework as a design model for newsgroups: PartRoOM, in Darses, F., Dieng, R., Simone, C., Zacklad, M., Cooperative Systems Design, IOS Press p. 243-257.

MALONEY-KRICHMAR D. & PREECE J. (2005). A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community, ACM Transactions on Computer-Human Interaction, vol. 12, p. 201-232.

NORMAN, D. A. (1999). Affordance, conventions, and design. interactions 6, 3 (May. 1999), p. 38-43.

PIXEL (2000). Etude PIXEL - L'entourage familial des patients atteints de la maladie d'Alzheimer. Novartis. http://www.mediathequenovartis.fr/novartis/spip.php?article107

PILLEMER K. & SUITOR J. J. (2002). Peer Support for Alzheimer's Caregivers: Is it Enough to Make a Difference? Research on Aging 2002; 24; p. 171

SALEM D. A., BOGAT G. A. AND REID C. (1997). Mutual help goes online Journal of Community Psychology, 25(2), 189-207.

SCHANK, R.C. & ABELSON, R. (1977). Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Earlbaum Assoc.

THOITS, P. A. (1986). Social support as coping assistance. Journal of Consulting and Clinical Psychology, 54, 4, (Aug. 1986), p. 416-423.

XIE B. (2008). Multimodal Computer-Mediated Communication and Social Support among Older Chinese Internet Users, Journal of Computer-Mediated Communication, vol. 13, 2008, p. 751-767.

Méthodologie assistée de conception d'une ontologie à partir d'une conceptualisation consensuelle semi-formelle

Michel Héon[1], Gilbert Paquette[1], Josianne Basque[1]

[1] Centre de recherche LICEF, Télé-Université, Montréal, Canada
(michel.heon, gilbert.paquette, josianne.basque)@licef.ca

Résumé : Cet article présente une méthodologie assistée de conception d'une ontologie à travers trois méthodes, soit une méthode d'élicitation des connaissances d'un domaine résultant en un modèle semi-formel de ces connaissances, une méthode de formalisation conduisant à la production d'une ontologie et une méthode de validation syntaxique et sémantique de l'ontologie. Les processus de formalisation et de validation sont assistés par un système expert à la formalisation dont la base de connaissances est une ontologie de transformation.

Mots-clés : Formalisation des connaissances, Ingénierie des connaissances, Ingénierie ontologique, Élicitation, Méta-représentation, Ontologie de transformation, Système expert, Conception assistée d'ontologie.

1 Introduction

Cet article apporte une contribution au domaine de l'ingénierie ontologique (Dietz, 2006 ; Gašević *et al.*, 2006 ; Gómez-Pérez *et al.*, 2003 ; Gruber, 1995 ; Uschold et Gruninger, 1996). Dans ses grandes lignes, la construction d'une ontologie nécessite une étape d'élicitation des connaissances du domaine visé, suivie d'une étape de formalisation et d'une étape de validation. La nature des connaissances à représenter, l'hétérogénéité du support des connaissances à formaliser (documents plus ou moins formels, communication orale ou écrite, représentations graphiques, connaissances tacites, etc.) et le manque d'assistance automatisée, rendent laborieuse la démarche de conception d'une ontologie. À l'intérieur des courants actuels, notre méthodologie préconise la conception d'ontologies génériques et réutilisables tel que le propose (Gangemi *et al.*, 2007) du projet NeOn (www.neon-project.org) et peut facilement s'intégrer à des infrastructures technologiques à base d'ontologies comme KAON (kaon.semanticweb.org), On-To-Knowledge (Davies *et al.*, 2003) et TelOs du projet LORNET (www.lornet.org).

Nous adoptons la définition d'une ontologie proposée par Gruber (1995) qui stipule qu'une ontologie est un document formel (au sens de Uschold et Gruninger (1996)) dont le contenu et la sémantique sont traitables par des systèmes informatiques et dont la connaissance qui y est exprimée est obtenue de manière consensuelle. La stratégie

de co-modélisation des connaissances à l'aide d'un langage graphique semi-formel expérimentée par Basque *et al*. (2008b) s'accorde bien à l'aspect consensuel de cette définition d'ontologie. Cette activité nous semble ainsi pouvoir être mise à profit pour la conception d'une ontologie. Ainsi, nous pensons que le processus de construction d'une ontologie doit être décomposé en deux phases bien distinctes: une phase *d'élicitation* de la connaissance dans un formalisme de degré semi-formel relativement évolué, puis une phase de *formalisation* des connaissances où le modèle semi-formel est transformé dans un formalisme ontologique.

Malgré le fait qu'un modèle semi-formel peut comporter des éléments d'ambigüité, la souplesse et le caractère moins artificiel d'un langage semi-formel permettent d'accéder plus facilement à l'expression de connaissances tacites, car la spontanéité n'est pas bloquée par la charge cognitive associée à une formalisation plus poussée (Basque *et al.*, 2008b). L'usage d'un système de représentation plus convivial permet aussi d'élargir le bassin des personnes aptes à représenter leurs connaissances en même temps qu'il évite la démobilisation des experts de leur tâche principale, laquelle coûte cher à une organisation. L'élicitation dans un formalisme de degré semi-formel peut ainsi représenter une économie de temps et surtout un gain de qualité dans la représentation des connaissances. De plus, la convivialité du langage semi-formel fait en sorte que des modèles de degré semi-formel peuvent être conçus par les experts de contenu sans l'assistance d'un ingénieur de la connaissance, lequel peut ensuite les formaliser avec la participation minimale des experts les ayant conçus ou encore d'autres experts. Les opérations d'élicitation peuvent ainsi être intégrées de manière plus souple dans les activités des experts de contenu.

Partant de l'hypothèse qu'un modèle semi-formel peut être formalisé sous la forme d'une ontologie sans perdre les distinctions entre les différents types de connaissances (notamment entre les connaissances de nature déclarative, procédurale et stratégique au sens de (Paquette, 2002 ; Paris *et al.*, 1983)), cet article présente une méthodologie de conception d'une ontologie, assistée par un système expert, qui inclut une méthode d'élicitation, une méthode de formalisation et une méthode de validation. Nous comptons, par cette méthodologie, rendre plus accessible et plus rapide la construction d'une ontologie.

2 Le système de représentation des connaissances de la méthodologie

Issu du domaine de l'ingénierie pédagogique, le langage de modélisation par objets typés MOT, de degré semi-formel, est celui qui a été utilisé pour développer la méthodologie proposée. Notre but est toutefois de faire en sorte que celle-ci puisse s'appliquer à d'autres langages semi-formels tels que ceux utilisés pour les cartes conceptuelles, les réseaux sémantiques ou les diagrammes UML. Le langage et le logiciel qui l'implémente (MOT*Plus*) ont été conçus par une équipe sous la direction du second auteur (Paquette, 2002) au Centre de recherche LICEF de la Télé-Université (www.licef.ca).

Le langage semi-formel MOT différencie les types de connaissances au moyen de symboles graphiques (voir le **tableau 1** et le **tableau 2**). Les connaissances peuvent être combinées au sein d'un même schéma de manière à produire des modèles mixtes de connaissances.

Tableau 1. Catégories des connaissances dans le langage MOT

Catégorie de connaissance	Connaissance abstraite		Connaissance factuelle	
Déclarative *le quoi des choses*	Concept		Exemple	
Action *Le comment de choses*	Procédure		Trace	
Stratégique *Le pourquoi, le quand*	Principe		Énoncé	

Le *concept* représente « le quoi » des choses. Il sert à décrire l'essence d'un objet concret. Il peut être associé à l'idée de classe ou de catégorie. En ce sens, il est l'abstraction d'un objet concret. L'*exemple* représente l'un de ces objets en énonçant un certain nombre de faits qui le décrivent. La *procédure* permet de décrire « le comment » des choses. Elle désigne des opérations, des actions pouvant être accomplies. La *trace* représente l'ensemble des faits concrets obtenus lors de l'exécution d'une procédure. Le *principe* désigne « le pourquoi », « le quand » d'une chose. Il est une connaissance stratégique qui permet de nommer une relation qui existe entre des objets, que ce soit des concepts, des procédures ou d'autres principes. Il sert notamment à représenter une condition pouvant s'appliquer à l'exécution d'une action. L'*énoncé* représente l'instanciation d'un principe à propos d'objets concrets.

Tableau 2. Sémantique des relations typées dans MOT

Type de lien	Signification
S	Le lien de ***s****pécialisation* associe deux connaissances abstraites de même type dont la première est une spécialisation de la seconde. Ce lien est notamment utile dans la description des taxonomies.
I	Le lien d'***i****nstanciation* associe à une connaissance abstraite l'ensemble des faits qui caractérisent une instance de cette connaissance.
I/P	Le lien ***i****ntrant/****p****roduit* sert à associer une connaissance procédurale à une connaissance conceptuelle afin de représenter l'intrant ou le produit d'une procédure. Ce lien est notamment utile dans la description des algorithmes, des processus et des méthodes.
P	Le lien de ***p****récédence* associe une connaissance à une autre qui la suit dans une séquence temporelle de procédures ou de règle de décision (principes).
R	Le lien de ***r****égulation* associe une connaissance stratégique (un *Principe* ou u*n Énoncé*) à une autre connaissance afin de préciser une contrainte, une restriction ou une règle qui régit la connaissance.
C, C*	Les liens de ***c****omposition* et de *composition multiple* permettent de représenter l'association entre une connaissance et des connaissances qui la composent.

Les *relations* sont des liens directionnels qui unissent des connaissances. Le langage MOT offre un ensemble de liens qui sont typés (voir le **tableau 2**). Chaque

type de lien possède une sémantique propre qui respecte des règles d'intégrité. Par exemple, un lien de spécialisation (lien S) unit deux connaissances abstraites qui doivent être de même nature. L'ensemble de ces règles d'intégrité est décrit dans Paquette (2002).

3 Méthodologie de conception d'une ontologie

La **fig. 1** présente les trois principales méthodes qui composent la méthodologie que nous proposons. Dans l'ordre, ces méthodes concernent les procédures : *concevoir un modèle semi-formel*, *formaliser en ontologie* et *valider l'ontologie*.

La méthode *concevoir un modèle semi-formel*, basée sur le principe de co-modélisation, a pour objectif la conception d'un *modèle semi-formel de domaine* impliquant la collaboration d'au moins deux *experts de contenu* et optionnellement d'un *ingénieur de la connaissance* qui pilote l'activité de modélisation. Un ou plusieurs novices du domaine peuvent remplacer l'un des experts d'un contenu ou se joindre à ceux-ci afin de favoriser à la fois le processus d'explicitation des connaissances par les experts (suscité par les questions pointues et contextualisées des novices) et le transfert d'expertise chez les novices (Basque *et al.,* 2008). À ce jour, cette étape a été réalisée en présence dans nos travaux, mais nous envisageons de développer une version collaborative de l'outil MOT*Plus* qui permettrait de la réaliser à distance.

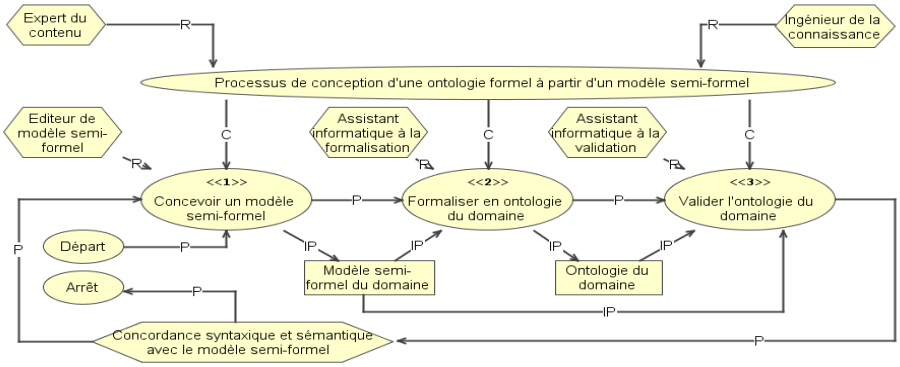


Fig. 1: Schéma MOT de la méthodologie de conception d'une ontologie

Lorsque le modèle semi-formel est jugé satisfaisant pour les acteurs participant aux séances d'élicitation, l'ingénieur a la responsabilité de piloter la *méthode de formalisation* qui a pour extrant une *ontologie du domaine*. Assisté par un *assistant informatique à la formalisation,* qui est en fait un système expert utilisant une *ontologie de transformation*, l'ingénieur formalise le modèle semi-formel de domaine en une ontologie de domaine prête à être validée.

Finalement, la *méthode de validation*, pilotée cette fois conjointement par l'ingénieur de la connaissance et l'expert de contenu, permet la *validation syntaxique et sémantique* de l'ontologie de domaine. La validation syntaxique a pour objectif

d'assurer une concordance entre les éléments déclarés dans le modèle semi-formel et la représentation de ces éléments dans l'ontologie. La validation sémantique a pour objectif de valider la concordance sémantique entre le modèle semi-formel et l'ontologie. *L'assistant informatique à la validation* permet de tirer des conclusions automatiques à partir de l'ontologie du domaine, qui sont comparées avec celles obtenues humainement à partir de l'interprétation du modèle semi-formel.

3.1 Méthode d'élicitation

L'élicitation des connaissances est un domaine de recherche vaste et complexe qui a fait l'objet de nombreux travaux depuis plus d'une vingtaine d'années (Hart, 1986). De récentes recherches de Basque *et al.*(2008a ; 2008b) ont montré que des activités de transfert d'expertise impliquant la co-modélisation entre experts et novices permettent de produire des modèles de connaissances dont le contenu ont les propriétés d'être consensuelles et riches en connaissances du domaine.

Afin de supporter et de stimuler la démarche intellectuelle des acteurs dans l'activité de co-modélisation, il importe de mettre à leur disposition un langage de représentation qui, par sa sémantique, impose une certaine structure à la pensée, de manière à en favoriser son déploiement, et ce, sans en entraver la créativité. Les langages de modélisation de degré semi-formel, tel que MOT, répondent à ces exigences Basque *et al.* (2008a ; 2008b). Des recherches menées dans des contextes d'apprentissage ont mis en évidence les médiations cognitives suscitées par l'usage d'un tel langage lors de l'élaboration de représentations graphiques de connaissances (Basque et Pudelko, sous presse).

Le schéma de la **fig. 2** présente la méthode utilisée pour concevoir un modèle semi-formel selon cette approche.

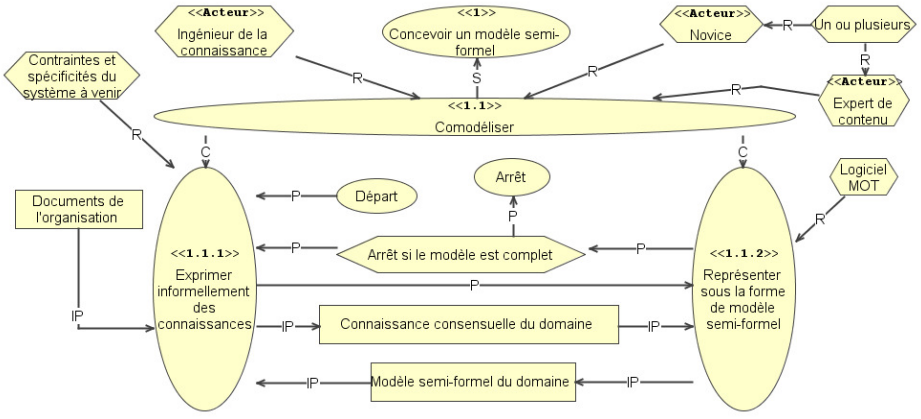


Fig. 2: La méthode de conception d'un modèle semi-formel fondée sur une stratégie de comodélisation

Le processus de *comodélisation*, qui est ici représenté en tant que sorte de processus d'*élicitation*, se compose d'une activité d'*expression informelle des*

connaissances et d'une activité de *représentation structurée semi-formellement des connaissances consensuelles de domaine*, ce qui conduit à la production, avec le logiciel MOT, d'un *modèle semi-formel du domaine*. Ces deux activités sont itératives et le produit de l'un sert d'intrant à l'autre. L'expression informelle est régie par les *contraintes et les spécifications du système à venir*. Des *documents contenant des savoirs et des savoir-faire de l'organisation* peuvent également servir en intrant à l'activité. Le processus de comodélisation se termine lorsque les éléments d'intérêts du domaine ont été représentés à la satisfaction des acteurs en présence.

3.2 Méthode de formalisation

La formalisation d'un modèle semi-formel en une ontologie est un processus délicat dont l'objectif est de produire une ontologie du domaine. Pour qu'elle soit opérable, l'ontologie du domaine doit à la fois respecter la syntaxe du langage formel et représenter la sémantique du domaine contenue dans le modèle semi-formel

L'application de la méthode de formalisation est sous la responsabilité de l'*ingénieur de la connaissance* qui est assisté par un *assistant informatique à la formalisation* (voir la **fig. 3**). La méthode de formalisation se décompose en trois processus: *importer dans l'espace de modélisation ontologique*, *désambiguïser* et *transformer en ontologie*.

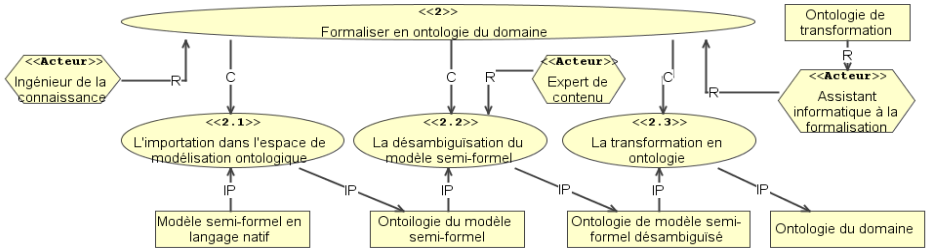


Fig. 3: Méthode de transformation du modèle semi-formel de domaine en ontologie de domaine

3.2.1 Architecture des composants de la méthode

Le diagramme de composants de la **fig. 4** présente les différents éléments et les relations qui composent l'assistant informatique à la formalisation. D'un point technologique, l'assistant est une application Eclipse (www.eclipse.org) qui hérite des propriétés de gestion des données, de conception d'interface graphique, d'intégration de "plug-in", d'interopérabilité par l'utilisation de Java et de capacité de développement des aspects procéduraux nécessaires à la tâche de formalisation et de communication avec l'utilisateur.

Une base de connaissances incarnée par l'*Ontologie de transformation* assure la cohérence entre les divers modules de l'assistant afin d'offrir un support intelligent à l'ingénieur. Quatre domaines distincts composent l'ontologie de transformation. Le domaine de *l'ontologie des langages semi-formels* contient un ensemble de

connaissances sur les langages semi-formels qui permettent aux modules d'importation d'intégrer les entités du modèle semi-formel dans l'espace de modélisation ontologique (transposition d'un modèle de format EMF en format OWL). L'ontologie contient les métamodèles des langages semi-formels supportés par la tâche de formalisation. L'*ontologie de désambiguïsation* contient les connaissances nécessaires à la désambiguïsation des éléments du modèle semi-formel qui seront transformés en une classification générique dans l'*ontologie de référence.* À chaque schéma de langage correspond une ontologie de désambiguïsation spécifique. Le domaine représenté par l'*ontologie de référence* permet de classifier de façon générique et non ambigüe les différents éléments du modèle semi-formel. L'ontologie de référence sert d'adaptateur entre les diverses représentations possibles du modèle semi-formel et sa représentation finale sous forme d'ontologie du domaine.

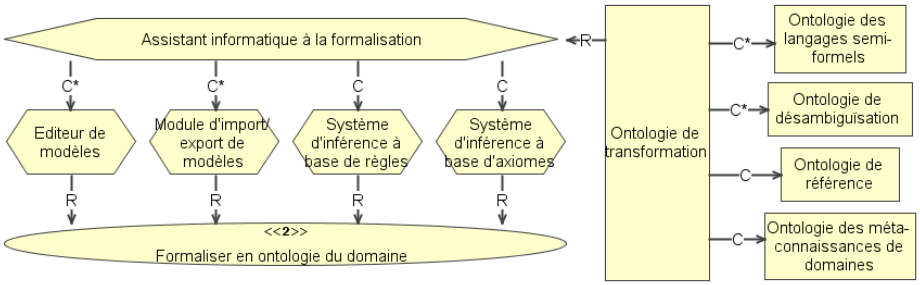


Fig. 4: Composants de l'assistant à la formalisation

Finalement, le domaine représenté par *l'ontologie des méta-connaissances de domaine,* qui est une ontologie de niveau générique au sens de Oberle (2006 p. 46), englobe un ensemble de méta-connaissances concernant les domaines d'application. Elle contient notamment des méta-connaissances au sujet de connaissances procédurales, déclaratives et stratégiques et des méta-propriétés telles que A-POUR-COMPOSANT, EST-LE-PRECEDENT-DE (voir la **fig. 5**). Cette ontologie générique est considérée comme un langage de représentation formel de domaines d'application et elle est importée par l'ontologie du domaine faisant ainsi partie résultat de la

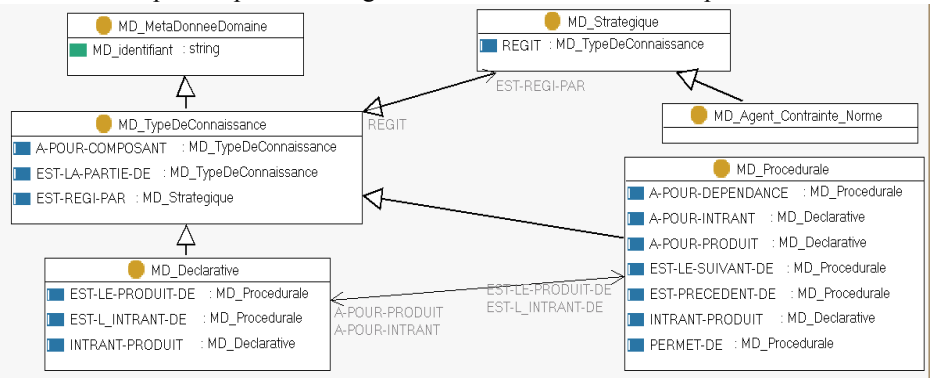


Fig. 5: Ontologie générique des méta-connaissances de domaines

méthode de formalisation.

3.2.2 Processus d'importation, de désambiguïsation et de transformation

Le processus automatique d'*importation* de modèles semi-formels assure la traduction d'un modèle de format natif à l'éditeur semi-formel vers un modèle de format OWL. La manipulation des éléments ontologiques (classes, propriétés, axiomes et individus) est alors interfacée par les classes Java générées. Nos travaux antérieurs ont permis de confirmer la validité de cette approche pour l'importation de modèles semi-formels MOT dans l'espace ontologique OWL (Héon *et al.*, 2008).

Le processus de *désambiguïsation* consiste à supprimer les ambigüités contenues dans le modèle semi-formel. Trois catégories de test sont réalisées, soit les tests *typologique, topologique* et *sémantique.* La *désambiguïsation typologique* est la plus automatique des désambiguïsations. Elle associe le type d'un composant du modèle semi-formel à un élément ontologique. Par exemple (voir le cas 1 du **tableau 3** et du **tableau 4**), le composant « lien S » (*sorte-de*) s'interprète comme une relation d'hyponymie et le composant « lien I » (*instance*) s'interprète comme une relation d'instanciation. La *désambiguïsation topologique* est plus complexe et peut, dans quelques cas, n'être que semi-automatisée. Cette désambiguïsation s'établit en caractérisant les composants du modèle par l'identification d'un patron de disposition. À l'exemple du cas 2 du **tableau 3** et du **tableau 4**, les concepts C4 et C5 sont interprétés comme des classes et le principe P1 comme une propriété (binaire) mettant en relation ces classes, grâce à l'identification du patron de disposition suivant: un *principe* est uni par un *lienR* à un *concept* en intrant et est uni par un *lienR* à un *concept* extrant. Également, les connaissances stratégiques P1 et A1 qui sont représentées par des *Principes* dans le modèle semi-formel sont désambiguïsées en *Propriété* pour P1 et en *connaissance stratégique* pour A1.

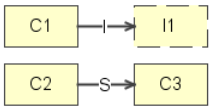
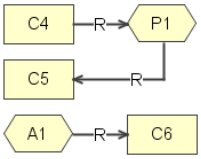
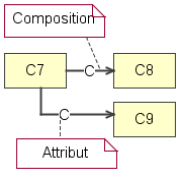
Du point de vue de l'ingénieur de la connaissance, l'étape de *désambigüisation selon la sémantique du domaine* est délicate, car elle nécessite une compréhension du domaine qui est modélisé, ce qui implique la collaboration de l'expert de contenu afin de répondre aux questions de l'ingénieur. L'exemple d'interprétation du *lien de composition* dans le langage MOT est une bonne illustration de ce type d'ambiguïté (voir le cas 3 du **tableau 3**). Le lien de composition s'interprète de deux façons. La première interprétation possible est la composition entre concepts (exemple: C7 *a pour partie* C8). La deuxième interprétation possible est l'utilisation du lien de composition pour assigner des attributs à un concept (exemple : C7 *a pour attribut* C9), c'est-à-dire que la relation unit une classe à un objet *DataType.* Seule une connaissance adéquate du domaine de connaissances permet de lever l'ambiguïté liée à cette double interprétation du *LienC*.

Le dernier processus de la méthode de formalisation est l'*application des règles de transformation* afin de transformer le modèle semi-formel désambiguïsé en ontologie formelle. Ce processus est réalisé par un système expert utilisant des axiomes et des règles du *Semantic Web Rule Language* (SWRL) (Horrocks *et al.*, 2004). Le **tableau 4** présente le résultat de la formalisation en ontologie du domaine du modèle semi-formel présenté au **tableau 3**.

3.2.3 Expérimentation

Le **tableau 3** présente trois cas de figure de modélisation semi-formelle à formaliser. Les résultats de la réalisation des étapes de modélisation (**tableau 3a**); d'importation (**tableau 3b**), de désambiguïsation (**tableau 3c**) et de transformation (**tableau 4**) y sont respectivement représentés en langage MOT et OWL dans la Notation 3 (N3) (Berners-Lee, 1998).

Tableau 3. Études de cas représentés en langage MOT et en langage OWL-N3.

<table>
<tr><th>Cas 1) Relations d'instanciation et de subsomption</th><th>Cas 2) Un principe en tant que propriété ou agent</th><th>Cas 3) Relations de composition ou d'attribution</th></tr>
<tr><td colspan="3">A) Représentation semi-formelle du domaine</td></tr>
<tr>
<td> 
</td>
<td> 
</td>
<td> 
</td>
</tr>
<tr><td colspan="3">B) Représentation après importation dans l'espace de modélisation ontologique</td></tr>
<tr>
<td>

```
:C1 a metaMot:MOT_Concept .
:C2 a metaMot:MOT_Concept .
:C3 a metaMot:MOT_Concept .
:I1 a metaMot:MOT_Exemple .
:LienI_C1_I1 a metaMot:MOT_LienI ;
   metaMot:MOT_connDestination :I1 ;
   metaMot:MOT_connSource      :C1 .
:LienS_C2_C3 a metaMot:MOT_LienS ;
   metaMot:MOT_connDestination :C3 ;
   metaMot:MOT_connSource      :C2 .
```

</td>
<td>

```
:C4 a metaMot:MOT_Concept .
:C5 a metaMot:MOT_Concept .
:C6 a metaMot:MOT_Concept .
:P1 a metaMot:MOT_Principe .
:A1 a metaMot:MOT_Principe .
:LienR_C4_P1 a metaMot:MOT_LienR ;
   metaMot:MOT_connDestination :P1 ;
   metaMot:MOT_connSource      :C4 .
:LienR_P1_C5 a metaMot:MOT_LienR ;
   metaMot:MOT_connDestination :C5 ;
   metaMot:MOT_connSource      :P1 .
```

</td>
<td>

```
:C7 a metaMot:MOT_Concept .
:C8 a metaMot:MOT_Concept .
:C9 a metaMot:MOT_Concept .
:LienC_C7_C8 a metaMot:MOT_LienC ;
   metaMot:MOT_connDestination :C8 ;
   metaMot:MOT_connSource      :C7 .
:LienC_C7_C8 a metaMot:MOT_LienC ;
   metaMot:MOT_connDestination :C9 ;
   metaMot:MOT_connSource      :C7 .
```

</td>
</tr>
<tr><td colspan="3">C) Représentation après la désambiguïsation</td></tr>
<tr>
<td>Cas 1)</td>
<td colspan="2">

```
:C1 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C2 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C3 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:I1 a oRef:OR_Entite_Observable_Declaratif , metaMot:MOT_Exemple .
:LienI_C1_I1 a oRef:OR_Relation_Instance , metaMot:MOT_LienI ; ...
:LienS_C2_C3 a oRef:OR_Relation_Hyponyme , metaMot:MOT_LienS ; ...
```

</td>
</tr>
<tr>
<td>Cas 2)</td>
<td colspan="2">

```
:C4 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C5 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C6 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:P1 a oRef:OR_Relation_Propriete_Objet , metaMot:MOT_Principe .
:A1 a oRef:OR_Entite_Abstraite_Strategique_Agent , metaMot:MOT_Principe .
:LienR_P1_C5 a oRef:OR_Relation_Propriete_Regulation , metaMot:MOT_LienR ; ...
:LienR_C4_P1 a oRef:OR_Relation_Propriete_Regulation , metaMot:MOT_LienR ; ...
:LienR_A1_C6 a oRef:OR_Relation_Propriete_Regulation , metaMot:MOT_LienR ; ...
```

</td>
</tr>
<tr>
<td>Cas 3)</td>
<td colspan="2">

```
:C7 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C8 a oRef:OR_Entite_Abstraite_Declaratif , metaMot:MOT_Concept .
:C9 a oRef:OR_Entite_Schema_String , metaMot:MOT_Concept ;
:LienC_C7_C8 a oRef:OR_Relation_Holonyme , metaMot:MOT_LienC ; ...
:LienC_C7_C9 a oRef:OR_Relation_Propriete_Attribut , metaMot:MOT_LienC ; ...
```

</td>
</tr>
</table>

Chacun des cas est associé à un type de désambiguïsation particulier. Ainsi, le premier cas, qui présente un modèle composé des liens de subsomption et d'instanciation, fait appel à une désambigüisation de type *typologique* qui classe: les *Concepts* C1, C2, C3 en tant qu'*owl:class,* avec C2 qui subsume C3; et l'*Exemple* I1 en tant qu'individu OWL appartenant à C1. Le deuxième cas, qui fait référence à une

désambiguïsation *topologique,* permet de désambiguïser les *Principes* P1 et A1 respectivement en *owl:ObjectProperty* et *owl:class* de catégorie *Agent_Contrainte_Norme.* Finalement, le troisième cas nécessite une compréhension du domaine afin de désambiguïser de façon *sémantique* l'interprétation du *LienDeComposition*, qui: entre C7 et C8 se formalise par une *owl:ObjectProperty* de catégorie *A-POUR-COMPOSANT* dont le domaine est l'*owl:class* C7 et l'*owl:class* C8; et entre C7 et C9 qui se formalise par une *owl:DatatypeProperty* de catégorie *A-POUR-ATTRIBUT* dont le domaine est l'*owl:class* C7 et l'image une *xsd:string.*

Tableau 4. Modèle semi-formel formalisés en ontologie et représentés en OWL-N3

Cas 1)

```
:C1 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C3 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C2 a owl:Class ;
  rdfs:subClassOf :C3 .
:I1 a :C1 .
```

Cas 3)

```
:C7 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C8 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C7_aPourComposant_C8 a owl:ObjectProperty ;
  rdfs:subPropertyOf metaDom:A-POUR-COMPOSANT .
:C7_aPourAttribut_C9 a owl:DatatypeProperty ;
  rdfs:domain :C7 ;
  rdfs:range xsd:string ;
  rdfs:subPropertyOf metaDom:A-POUR-ATTRIBUT .
```

Cas 2)

```
:A1 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Agent_Contrainte_Norme .
:C4 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C5 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:C6 a owl:Class ;
  rdfs:subClassOf owl:Thing , metaDom:MD_Declarative .
:P1 a owl:ObjectProperty ;
  rdfs:domain :C4 ;
  rdfs:range :C5 .
:A1_regit_C6 a owl:ObjectProperty ;
  rdfs:domain :A1 ;
  rdfs:range :C6 ;
  rdfs:subPropertyOf metaDom:REGIT .
```

3.3 Méthode de validation

La méthode de validation se divise en deux thèmes, soit la validation syntaxique et la validation *sémantique.* La validation syntaxique a pour but d'assurer que tous les éléments du modèle semi-formel sont représentés dans l'ontologie selon les règles du langage ontologique utilisé, dans notre cas, OWL-DL. La réalisation de cette validation se décompose en deux étapes, soit *générer un modèle semi-formel* à partir de l'ontologie du domaine et *comparer les éléments* du modèle semi-formel généré avec les éléments de modèle semi-formel d'origine. Si la concordance est exacte, alors il est véridique de conclure que l'ontologie du domaine représente syntaxiquement l'ensemble des éléments du modèle semi-formel du domaine.

La validation sémantique s'attarde sur le sens des représentations qui sont consignées dans l'ontologie. À titre d'exemple, une erreur occasionnellement commise par les concepteurs de modèles peu familiers au langage MOT est issue d'une confusion entre le lien S ("sorte de" [*is-a*]) et le lien C ("composé de" [part-of]). Par exemple, ils pourraient être tentés d'exprimer que *planète* se compose de *vénus*, *terre*, *mars*, etc., au lieu d'exprimer que *vénus*, *terre*… sont des sortes de *planètes*. Une inférence sur l'ontologie de domaine qui déterminerait que la propriété inverse de *composeDe* est *faitPartieDe* permettrait de conclure, que *vénus, terre, mars…* font partie de *planète*. Ce type de conclusion erronée devrait servir de signal à l'ingénieur et aux experts concernant une possible erreur de sémantique dans la représentation du domaine.

La première étape à réaliser pour une validation sémantique est la production de *conclusions formelles* en appliquant des inférences sur l'ontologie du domaine. Des scénarios de test de déductions peuvent êtres utilisés afin de formaliser d'avantage l'étape. L'étape de *comparaison des conclusions* entre les réponse qu'auraient données l'ingénieur et l'expert aux déduction et les conclusions automatiques permettent d'entreprendre une réflexion qui peut , d'une part, amener l'expert de contenu à revoir la façon de représenter ce qu'il exprime, ou d'autre part, mettre fin au processus de construction de l'ontologie du domaine. Gómez-Pérez (2004) présente les critères selon lesquels l'ontologie devrait être validée: la *consistance* identifie s'il y a des contradictions entre éléments ontologiques; la *complétude* assure que tous les éléments ontologiques sont soit explicitement déclarés ou soit inférables; la *concision* est un principe qui stipule que seuls les éléments à être définis doivent être définis; l'*expansibilité* est la capacité d'ajouter des nouvelles connaissances sans modifier les anciennes; la *sensibilité* est la capacité à réagir à des modifications.

4 Conclusion

Dans cet article, nous avons présenté une méthodologie de conception d'une ontologie à partir d'une représentation semi-formelle de domaine. Son originalité réside notamment dans le fait qu'elle est assistée par un système expert à la formalisation. Ce système expert assiste l'ingénieur au cours du processus de formalisation et assiste l'ingénieur et l'expert dans la validation syntaxique et sémantique de l'ontologie du domaine. La méthodologie intègre les principes de consensualité et de formalité propres aux ontologies par l'utilisation de la méthode de comodélisation pour la production de modèles semi-formels à l'étape d'élicitation. Finalement, il est démontré que la méthode permet la formalisation de connaissances procédurale et stratégique.

Dans cet article, la méthodologie utilise les langages MOT et OWL dans son application. Cependant, bien que ce ne soit pas ici démontré, on notera que la méthodologie est conçue afin qu'elle soit généralisée dans son application à d'autres langages semi-formel et formel. Cette particularité sera démontrée dans de futurs travaux.

Présentement, la méthodologie est validée de manière fonctionnelle à partir de modèles fictifs. Dans un avenir proche, nous comptons appliquer la méthodologie à une évaluation en laboratoire avec des cas, des experts non fictifs, et des modèles issus d'activités de co-modélisations.

Références

BASQUE J., DESJARDINS, C., PUDELKO, B. & LÉONARD, M. (2008a). Gérer les connaissances stratégiques dans des entreprises manufacturières de la Montérégie: expérimentation de la co-modélisation des connaissances dans 3 PME. Montréal/Canada, Rapport de recherche. Montréal: CEFRIO.: 118 p En ligne. <https://www.cefrio.qc.ca/upload/1599_rapportcefriovalotechfinal12nov08.pdf>.

BASQUE J., PAQUETTE, G., PUDELKO, B. & LÉONARD, M. (2008b). Collaborative Knowledge Modeling with a Graphical Knowledge Representation Tool: A Strategy to Support the Transfer of Expertise in Organizations. In, Alexandra Okada, Simon Buckingham Shum et Tony Sherborne Eds. *Knowledge Cartography. Mapping Techniques and Software Tools*. London: Springer-Verlag.

BASQUE J. & PUDELKO, B. (sous presse). Intersubjective Meaning-Making in Dyads Using Object-Typed Concept Mapping. In, P.L. Torres et R.C.V. Marriott Eds. *In Handbook of Research on Collaborative Learning Using Concept Mapping*: IGI Global.

BERNERS-LEE T. (1998). «Notation 3». En ligne. <http://www.w3.org/DesignIssues/Notation3.html>.

DAVIES J., FENSEL, D. & HARMELEN, F. V. (2003). *Towards The Semantic Web : Ontology-Driven Knowledge Management*: John Wiley & Sons.

DIETZ J. L. G. (2006). *Entreprise Ontology: Theory and Methodology*. Coll. «Computer Science». Berlin Heidelberg: Springer-Verlag.

GANGEMI A., GOMEZ-PEREZ, A., PRESUTTI, V. & SUAREZ-FIGUEROA. (2007). «Towards a Catalog of OWL-based Ontology Design Patterns». In *12 Conference of the Spanish Association for Artificial Intelligence* (12-16 November 2007): Springer.

GAŠEVIĆ D., DJURIĆ, D. & DEVEDŽIĆ, V. (2006). *Model Driven Architecture and Ontology Development*. New York, Inc.: Springer-Verlag.

GÓMEZ-PÉREZ A. (2004). Ontology Evaluation. In, R. Studer S. Staab Eds. *Handbook on Ontologies,* p. 251-274. New York: Springer.

GÓMEZ-PÉREZ A., FERNÁNDEZ-LÓPEZ, M. & CORCHO, O. (2003). *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, First edition. New York: Springer.

GRUBER T. R. (1995). «Toward principles for the design of ontologies used for knowledge sharing». *Int. J. Hum.-Comput. Stud.* vol. 43, no 5-6, p. 907-928.

HART A. (1986). *Knowledge acquisition for expert systems*. New York: McGraw-Hill.

HÉON M., PAQUETTE, G. & BASQUE, J. (2008). «Transformation de modèles semi-formels en ontologies selon les architectures conduites par les modèles». In *2èmes Journées Francophones sur les Ontologies* (1-2 Décembre).

HORROCKS I., BOLEY, H., TABET, S., GROSOF, B. & DEAN, M. (2004). «SWRL: A Semantic Web Rule Language Combining OWL and RuleML». W3C. En ligne. <http://www.w3.org/Submission/SWRL/>. Consulté le 29 mai.

OBERLE D. (2006). *Semantic Management of Middleware* Coll. «Semantic Web And Beyond Computing for Human Experience»: Springer US.

PAQUETTE G. (2002). *Modélisation des connaissances et des compétences : un langage graphique pour concevoir et apprendre*. Sainte-Foy: Presses de l'UQ.

PARIS S., LIPSON, M. Y. & WIXSON, K. K. (1983). «Becoming a Strategic Reader». *Contemporary Educational Psychology*. vol. 8, p. 293-316.

USCHOLD M. & GRUNINGER, M. (1996). «Ontologies: Principles, Methods and Applications». *Knowledge Engineering Review*. vol. 11, no 2, p. 93-155

Construction automatique d'ontologie à partir de bases de données relationnelles: application au médicament dans le domaine de la pharmacovigilance

Sonia Krivine[1], Jérôme Nobécourt[1], Lina Soualmia[1], Farid Cerbah[2], Catherine Duclos[1]

[1] Laboratoire d'informatique médicale & Bioinformatique, Université Paris 13
UFR de Santé Médecine et Biologie humaine, 74, rue Marcel Cachin 93017 Bobigny France
sonia.krivine@free.fr, j.nobecourt@smbh.univ-paris13.fr, lina.soualmia@univ-paris13.fr, catherine.duclos@avc.aphp.fr

[2] Dassault aviation, Département des Etudes Scientifiques
78, quai Marcel Dassault 92552 Saint-Cloud, France
farid.cerbah@dassault-aviation.fr

Résumé : Afin de construire une ontologie des médicaments en OWL dans le cadre d'un projet sur la pharmacovigilance, nous envisageons la possibilité de réutiliser les connaissances contenues dans une base de données. L'étude des outils permettant cette transition montre qu'ils ne sont pas entièrement satisfaisants. En effet, ils ne considèrent pas chaque enregistrement comme étant un concept et de ce fait les hiérarchies implicites contenues dans la base ne sont pas restituées. Nous proposons une extension de l'outil RDBToOnto afin d'empêcher l'instanciation des enregistrements de certaines tables de la base et de reproduire les hiérarchies de concepts.

Mots-clés : Ontologies, bases de données relationnelles, pharmacovigilance, médicament.

1. Introduction

Les médicaments sont responsables d'effets indésirables. Ces effets peuvent être connus et observés pendant les phases expérimentales de leur développement, ou encore apparaître lors de leur commercialisation et de leur usage à grande échelle. Les agences sanitaires ont mis en place un processus de pharmacovigilance qui permet de surveiller à l'échelle régionale, nationale et supra-nationale l'apparition d'effets indésirables, et d'avoir un système d'alerte identifiant les médicaments dont le rapport bénéfice/risque est trop défavorable. Le processus vise à colliger, documenter et imputer tout cas d'effet indésirable lié à l'utilisation de médicaments déclaré par un médecin notificateur. Les informations sur ces cas sont enregistrées dans des bases de données de pharmacovigilance sur lesquelles des algorithmes de détection du signal identifient une présence statistiquement atypique de couples {médicament, effet

indésirable}(Hauben, 2003). Cela permet par la suite de mener des enquêtes spécifiques sur ces médicaments afin d'en réévaluer le bénéfice/risque.

Henegar et al (Henegar, 2006) ont montré que la détection du signal était améliorée si certains cas étaient regroupés sur la base de la proximité conceptuelle des effets indésirables. Pour cela, ils ont développé une ontologie des effets indésirables qui permet de reclasser les concepts de la nomenclature MedDRA[1] utilisée notamment pour coder les effets indésirables dans les cas de pharmacovigilance. Par exemple, si le médicament M1 est déclaré dans un cas comme donnant une élévation des transaminases (marqueur d'inflammation hépatique) et dans un autre cas comme donnant une hépatite (atteinte du foie par un processus inflammatoire), les deux cas peuvent être regroupés comme associant M1 et une inflammation du foie.

Faisant suite à ces travaux, il a été émis, dans le cadre du projet VigiTermes[2], l'hypothèse que les médicaments pourraient bénéficier de modalités de regroupement. Par exemple, un cas décrivant un effet indésirable de type rhabdomyolyse après la prise de Zocor[3] et un autre décrivant le même effet indésirable après la prise de Lipanthyl[4] seraient regroupés car ils impliquent tous les deux des médicaments hypolipémiants. Cela permettrait ensuite de déduire que l'effet indésirable de rhabdomyolyse est lié à la classe des molécules hypolipémiantes.

Pour tester cette hypothèse, il faut disposer d'une ressource ontologique décrivant les médicaments à l'aide des propriétés susceptibles d'être à l'origine d'effets indésirables pour permettre un processus de classification des médicaments selon leur risque iatrogène[5]. En effet, un certain nombre de propriétés du médicament peuvent être identifiées comme ayant un lien potentiel avec la survenue d'effets indésirables (par exemple le fait qu'un médicament est fortement métabolisé par une enzyme hépatique peut conduire à un surdosage et donc à un effet indésirable chez un patient qui est déficitaire en cette enzyme, ou chez qui cette enzyme est inhibée par un autre médicament).

Plusieurs ontologies traitant du médicament sont disponibles comme Drug Ontology (Solomon, 1999), National Drug File – Reference Terminology (NDF-RT) (Chute, 2003), SNOMED-CT (Kim, 2001). Leurs limites sont soit un point de vue classificatoire orienté vers la prescription (description de la composition en ingrédients, appartenance à des classes pharmaco-thérapeutiques), ou lorsque des propriétés supplémentaires sont décrites, un abandon dans leur maintenance conduisant à leur obsolescence et à la non exhaustivité de leur contenu.

Les propriétés du médicament sont par ailleurs complètement décrites dans les banques de données commerciales dans un but de diffusion de l'information sur le médicament ou d'opérationnalisation dans les systèmes d'aide à la prescription. Ces

[1] Medical Dictionary for Regulatory Activities (http://meddramsso.com/MSSOWeb/index.htm)
[2] VigiTermes (ANR-07-TECSAN-026-04)
[3] Zocor est une simvastatine, molécule hypolipémiante de la famille des statines
[4] Lipanthyl est un fénofibrate, molécule hypolipémiante de la famille des fibrates
[5] Iatrogène: "se dit d'un trouble ou d'une maladie provoquée par un acte médical ou par les médicaments, même en l'absence d'erreur du médecin" (Petit Larousse)

banques ont l'intérêt d'avoir un contenu exhaustif pour décrire des propriétés liées à la iatrogénie, extrêmement structuré et maintenu à jour.

Notre objectif, dans le cadre du projet VigiTermes, est de construire de façon la plus automatisée possible une ressource ontologique permettant de classer les médicaments selon leur risque iatrogène. Il semble intéressant de partir d'une base de données structurée sur le médicament et d'outiller sa transition vers un formalisme ontologique (Gomes-Perez, 2004).

Nous présentons dans cet article tout d'abord le périmètre des propriétés liées au risque iatrogène du médicament et la sélection d'un sous-ensemble de tables de la base de données sur le médicament Thésorimed[6]. Nous décrivons ensuite les outils existants permettant de faire la transition entre le formalisme base de données relationnelle et le formalisme OWL et les utilisons sur l'extrait de la base Thésorimed. Confrontant les résultats à nos attentes, nous proposons une méthode pour enrichir ces outils de l'aspect important de la conservation des taxonomies. L'analyse des résultats de ce travail exploratoire est abordée en discussion.

2. Description du domaine

2.1. Modèle de la iatrogénie du médicament

Les effets indésirables peuvent se comprendre grâce à certaines propriétés qu'ont les médicaments. Un effet indésirable peut (a) être spécifique d'une substance active (*l'amoxicilline peut donner une dyschromie de l'émail dentaire*) ou d'une substance auxiliaire entrant dans la composition du médicament mais n'ayant pas de propriétés thérapeutiques (*l'aspartam contenu dans une suspension buvable d'amoxicilline peut être dangereux chez les sujets atteints de phénylcétonurie*); (b) être un effet de classe se produisant avec tous les médicaments appartenant à une classe chimique ou pharmacologique (*les amidinopénicillines peuvent donner des diarrhées ; les antiulcéreux inhibiteurs de la pompe à proton peuvent donner des douleurs abdominales*); (c) être la manifestation d'une interaction générique avec un organe ou d'une interaction expliquée au niveau moléculaire avec un site effecteur recherché ou secondaire (*le captopril dont le mécanisme d'action est d'inhiber de l'enzyme de conversion rénine angiotensine pour diminuer la pression artérielle risque d'entraîner une hypotension artérielle directement liée à son mécanisme d'action mais aussi une toux dont le mécanisme de survenue est inconnu*) ; (d) survenir à dose thérapeutique ; (e) être la manifestation d'un surdosage se traduisant par une majoration de l'ensemble des effets du médicament et pouvant conduire à une toxicité du médicament.

L'effet indésirable qui survient à dose thérapeutique peut apparaître dans certains contextes (a) patients (physiologique, pathologique, génétique ou allergique) ; (b) de doses (faible, forte) ; (c) d'administration (forme, voie, débit) ; (d) d'exposition

[6] Thesorimed est une base de données sur le médicament développée par le GIE SIPS (Système d'Information sur les Produits de Santé), (http://www.giesips.org/)

(administration prolongée). Le surdosage peut être lié à (a) une posologie erronée (dose par prise, dose par jour, fréquence d'administration, durée) ;(b) une altération des paramètres pharmacocinétiques[7] en raison d'interactions médicamenteuses, d'un contexte patient particulier ; (c) une altération de la pharmacodynamie[8] en raison d'une interaction médicamenteuse, d'un contexte patient particulier. La figure 1 illustre le modèle de la iatrogénie du médicament.

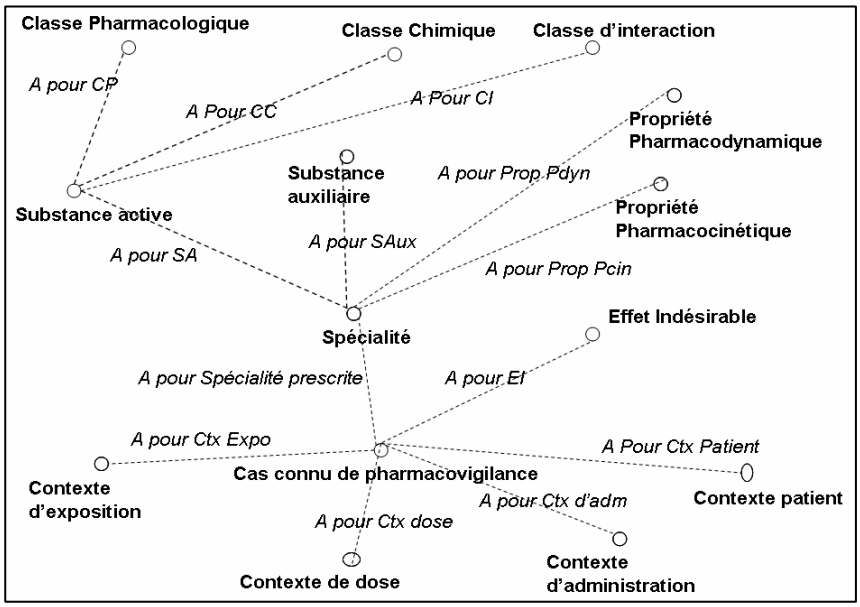


Fig. 1 - Modèle de la iatrogénie du médicament pour documenter un cas de pharmacovigilance (les lignes en pointillés représentent des relations entre concepts).

2.2 – Etude de Thésorimed et constitution de la base de données sur la iatrogénie du médicament

La base de données sur le médicament Thésorimed contient les informations illustrant les propriétés précédemment décrites (figure 1). Ces propriétés sont représentées soit en utilisant une terminologie contrôlée (termes d'indexation de Thésorimed), soit sous forme de texte. Nous nous intéresserons ici aux éléments décrits par une terminologie contrôlée (figure 2), à savoir : (i) les substances actives, auxiliaires, classes chimiques, classes pharmacologiques donnant des effets indésirables à dose thérapeutique ou en cas de surdosage, (ii) les hiérarchies de classes chimiques, pharmacologiques et d'effets indésirables.

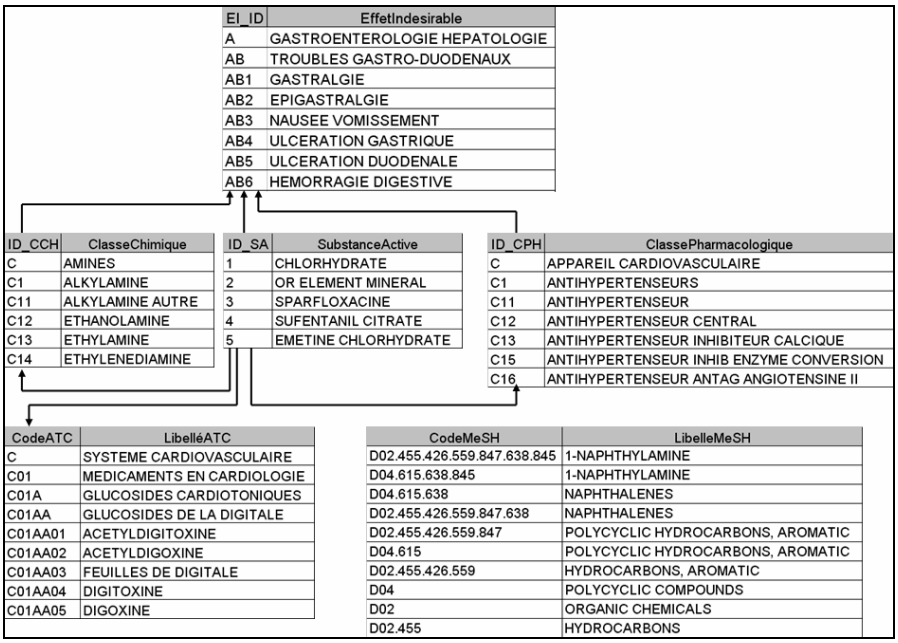
[7] La pharmacocinétique étudie le devenir d'un médicament dans l'organisme
[8] La pharmacodynamie décrit ce que le médicament fait à l'organisme

A cela il nous semble intéressant d'adjoindre des hiérarchies comme le MeSH[9] pour enrichir la description chimique ou comme l'ATC[10] pour la description pharmacothérapeutique des médicaments.

Les hiérarchies dans ces tables sont exprimées de trois façons différentes. La hiérarchie de type ATC est représentée à l'aide d'un code alphanumérique sans séparateur qui s'étend de père en fils et pour lequel un fils a un seul père. Chaque caractère du code représente un niveau de la hiérarchie. Ce codage est retrouvé dans Thésorimed pour les classes chimiques, pharmacologiques et d'effets indésirables.

La hiérarchie de type MeSH reprend le principe d'un code alphanumérique avec séparateur qui s'étend de père en fils. Elle a la caractéristique particulière qu'un fils peut avoir plusieurs pères selon le principe de l'héritage multiple.

Dans la hiérarchie de type Thésorimed le lien de subsomption est exprimé grâce la relation entre la table des substances et celle des classes chimiques ou des classes pharmacologiques.



EI_ID	EffetIndesirable
A	GASTROENTEROLOGIE HEPATOLOGIE
AB	TROUBLES GASTRO-DUODENAUX
AB1	GASTRALGIE
AB2	EPIGASTRALGIE
AB3	NAUSEE VOMISSEMENT
AB4	ULCERATION GASTRIQUE
AB5	ULCERATION DUODENALE
AB6	HEMORRAGIE DIGESTIVE

ID_CCH	ClasseChimique
C	AMINES
C1	ALKYLAMINE
C11	ALKYLAMINE AUTRE
C12	ETHANOLAMINE
C13	ETHYLAMINE
C14	ETHYLENEDIAMINE

ID_SA	SubstanceActive
1	CHLORHYDRATE
2	OR ELEMENT MINERAL
3	SPARFLOXACINE
4	SUFENTANIL CITRATE
5	EMETINE CHLORHYDRATE

ID_CPH	ClassePharmacologique
C	APPAREIL CARDIOVASCULAIRE
C1	ANTIHYPERTENSEURS
C11	ANTIHYPERTENSEUR
C12	ANTIHYPERTENSEUR CENTRAL
C13	ANTIHYPERTENSEUR INHIBITEUR CALCIQUE
C15	ANTIHYPERTENSEUR INHIB ENZYME CONVERSION
C16	ANTIHYPERTENSEUR ANTAG ANGIOTENSINE II

CodeATC	LibelléATC
C	SYSTEME CARDIOVASCULAIRE
C01	MEDICAMENTS EN CARDIOLOGIE
C01A	GLUCOSIDES CARDIOTONIQUES
C01AA	GLUCOSIDES DE LA DIGITALE
C01AA01	ACETYLDIGITOXINE
C01AA02	ACETYLDIGOXINE
C01AA03	FEUILLES DE DIGITALE
C01AA04	DIGITOXINE
C01AA05	DIGOXINE

CodeMeSH	LibelleMeSH
D02.455.426.559.847.638.845	1-NAPHTHYLAMINE
D04.615.638.845	1-NAPHTHYLAMINE
D04.615.638	NAPHTHALENES
D02.455.426.559.847.638	NAPHTHALENES
D02.455.426.559.847	POLYCYCLIC HYDROCARBONS, AROMATIC
D04.615	POLYCYCLIC HYDROCARBONS, AROMATIC
D02.455.426.559	HYDROCARBONS, AROMATIC
D04	POLYCYCLIC COMPOUNDS
D02	ORGANIC CHEMICALS
D02.455	HYDROCARBONS

Fig. 2 - Tables et champs retenus dans la banque Thésorimed pour représenter les éléments initiaux du modèle de la iatrogénie des médicaments et présentation d'extraits de leur contenu, les tables reliées à la table effet indésirable servent à la construction de cas connus de pharmacovigilance

[9] Medical Subject Headings (http://www.nlm.nih.gov/mesh/)

[10] Classification Anatomique, Thérapeutique et Chimique (http://www.whocc.no/atcddd/)

3. Outils de transition entre bases de données et OWL

DataMaster[11], KAON2[12], et RDBToOnto[13] sont trois outils permettant de faire la transition entre les formalismes base de données et OWL. Pour chacun de ces outils, nous avons étudié comment se fait le reverse engineering pour créer un concept, une relation, un héritage, une propriété, une instance.

3.1 Description des outils

3.1.1 DataMaster

DataMaster est un plug-in de Protégé spécialisé dans l'import des structures et des données de bases de données relationnelles (Nyulas, 2007) qui propose une méthode d'import des tables de la base de données relationnelle comme des concepts OWL.

Avec DataMaster, le paradigme relationnel est conservé puisque les données extraites de la base sont simplement inscrites dans l'ontologie générique *Relational.OWL* [de Laborda *et al.* 2005]. Cet outil n'est pas dimensionné pour effectuer une restructuration profonde du modèle et des données sources, il permet néanmoins de migrer les données vers une description dans une ontologie.

En utilisant DataMaster avec notre base de données, chaque table a été convertie en un concept et chaque ligne de la table a été convertie en une instance du concept correspondant. Les valeurs des attributs ont été instanciées avec les valeurs des champs correspondants de la table.

3.2.2 KAON2

KAON2 est une plateforme de construction d'ontologies développée à l'université de Karlsruhe.

Cette plateforme est en fait équipée d'une fonctionnalité de mise en correspondance entre bases de données et ontologies, dans la lignée des outils et formalismes de mapping, tels que D2RQ [Bizer 2003] et R2O [Barrasa *et al.* 2004]. Dans cette perspective, il ne s'agit pas d'automatiser la construction du modèle de classes et de propriétés. L'objectif visé est d'offrir des moyens déclaratifs pour décrire des procédés d'instanciation à partir de bases relationnelles d'ontologies prédéfinies manuellement. KAON2 permet de fournir une « vue » sous forme d'ontologie (ce que ses concepteurs appellent une « ontologie virtuelle »), alimentée à la volée par des instances extraites d'une base de données.

La documentation très succincte fournie sur le site Web de la plateforme[14] et dans les exemples présents dans la distribution du logiciel laisse prévoir un type de résultat très proche de celui de DataMaster.

[11] http://protegewiki.stanford.edu/index.php/DataMaster
[12] http://kaon2.semanticweb.org/
[13] http://www.tao-project.eu/researchanddevelopment/demosanddownloads/RDBToOnto.html
[14] http://kaon2.semanticweb.org/#documentation

Les procédés de mapping proposés présupposent l'utilisation de structures ontologiques simples, sans axiomes, quasi-calquées sur les schémas relationnels sources.

3.2.3 RDBToOnto

RDBToOnto [Cerbah 2008a] est un outil dédié à la conversion de bases de données en ontologies, développé dans le cadre du projet européen TAO (Transitioning Applications to Ontologies).11.
RDBToOnto permet à l'utilisateur un paramétrage du projet de conversion à savoir :

- la définition interactive de contraintes locales, attachées aux tables d'entrée, pour orienter le processus de transformation. En particulier, ces contraintes permettent de spécifier des motifs de nommage des classes et des instances (en combinant des valeurs d'attributs pour former les noms) ou encore d'exclure des tables et surtout des colonnes lors de la conversion,
- Le choix d'un convertisseur (implémentation d'une méthode de transformation), parmi ceux prédéfinis, ou le recours à une nouvelle méthode obtenue en spécialisant un convertisseur déjà intégré dans la plateforme.

Le développement de RDBToOnto est orienté vers la reconnaissance de structures de catégorisation en analysant conjointement le schéma de la base et les données stockées. Ainsi, le convertisseur RTAXON, central dans cet outil, implémente une méthode générique de reconnaissance des attributs de catégorisation, se basant d'une part sur le nom de l'attribut, d'autre part sur la redondance dans les extensions des attributs à l'aide d'une méthode basée sur l'entropie (Cerbah, 2008b).

Le paramétrage de l'outil en « Catégorisation forcée» permet déjà d'obtenir un second niveau de profondeur hiérarchique, qu'aucun des autres outils examinés ne propose et témoigne d'une orientation dans la conception de l'outil vers l'extraction de relations de subsomption, particulièrement intéressante pour notre projet.

Chaque table est convertie en une classe et chaque ligne de la table donne lieu à la création d'une sous-classe.

3.2 Limites des outils existants

Notre étude a révélé que les outils existants partagent certaines limitations majeures.

Une conception orientée vers le peuplement d'ontologies c'est-à-dire l'instanciation de concepts représentés par les tables de la base de données plutôt que vers l'apprentissage d'une structure conceptuelle (avec hiérarchies multi niveaux, propriétés, héritage multiple, etc.) à partir de bases de données. Ceci à l'exception de RDBToOnto, dont la méthode de conversion RTAXON est orientée vers l'extraction de relations d'héritage contenues dans certains champs de la base de données.

La quasi-absence de paramétrage de l'importation par l'utilisateur : à l'exception de RDBToOnto, ces outils s'intègrent au sein d'un logiciel (ou d'une

suite logicielle) d'édition d'ontologies, sous la forme d'une fonctionnalité qui laisse peu de possibilités de paramétrage de la conversion à l'utilisateur comme par exemple l'exclusion de certaines tables, des traitements spécifiques à certaines tables, le choix de conventions sur le nommage des concepts et propriétés extraits.

L'héritage multiple est exclu.

L'extraction d'une structure ontologique « plate » : la structure obtenue reste très proche du schéma de la base de données. Toutes les tables subissent un traitement identique et rudimentaire : une table de la base source est convertie en un concept et chaque ligne de cette table est convertie en instance de ce même concept, les valeurs des colonnes étant reportées sous forme de valeurs d'attributs. Les structures hiérarchiques extraites sont de ce fait peu profondes : deux niveaux de profondeur maximum obtenus avec l'option de catégorisation de RDBToOnto.

4. Enrichissement de RDBToOnto

Nous avons choisi de procéder à une adaptation de RDBToOnto en spécialisant le convertisseur RTAXON de manière à en contourner les limitations exposées précédemment (figure 3).

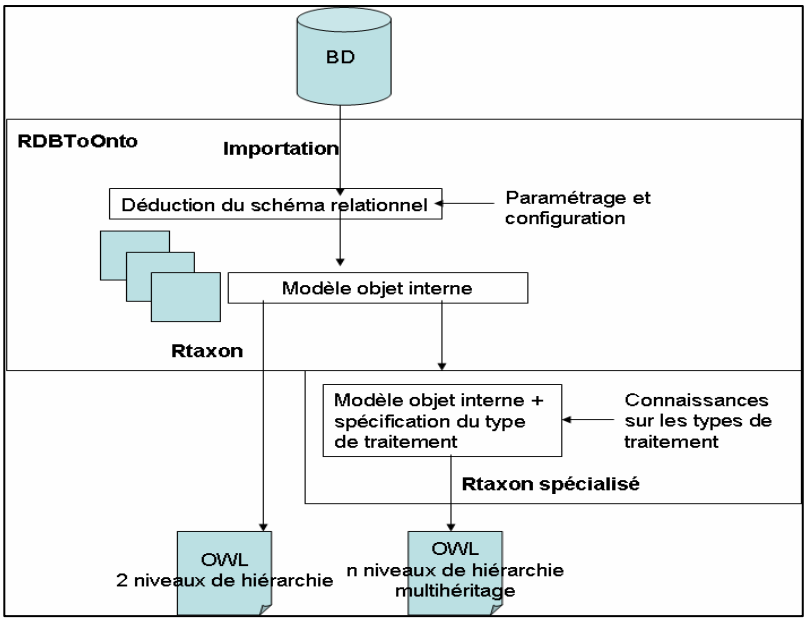


Fig. 3 – Méthode d'enrichissement de RDBToOnto

Le développement de quelques composants Java supplémentaires est requis afin que cet outil :

- génère des hiérarchies de concepts d'une profondeur supérieure à 2 dans le cas de hiérarchies de type ATC

- autorise le multihéritage dans le cas des tables de type MeSH,
- prévoit l'import spécifique de tables contenant des relations entre concepts, par exemple « SubstanceActive (aPourClasseChimique) Classe_Chimique ».

4.1 – Traitement de type ATC pour la construction des taxonomies de profondeur fixe ou variable, sans multihéritage

L'utilisateur identifie une table comme étant hiérarchique et l'attribut qui contient l'information sur les niveaux hiérarchiques. Cette information doit être décodée pour en extraire la taxinomie. Dans notre cas, l'information sur ces niveaux hiérarchiques est représentée par un code alphanumérique qui s'étend de père en fils. La hiérarchie peut se déduire automatique par parcours récursif du code.

Dans ces tables, les libellés ne représentent pas le terme préféré pour nommer un concept (par exemple le libellé « associations » est utilisé quelque que soit le type d'association (plusieurs antibiotiques ou plusieurs antidiabétiques)), le sens de l'association est explicité lorsque la hiérarchie est déployée (on déduit que « association » a le sens « association d'antibiotiques » parce qu'elle est associée au code J01CA20, qui est dans la hiérarchie des antibiotiques). Pour nommer un concept, nous avons donc choisi d'associer le code et le libellé du code (figure 4).

CodeATC	LibelléATC
C	SYSTEME CARDIOVASCULAIRE
C01	MEDICAMENTS EN CARDIOLOGIE
C01A	GLUCOSIDES CARDIOTONIQUES
C01AA	GLUCOSIDES DE LA DIGITALE
C01AA01	ACETYLDIGITOXINE
C01AA02	ACETYLDIGOXINE
C01AA03	FEUILLES DE DIGITALE
C01AA04	DIGITOXINE
C01AA05	DIGOXINE

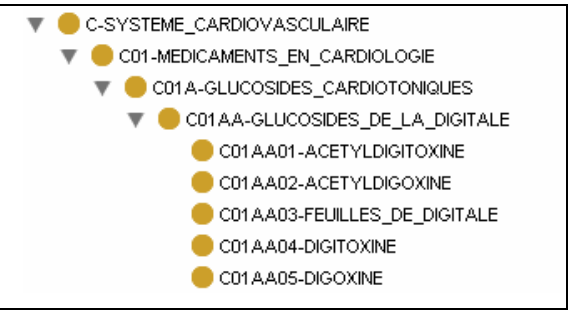
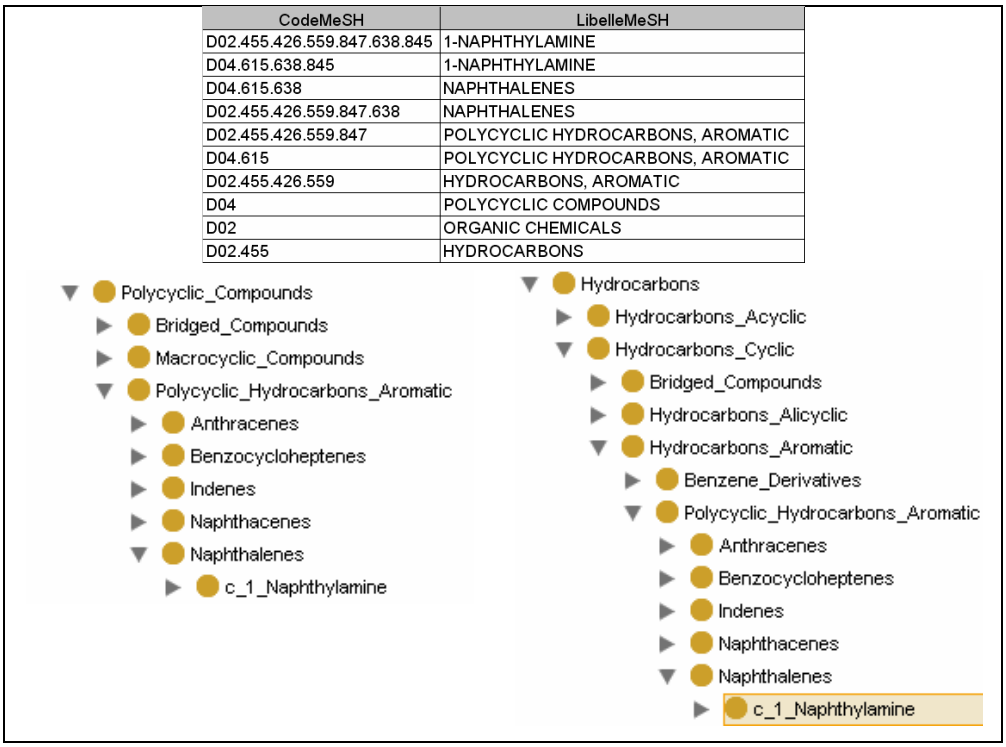


Fig. 4 – Exemple de transformation de type ATC à partir de la table de « classification ATC » et visualisation de la taxinomie résultante dans Protégé

4.2 Traitement de type MeSH pour la construction des taxonomies avec multihéritage

De la même manière qu'en 4.1, la table doit être identifiée comme étant hiérarchique, un attribut doit contenir une information sur un des niveaux hiérarchiques (code) et un autre attribut doit comporter le terme. Il existe donc plusieurs lignes ayant le même terme mais des codes différents. Le terme est pris comme label du concept, ses pères sont calculés en fonction de chacun de ses codes (figure 5).



CodeMeSH	LibelleMeSH
D02.455.426.559.847.638.845	1-NAPHTHYLAMINE
D04.615.638.845	1-NAPHTHYLAMINE
D04.615.638	NAPHTHALENES
D02.455.426.559.847.638	NAPHTHALENES
D02.455.426.559.847	POLYCYCLIC HYDROCARBONS, AROMATIC
D04.615	POLYCYCLIC HYDROCARBONS, AROMATIC
D02.455.426.559	HYDROCARBONS, AROMATIC
D04	POLYCYCLIC COMPOUNDS
D02	ORGANIC CHEMICALS
D02.455	HYDROCARBONS

Fig. 5 – Exemple de transformation de type MeSH à partir de la table de « classification MeSH » et visualisation du multihéritage dans Protégé

4.4 Traitement de type Thesorimed pour la construction des relations et des hiérarchies contenues dans des tables de relation binaire entre substance et classification

L'utilisateur identifie trois tables : la table de classification, la table de substance et la table de jointure entre substance et classification. A partir de la table de classification le traitement décrit en 4.1 est réalisé, puis on décrit des concepts de regroupement de substances. Ces concepts sont définis en utilisant un rôle dont le domaine est la substance et le co-domaine le code de classification.

Pour chaqque concept de substance, nous définissons une propriété « aPourClassification » grâce à la table de jointure.

A l'issue de ces descriptions, nous utilisons le classifieur Fact++ pour reclasser les substances dans les regroupements de substances.

Nous avons pour l'instant réalisée cette classification sur 5234 concepts (concepts ATC et concepts substances), et 60 concepts de regroupements. La figure 6 illustre ce traitement.

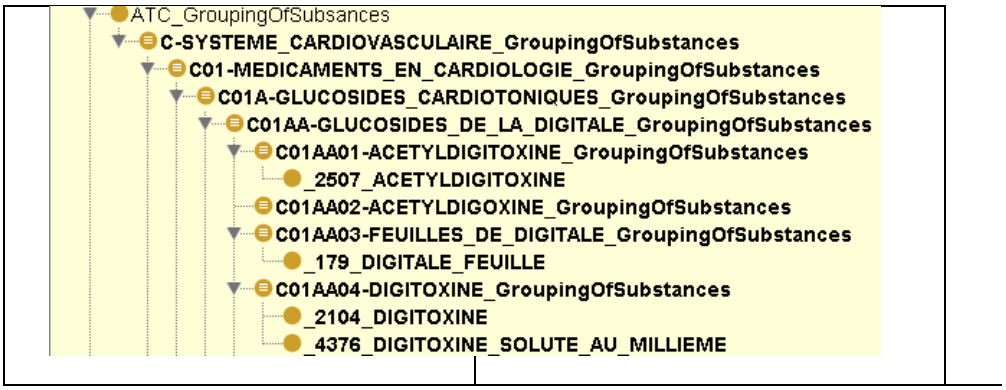


Fig. 6 – Exemple de transformation de type Thesorimed à partir des tables de « classification ATC », « Substances » et « SubstanceApourATC » et visualisation de la hiérarchie inférée par Fact++ dans Protégé

5. Discussion et conclusion

Les outils disponibles pour faire la transition entre les bases de données relationnelles et OWL supportent principalement le postulat que les enregistrements trouvés dans les bases sont des instances du concept constitué par la table. Dans le cas des bases de données sur le médicament, il s'avère qu'un certain nombre de tables contiennent des termes normalisés servant à la description des propriétés du médicament et qu'ils peuvent être considérés non comme des instances de concepts mais comme des concepts. Ces concepts entretiennent en particulier des relations de subsumption qu'il faut pouvoir rendre dans l'ontologie.

La création de notre ontologie est guidée à la fois par la connaissance sur les données mais également par l'usage envisagé pour cette ontologie. Il existe d'autres approches de reverse engineering qui intègrent l'usage de la base de données pour augmenter la spécificité des traitements (Benslimane 2007).

Nous avons distingué trois types de traitement à appliquer à notre base de données sur le médicament pour : (a) reproduire une hiérarchie contenue dans un code par parcours récursif du code, (b) créer des liens de hiérarchie multiple, (c) créer des associations décrites par la relation entre deux entités.

Cela nous a conduit, dans une première approche exploratoire, à spécialiser l'outil RDBToOnto afin de proposer à l'utilisateur un nouveau convertisseur lui permettant d'indiquer des propriétés des tables pour réaliser des traitements spécifiques aux différents types de hiérarchies.

Cette spécialisation de l'outil remet en cause le principe de généricité de RDBToOnto et nous amène à reconsidérer le problème général de la construction de nouveaux de points de vue sur l'utilisation des connaissances sur le médicament tout en réutilisant le contenu de bases de données commerciales sur le médicament. Cela

devrait nous orienter vers le développement d'outils dédiés spécifiquement à la réutilisation d'une base médicaments donnée.

Le fait de s'adosser à une base de données mise à jour par des éditeurs nous permet d'envisager une maintenance plus aisée de l'ontologie grâce à des traitements les plus automatisés possibles réalisés sur cette base.

6. Références

BARRASA J, CORCHO O, GÓMEZ-PÉREZ A. (2004) "R2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language". *Second Workshop on Semantic Web and Databases (SWDB2004)*. Toronto, Canada.

BENSLIMANE S.M, BENSLIMANE D, MALKI M, AMGHAR Y, GARGOURI F.(2006) Construction d'une ontologie à partir d'une base de données relationnelle: approche dirigée par l'analyse des formulaires HTML. INFORSID'06, Hammamet, Tunisie pp. 991-1010.

BIZER, C.(2003) "D2R MAP - A Database to RDF Mapping Language", WWW03, Budapest.

CERBAH, F. (2008a) Learning Highly Structured Semantic Repositories from Relational Databases - RDBtoOnto Tool, in *Proceedings of the 5th European Semantic Web Conference (ESWC 2008),* Tenerife, Spain, June, 2008.

CERBAH, F. (2008b) Mining the Content of Relational Databases to Learn Ontologies with deeper Taxonomies. *Proceeding of IEEE/WIC/ACM International Joint Conference on Web Intelligence (WI'08) and Intelligent Agent Technology (IAT'08)*, Sydney, Australia, 9-12.

CHUTE ET AL (2003) Integrating pharmacokinetics knowledge into a drug ontology: as an extension to support pharmacogenomics. AMIA Annu Symp Proc. 170-4.

DE LABORDA, C.P., CONRAD, S. (2005) Relational.OWL: a data and schema representation format based on OWL. *In: APCCM '05: Proc. of the 2nd Asia-Pacific conference on Conceptual modelling, Darlinghurst*, Australian Computer Society, Inc.

GOMEZ-PEREZ A., FERNANDEZ-LOPEZ M., CORCHO O. (2004) Ontological engineering, with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web, Springer-Verlag London.

HAUBEN M, ZHOU X.(2003) Quantitative methods in pharmacovigilance: focus on signal detection. Drug Saf. ;26(3):159-86.

HENEGAR C, BOUSQUET C, LILLO-LE LOUËT A, DEGOULET P, JAULENT MC (2006). Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Comput Biol Med. 36(7-8):748-67.

KIM JM, FROSDICK P (2001).Description of a drug hierarchy in a concept-based reference terminology. Proc AMIA Symp:314-8.

NYULAS C. O'CONNOR, M. TU, S. (2007), DataMaster – a Plug-in for Importing Schemas and Data from Relational Databases into Protégé, (http://protege.stanford.edu/conference/2007/presentations/10.01_Nyulas.pdf).

SOLOMON ET AL (1999).A reference terminology for drugs. Proc AMIA Symp. 152-5.

Construction automatique d'ontologies à partir de spécifications de bases de données

Mouna Kamel, Nathalie Aussenac-Gilles

Laboratoire IRIT, Université Paul Sabatier de Toulouse
{kamel, aussenac}@irit.fr

Résumé : Les méthodes classiques de construction automatiques d'ontologies à partir de textes exploitent le texte proprement dit. Nous étendons ces approches en prenant en compte la structure du texte, élément porteur d'information. Pour cela, nous nous basons sur des documents de spécifications de bases de données au format XML, pour lesquels le découpage structurel du texte correspond à une caractérisation sémantique de son contenu. L'idée est de tirer profit à la fois de la structure du texte et du texte rédigé. La méthode proposée consiste à utiliser la sémantique des balises et à caractériser leurs relations pour définir des règles de création de concepts et de relations sémantiques. Un noyau d'ontologie a été ainsi construit automatiquement à l'aide de ces règles, noyau ensuite enrichi par l'exploitation du texte en langage naturel à l'aide de patrons lexico-syntaxiques définis. Règles et patrons ont été implémentés sous Gate.

Mots-clés : Ingénierie des connaissances, extraction de relations, ontologie, documents structurés, traitement automatique de textes.
Communication appliquée

1 Introduction

Les méthodes de construction d'ontologies à partir de textes privilégient souvent l'analyse du texte proprement dit, que ce soit selon une approche statistique ou linguistique (Nédellec et Nazarenko, 2003), (Aussenac et al., 2008), (Maedche, 2002), (Buitelaar et al., 2005). La plupart de ces travaux montrent la nécessité d'intégrer différents outils de TAL, et soulignent la complémentarité entre identification de concepts et extraction de relations. Notre objectif est d'étendre ces approches classiques en prenant en compte la structure explicite des textes, lorsque cette structure caractérise la sémantique d'unités textuelles repérées et leur hiérarchisation. C'est le cas notamment des documents de spécifications de bases de données, riches en descriptions de concepts. Ce type de document nous permet alors d'identifier i) les concepts spécifiés, ii) les relations entre ces concepts à l'aide de la structure hiérarchique du document et iii) de nouveaux concepts et/ou de nouvelles relations en exploitant le texte rédigé. Le format XML est un bon format de représentation de tels documents : il intègre les notions de description sémantique de la structure d'un document (le même fichier situe le texte en langage naturel au sein de cette structure) et de structure hiérarchique. En exploitant donc la sémantique véhiculée par les balises, la structure hiérarchique du document et l'information contenue entre les

balises, la complémentarité entre identification de concepts et extraction de relations est assurée, à l'instar des approches classiques.
Comme XML est un langage de balisage non prédéfini, nous nous plaçons dans une perspective d'Ingénierie des Connaissances. Nous nous limitons à un domaine spécifique et nous tenons compte à la fois de la sémantique associée aux balises et à leurs relations, et de connaissances d'arrière-plan. Une correspondance peut alors être établie entre les fragments de textes balisés et des éléments d'ontologie.
La méthode que nous présentons consiste à extraire en priorité et à l'aide de cette correspondance (exprimée par des règles) les concepts et les relations sémantiques, étape essentielle pour définir un noyau de l'ontologie, puis à enrichir ce noyau en exploitant le texte en langage naturel présent dans le document. Nous rappelons d'abord les différentes recherches sur l'identification des relations sémantiques à partir du contenu du texte ou de leur structure (partie 2). La partie 3 décrit comment notre méthode conjugue certaines de ces approches pour extraire des connaissances à partir de la structure du texte et de son contenu. La partie 4 présente la mise en œuvre de notre méthode au sein du projet GEONTO, dont l'objet est de construire automatiquement des ontologies à partir de spécifications de bases de données géographiques. La 5ème partie donne une évaluation de notre méthode dans ce contexte. Nous dressons enfin le bilan actuel de nos travaux au sein du projet (partie 6) et présentons les perspectives pour les améliorer.

2 Méthodes d'identification de relations sémantiques

L'identification de relations est utile pour construire automatiquement une ontologie ou pour l'enrichir par des relations entre instances. Deux familles de techniques extraient des relations sémantiques à partir de textes : les approches statistiques et les approches linguistiques. Les approches statistiques consistent à étudier les termes co-occurrents et la similarité entre leurs contextes syntaxiques (Hindle, 1990), (Grefenstette, 1994), à prédire les relations à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko, 2007) ou de techniques de Text Mining (Grcar et al., 2007), ou encore à inférer des connaissances à l'aide d'algorithmes d'apprentissage (Guiliano et al., 2006). Ces méthodes sont efficaces, mais n'identifient pas toujours la sémantique de la relation. L'approche linguistique fait appel à des analyses syntaxiques ou des calculs de dépendance pour identifier les relations argumentatives (sujet, verbe, objet) (Jacquemin, 1997), (Bourigault, 2002), ou définit des patrons lexico-syntaxiques pour reconnaître les marques linguistiques des relations sémantiques (Aussenac et Seguela, 2000). Ainsi la sémantique des relations est bien identifiée, mais la variabilité de leur sémantique et de leur expression en corpus oblige à multiplier les patrons et rend l'approche coûteuse.
Ces techniques s'appliquent au niveau interne de la phrase, alors que d'autres études ont pour niveau d'analyse le texte lui-même. L'objectif est alors assez différent : il ne s'agit plus de trouver des relations entre concepts, mais des relations sémantiques plus diffuses entre les différentes unités textuelles repérées. Ces liens peuvent être décelés soit à l'aide de marqueurs linguistiques (Asher et al., 2001), soit en exploitant la

structure matérielle du texte, (Virbel & Luc, 2001) ayant montré que la matérialité d'un texte participe à son sens, soit en combinant les marqueurs linguistiques et la structure du texte (Charolles, 1997).
Les traitements statistiques et linguistiques sont largement utilisés pour la construction automatique d'ontologies (Buitelaar et Ciminao, 2005), (Maedche, 2002), alors qu'il n'existe à notre connaissance que très peu de travaux qui aient exploité les relations du discours ou la mise en forme matérielle d'un texte dans ce but (Laurens, 2006).
L'idée que nous développons ici est de combiner une approche basée sur la structure du document à une approche linguistique pour tirer profit à la fois de la structure du texte et du texte lui-même. Notre objectif est d'élaborer des ontologies plus riches que ne pourraient fournir ces approches prises indépendamment. Nous présentons dans le paragraphe suivant notre méthode pour l'analyse d'un document de type spécifications de bases de données (au format XML) en vue de la conception d'une ontologie.

3 Méthodologie

La méthode proposée pour la construction d'ontologies consiste à exploiter un document de spécifications de bases de données au format XML en associant une analyse de la structure du texte (sémantique des balises et de leur hiérarchie) à une analyse linguistique du texte présent entre les balises (exploitation du langage naturel, de signes typographiques). Ces deux traitements sont réalisés indépendamment, sauf dans quelques règles où la détection de concepts ou de relations découle à la fois de la sémantique d'une balise et du texte qu'elle marque.

3 .1 Exploitation des balises

Le langage de balises fournit à la fois une description du texte et, par le processus d'imbrication des balises, des liens existant entre les unités textuelles balisées. Dans le cas où ces unités textuelles sont simples et ne renvoient chacune qu'à un seul concept, l'analyse des balises et de leur imbrication hiérarchique permet d'identifier des concepts et des relations sémantiques. La règle générique mise en œuvre pour l'identification des relations est la suivante *(">>" est l'opérateur de recouvrement)* :

si	*A, B et C sont des balises, B et C introduisant des concepts*
	A >> B
	A >> C
alors	*Il existe une relation conceptuelle R_{BC} entre B et C.*

Une étude préalable de la nature des balises et de leur hiérarchisation a pour but de déterminer les balises B et C qui introduisent des concepts et la sémantique des relations conceptuelles R_{BC} (hiérarchiques, méronymiques, et autres) associées aux relations entre balises B et C. Cette étude permet de définir autant de règles du format de la règle générique que nécessaire. L'analyse automatisée du corpus à l'aide de ces règles fournit un noyau d'ontologie.

3 .2 Exploitation du texte en langage naturel

Le corps du document XML correspond au texte en langage naturel et peut contenir de l'information intéressante à exploiter pour enrichir l'ontologie obtenue à l'issue de l'exploitation de sa structure (section 3.1). Selon C. Barrière (Barrière et Agbado, 2006) et Hearst (Hearst, 1992), un des moyens de qualifier "des contextes riches en connaissances" est qu'ils contiennent des marques linguistiques de relations sémantiques. Nous avons choisi d'utiliser des patrons lexico-syntaxiques pour repérer des relations sémantiques (Auger et Barrière, 2008). Un patron lexico-syntaxique décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte répondant à ce format. Dans le cas particulier de la recherche de relations, le patron caractérise un ensemble de formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes (Rebeyrolle & Tanguy, 2000). L'application de tels patrons nécessite de traiter préalablement le texte en appliquant différents outils du TAL (tokenizer, lemmatiseur, analyseur syntaxique, etc.). Les patrons exploitent les étiquettes morpho-syntaxiques ou sémantiques attribuées par ces logiciels. Ainsi, la forme des patrons dépend à la fois du logiciel de définition et de projection des patrons, et des analyses et étiquetages effectués sur les textes.
Notre approche complète l'exploitation de la structure des documents (§ 3.1) en identifiant, à l'aide de patrons préalablement définis, de nouveaux concepts, des relations, voire des propriétés dans les parties rédigées du document. Elle peut être reproduite sur tout document de ce genre, moyennant la définition de règles d'interprétation des balises et des patrons.

4 Cadre d'application : projet GEONTO

Au sein du projet GEONTO[1], un des partenaires dispose de bases de données géographiques hétérogènes et a pour objectif l'interopérabilité de ces bases. Pour cela, le projet prévoit de fournir une ontologie par base de données, et d'aligner les ontologies obtenues en vue de produire une ontologie de référence. Les ontologies produites seront issues des spécifications de ces bases de données, et non de leurs schémas comme dans les travaux de (Tirmizi et al., 2008), (Gardarin et al., 2008)

4.1 Description des données

L'expérimentation décrite ici porte sur la base de données BDTopo qui sert de référence pour la localisation de l'information relative aux problématiques d'aménagement, d'environnement ou d'urbanisme. Les spécifications de cette base de données, disponibles au format WORD utilisant un style pour chaque type d'information, ont été automatiquement traduites en XML : c'est ce document qui

[1] Projet ANR-07-MDCO-005, http://www.lri.fr/geonto : collaboration entre le COGIT, le LRI (Université paris Sud), le LIUPPA (Université de Pau) et l'IRIT (Université de Toulouse)

servira à la construction de l'ontologie. Un extrait du document de spécification de la base BDTopo (classe Tronçon de chemin) est présenté figure 1, et la portion du fichier XML correspondant à ces spécifications est donnée figure 2.

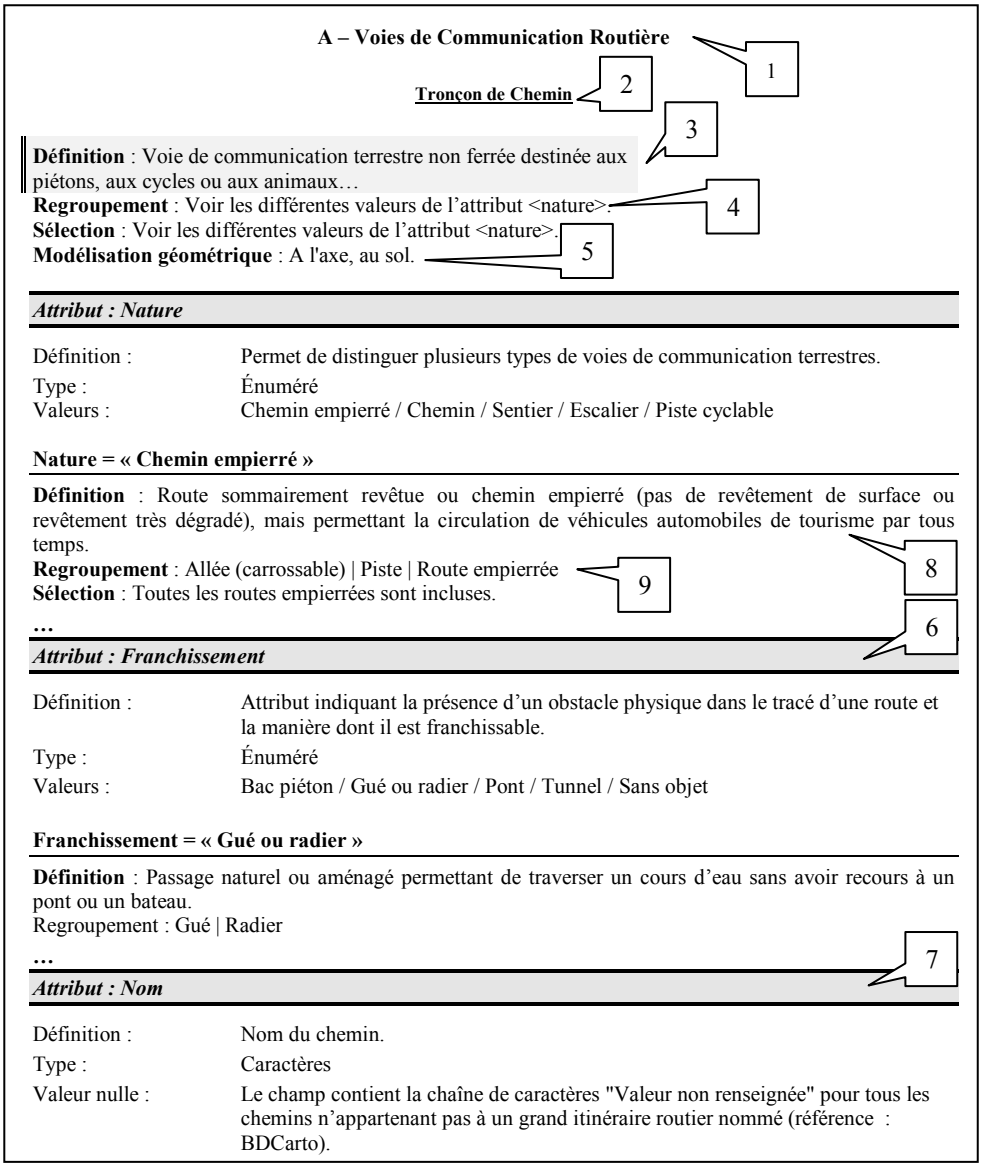


Figure 1 : extrait des spécifications concernant la classe **Tronçon de chemin**.

Sur la figure 1, les classes d'objets (2) sont réparties dans 9 domaines d'information (1). Les objets d'une même classe mentionnés dans le champ *Regroupement* (4) partagent une même définition (3), un même type de géométrie (5) et une même liste

d'attributs. Ces attributs supportent des informations à caractère qualitatif (liste d'objets) (6) ou quantitatif (attribut de type non énuméré) (7). Chaque valeur d'attribut a sa propre définition (8) et peut représenter une liste d'objets (9).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<document> <domaine>
  <nom_domaine>A - Voies de communication routière</nom_domaine>
  <classe>
    <nom_classe>Tronçon de chemin</nom_classe>
    <definition>Voie de communication terrestre … </definition>
          …
    <regroupement> Voir les différentes valeurs de l'attribut &lt;nature&gt;
    </regroupement>
          …
    <attributs>
    <attribut>
      <nom_attribut> Nature </nom_attribut>
      <definition> Permet de distinguer plusieurs… </definition>
      <type> Énuméré</type>
      <valeurs> Chemin empierré/Chemin/Sentier/Escalier … </valeurs>
      <attribut_valeur>
        <valeur> Chemin empierré </valeur>
        <definition> Route sommairement revêtue … </definition>
        <regroupement> Allée (carrossable)|Piste|Route empierrée
         </regroupement>
      </attribut_valeur>
          …
    </attribut>
  </classe></domaine> </document>
```

Figure 2 : extrait du document XML correspondant à la classe **Tronçon de chemin**

4.2 Traitement du document XML

4.2.1 Exploitation des balises et de leur imbrication

Une étude systématique des balises du document XML et de leur imbrication a permis de déterminer comment identifier concepts, relations conceptuelles et propriétés.
Concepts : chacun des termes existant entre les balises <*nom_domaine*>, <*nom_classe*>, <*regroupement*>, <*nom_attribut*> (lorsque l'attribut est qualitatif) et <*valeur*> donne lieu à la définition d'un concept dont le label correspond à ce terme
Relations hiérarchiques : émanent des relations entre les balises identifiées (à savoir <*nom_domaine*> et < *nom_classe*>, < *nom_classe*> et <*regroupement*>, <*nom_attribut*> et <*valeur*>, <*valeur*> et <*regroupement*> et des concepts associés.
Propriétés : portées par les termes encadrés par les balises <*attribut*> lorsque ceux-ci sont quantitatifs, et associées aux concepts encadrés par les balises <*nom_classe*>
Relations conceptuelles autres que hiérarchiques : portées par les termes encadrés par les balises <*attribut*> lorsqu'ils sont qualitatifs, et associées aux concepts dont les noms sont encadrés par les balises <*nom_classe*>
La sémantique des propriétés et des relations autres que hiérarchiques ne peut être déterminée que suite à une analyse de la balise <type> et du texte qu'elle marque.

Cette analyse ne présente pas de difficulté majeure car les balises véhiculent elles-mêmes leur sémantique et les relations découlent de la connaissance du domaine.

4.2.2 Exploitation du texte à l'aide de patrons lexico-syntaxiques

Le document de spécification de la base de données BDTopo contient des textes très courts et très synthétiques, donc assez pauvres en matière d'expression de relations entre concepts. Le champ *définition* renferme néanmoins quelques expressions de la relation de méronymie ou de définition de propriétés. Soit l'extrait suivant :

<classe>
<nom_classe> Tronçon de route </nom_classe>
<définition> **Portion de** voie de communication destinée aux automobiles >/definition>
</classe>

Figure 3. Exemple de définition de la classe/concept **Tronçon de route**

a) La relation de méronymie peut être caractérisée par les termes *partie de*, *portion de, constitué de, formé de, composé de,* etc. Le patron permettant d'identifier cette relation et écrit selon le formalisme JAPE est le suivant :

(({Token.lemme== "portion"}|{Token.lemme== "partie"}|...)
({Token.lemme== "de"}) ({Terme}) :annot
) - - > annot.ANNOT = {kind="Partie", rule="Rule1"}

Ce patron recherche un des mots *partie*, *portion*, *composer*, etc. suivi du mot *de*, suivi d'un *Terme* (obtenu à l'aide d'un extracteur de termes, et annoté comme tel). Ce terme réannoté *Partie* (partie droite de la règle) à l'aide de la règle *Rule1*, sera relié à la classe qui recouvre le champ <définition> où s'applique le patron, par la relation *partie-de*. Dans notre exemple, le concept *Tronçon de route* sera relié au concept *Voie de communication* par la relation *partie-de.*

b) Une propriété peut être identifiée par une forme lexico-syntaxique composée d'un terme désignant un concept, suivi d'un participe passé (PP) puis d'une préposition (PREP). Le patron suivant, appliqué à l'exemple de la figure 3 permet d'identifier une nouvelle propriété *destiné aux automobiles* pour le concept *Tronçon de route* :

({Concept}
({Token.category==PP} {Token.category==PREP} {Terme}) : annot
) - - > annot.ANNOT = {kind="Propriete", rule="Rule2"}

4.2.3 Mise en œuvre

Ces traitements ont été réalisés à l'aide de la plate-forme GATE[2] dont le principe est d'appliquer successivement sous forme de pipeline des ressources linguistiques et/ou des ressources de traitement sur un corpus. Le résultat est un corpus annoté, et ces annotations peuvent faire l'objet de divers traitements à l'aide de règles écrites en JAPE ou en Java. Comme GATE dispose d'une *Ontology API,* une ontologie peut être construite à partir de l'exploitation de ces annotations. Par ailleurs, GATE considère une balise XML du document original comme une annotation, ce qui permet :

1) d'exploiter les balises XML comme des annotations, l'imbrication des balises comme des recouvrements d'annotations, pour construire une première ontologie

[2] General Architecture for Text Engineering : plate-forme d'ingénierie linguistique développée à l'Université de Sheffield (http://gate.ac.uk)

2) d'appliquer des outils du TAL pour produire des étiquettes morpho-syntaxiques, puis de les exploiter par des patrons pour enrichir l'ontologie

Nous avons défini 6 règles exploitant les balises pour réaliser le noyau de l'ontologie et 5 règles correspondant aux patrons pour enrichir ce noyau. Nous donnons en figure 5, un extrait de l'ontologie obtenue, qui correspond aux spécifications de la figure 1.

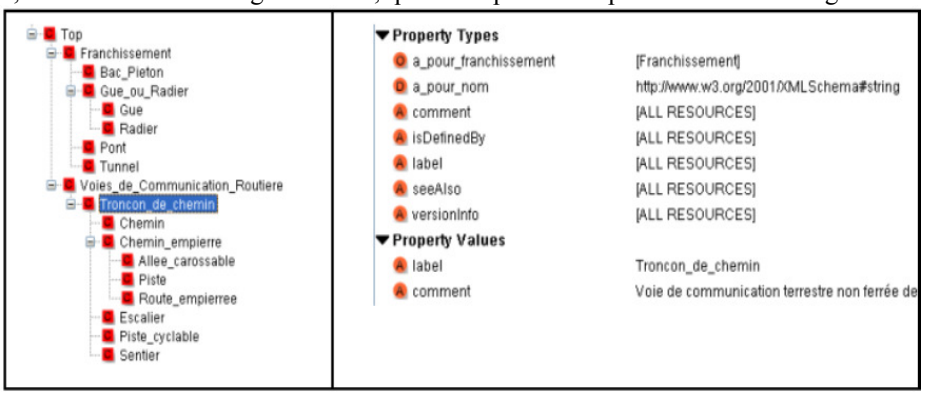


Figure 5. Extrait de l'ontologie obtenue avec GATE, conforme aux spécifications de la figure1.

Le concept *Tronçon de chemin* (<nom_classe>) est un fils du concept *Voies de communication routière* (<nom_domaine>) et a pour fils *Chemin*, *Escalier*, *Piste cyclable*, *Sentier* et *Chemin empierré* (<valeur>), qui à son tour a pour fils *Allée carrossable*, *Piste* et *Route empierrée* (<regroupement>). *Tronçon de chemin* est lié au type *String* par la relation *a-pour-nom* de type DataProperty, et au concept *Franchissement* par la relation *a-pour-franchissement* de type ObjectProperty.

5 Evaluation

Nous avons évalué l'apport de notre méthode sur ce type de corpus, par rapport aux approches connues. Au vu des spécifications, une approche statistique ne peut fournir de résultats significatifs (les textes étant très courts et très succincts, les concepts ne sont cités que peu de fois). De même, l'approche linguistique ne peut être satisfaisante car il y a relativement peu de texte rédigé (la plupart des expressions contenues entre les balises sont des noms de concepts). Par contre, une approche basée uniquement sur la structure visuelle du texte pour construire automatiquement des ontologies à partir de ces mêmes spécifications existe (Laurens, 2006). Nous avons donc comparé les deux méthodes à travers les ontologies obtenues.

5.1 Ontologie basée sur la structure visuelle

Cette étude (Laurens, 2006) se base sur un document HTML généré automatiquement à partir du document Word correspondant aux spécifications d'une base de données géographique, et exploite uniquement la structure visuelle du texte (style, caractères gras, soulignement, encadrement). Le document HTML est converti au format XML

en imbriquant les éléments du texte selon leur structure visuelle. Parallèlement, les groupes nominaux présents dans le texte sont extraits par des transducteurs et soumis à un expert pour validation comme termes du domaine : les termes retenus sont projetés sur le document XML. Le résultat, une taxonomie de concepts, correspond alors à la hiérarchie des termes associés dans le document XML.

5.2 Comparaison des deux ontologies obtenues

Les deux ontologies Onto_SV et Onto_ST obtenues respectivement par l'approche de F. Laurens et notre approche sont issues des mêmes spécifications (base BDTopo). Dans les deux cas, la mise au point des outils d'analyse du document s'appuie sur une interprétation approfondie de la sémantique des balises et de leur imbrication. La construction de Onto_SV a nécessité une intervention humaine à différents stades de sa conception (sélection/validation des expressions géographiques, nettoyage manuel de la hiérarchie XML obtenue, nettoyage/réorganisation manuel de la taxonomie OWL), alors que l'ontologue n'intervient dans Onto_ST que pour corriger, une fois l'ontologie construite, les incohérences dues à des erreurs de spécification (voir §5.3). Nous donnons figure 6 un état comparatif de ces deux ontologies.

	Onto_SV	Onto_ST
Nombre de concepts	615	1251
Profondeur	6	6
Relation hiérarchique "est-un"	oui	oui
Propriétés	non	oui
Relation de méronymie	non	oui
Relations conceptuelles autres	non	oui
Mode de construction	Supervisé	Non supervisé

Figure 6 : tableau comparatif des deux ontologies

Onto_ST est construite de manière automatique et est plus riche que Onto_SV en termes de concepts (notre méthode différencie les concepts portant un même label, lorsque ceux-ci se différencient par leurs propriétés) et de relations (relations autres que hiérarchiques).
Onto_ST n'est certainement pas la meilleure ontologie du domaine que l'on puisse obtenir, mais c'est la plus proche des spécifications.

5.3 Limites et intérêts de notre approche

La qualité de l'ontologie obtenue dépend entièrement de la qualité des spécifications : lorsque des incohérences existent au niveau des spécifications, une intervention humaine s'impose pour corriger l'ontologie. Et c'est là un des intérêts de la formalisation : aider à repérer des informations trop peu précises ou des incohérences au sein de documents a priori très structurés comme des spécifications. Or les variations de sens (au niveau du lexique, de la structure ou de la mise en forme) sont une des caractéristiques des textes. L'analyse du document a soulevé des cas pour

lesquels soit la relation identifiée n'avait pas la sémantique attendue, soit un des éléments d'une énumération avait un statut différent des autres, etc. Nous pointons ici quelques unes de ces anomalies.

5.3.1 Problèmes au niveau de la hiérarchie des concepts

Prenons pour exemple l'attribut "Autre classement" qui qualifie "route" :

Classement = « Autre classement »

Définition : Route qui ne fait partie ni du réseau autoroutier, ni du réseau national, ni du réseau départemental (voir ci-dessus).
Regroupement : Voies goudronnées (voies communales, chemins ruraux ou voies privées) | Rues | Rues piétonnes

Le champ *Regroupement* (qui désigne des concepts fils de ce type de route) met au même niveau *Rues* et *Rues piétonnes*, alors qu'il serait naturel de caractériser *Rues piétonnes* comme une spécialisation de *Rues*. De nombreux autres cas de ce type ont été rencontrés dans les énumérations, qui devraient ne pas être considérées comme contenant des concepts systématiquement de même niveau.

5.3.2 Incohérence au niveau des relations conceptuelles

Tronçon de route

Définition : Portion de voie de communication destinée aux automobiles, homogène pour l'ensemble des attributs et des relations qui la concernent. Représente uniquement la chaussée, délimitée par les bas-côtés ou les trottoirs.

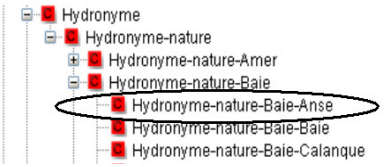
Géométrie : Linéaire

Attributs
- Identifiant (1)
- Source géométrique des données (1)
- Nature
- Classement
- Département gestionnaire
- Fictif

Le concept *Tronçon de Route* est défini à partir d'une <classe> du <domaine> *Voies de communication routière.* La règle d'interprétation des balises conduit à définir deux concepts liés par la relation hiérarchique "est-un". Or un patron de méronymie projeté sur le champ *définition* montre que *Tronçon de route* est une *partie-de voie de communication*. Ce genre d'incohérence (deux relations hiérarchiques entre les mêmes concepts) ne peut être levé que par intervention humaine.

5.3.3 Présence d'un même concept à différents niveaux de la hiérarchie

Les spécifications font qu'un même terme peut se retrouver à différents niveaux.



On retrouve ainsi le terme *Anse* comme désignant un concept fils de *Tronçon de laisse* et comme un concept fils de *Baie*, lui-même concept fils de *Hydronyme*. Une première solution est de créer un seul concept *Anse* et de lui associer plusieurs concepts pères. Or les spécifications donnent des définitions et des propriétés différentes dans chaque cas. Pour respecter la structure du document, nous avons choisi de concaténer le nom du concept courant à celui de ses concepts pères. Ceci permet de différencier les deux concepts *Anse*, et de fournir par ailleurs une traçabilité de l'ontologie vers le document de spécification qui a servi à la construire.

6 Conclusion et perspectives

Nous avons montré que, dans le cas favorable où des textes sont structurés à l'aide de balises dont la sémantique est claire, et dont la hiérarchisation porte aussi une sémantique précise, il est possible de définir une chaîne de traitements efficace pour construire automatiquement une ontologie. Ces traitements s'appuient sur des règles exploitant à la fois la structure des documents, le texte en langage naturel et en partie la mise en forme matérielle. Ils étendent donc les informations habituellement exploitées pour l'extraction de relations à partir de texte. L'ontologie ainsi obtenue à partir de textes s'avère riche en concepts et relations, et un lien précis est assuré entre éléments d'ontologie et textes. Nous sommes conscients que l'ontologie contient des incohérences qu'il faudra corriger manuellement. Dans le cadre de GEONTO, ces ajustements se feront après une étape d'alignement entre les différentes ontologies obtenues. Notre approche a été implémentée avec la plate-forme GATE.
L'évolution majeure envisagée pour notre méthode sera d'en assurer la portabilité à tout document de spécification de bases de données. Pour cela, nous pensons paramétrer les règles génériques en fonction des types de balises. Pour le moment, notre objectif sera d'être capable d'analyser tous les documents de spécification à traiter dans le projet GEONTO. Par ailleurs, nous comptons enrichir l'ontologie obtenue, d'une part en analysant plus systématiquement la mise en forme matérielle du texte (interprétation d'autres types d'énumération (Luc, 2001), des parenthèses, etc.), et d'autre part à l'aide de concepts, de relations ou de termes provenant de ressources externes. Dans le cas de GEONTO, il s'agira de concepts, relations et termes tirés de textes grand public, étude réalisée par le LIUPPA.

Références

ASHER N., BUSQUET J. ET VIEU L. (2001), La SDRT: une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum* 23, 73-101.

AUGER A., BARRIERE C. (2008), Pattern based approaches to semantic relation extraction : a state-of-the-art. *Terminology,* John Benjamins*, 14-1,1-19.*

AUSSENAC-GILLES N., SEGUELA P. (2000), Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire.* Numéro spécial linguistique de corpus. A. Condamines (Ed.). Toulouse : Presse de l'UTM. 25, 175-198.

AUSSENAC-GILLES N., DESPRES S., SZULMAN S. (2008), The TERMINAE Method and Platform for Ontology Engineering from texts. *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), IOS Press, p. 199-223.

BARRIÈRE C., AGBADO A. (2006), TerminoWeb : a software environment for term study in rich contexts. *International Conference on Terminology, Standardization and Technology Transfert (TSTT 2006), Beijing (China)*, p. 103-113.

BOURIGAULT D. (2002), UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *TALN 2002,* Nancy*, 24-27* juin *2002*

BUITELAAR P., CIMIANO P., MAGNINI B. (2005), Ontology Learning From Text: Methods, Evaluation and Applications. *IOS Press.*

CHAROLLES M. (1997), L'encadrement du discours : Univers, Champs, Domaines et Espaces. *Cahier de Recherche Linguistique, LANDISCO, URA-CNRS 1035, Univ. Nancy 2, n°6, 1-73.*

GARDARIN G., BEDINI I., NGUYEN B. (2008), B2B Automatic Taxonomy Construction, ICES (3-2) 2008 : 325-330

GIULIANO C., LAVELLI A., ROMANO L. (2006), Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. *In Proc. EACL 2006*.

GRCAR M., KLEIN E., NOVAK B. (2007) Using Term-Matching Algorithms for the Annotation of Geo-services. Post-proceedings of the ECML-PKDD 2007 Workshops, Springer, Berlin − Heidelberg − New York.*Boston, MA : Kluwer Academic Plubisher* .

GREFENSTETTE G. (1994), Explorations in Automatic Thesaurus Discovery. *Boston, MA : Kluwer Academic Plubisher.*

HEARST M.A. (1992), Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics*, Morristown, NJ, USA, ACL, 539–545.

HINDLE D. (1990), Noun classification from predicate argument structures. *In Actes, 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Berkeley USA.*

JACQUEMIN C. (1997), Présentation des travaux en analyse automatique pour la reconnaissance et l'acquisition terminologique. *In Séminaire du LIPN,* Université Paris 13, Villetaneuse.

LAURENS F. (2006), Construction d'une Ontologie à partir de Textes en Langage Naturel. *Rapport de Stage Master 1 en linguistique-Informatique, Septembre 2006*

LUC C. (2001), Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte*. TALN2001*, *Université de Tours*, juillet 2001, p. 263-272.

MAEDCHE A. (2002), *Ontology learning for the Semantic Web*, vol 665. Kluwer Academic Pub.

NÉDELLEC C., NAZARENKO A. (2003). Ontology and Information Extraction. in S. Staab & R. Studer (eds.) *Handbook on Ontologies in Information Systems*, Springer.

REBEYROLLE J., TANGUY L. (2000), Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25, 153-174

TIRMIZI S., SEQUEDA S., MIRANKER J.F (2008), Translating SQL Applications to the Semantic Web. Dexa 2008, Turin , Italie, 450-464

VIRBEL J., LUC C. (2001), Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum, Vol. XXIII, N. 1, p. 103-123.*

WEISSENBACHER D., NAZARENKO A. (2007), Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. *TALN 2007*, Toulouse, Juin 2007.

Approche générique pour l'extraction de relations à partir de textes

Seif Eddine Kramdi, Ollivier Haemmerlé et Nathalie Hernandez

IRIT – Université de Toulouse, UTM, 5 allées Antonio Machado, F-31058 Toulouse Cedex 9
{kramdi, haemmer, hernande}@irit.fr

Résumé : Cet article s'intéresse à l'extraction de relations dans le contexte du web sémantique, en vue de procéder à de la construction d'ontologies aussi bien qu'à de l'annotation automatique de documents. Notre approche permet l'extraction de relations entre entités à partir de textes. Elle ne fait pas d'hypothèse sur les entités, de manière à la rendre aussi générique que possible, et à autoriser par exemple l'extraction de relations entre concepts aussi bien que l'extraction de relations entre instances de concepts. Pour atteindre cet objectif, nous nous fondons sur l'algorithme LP^2. Afin d'adapter cet algorithme à l'extraction de relations, nous proposons une nouvelle notion de contexte reposant sur un graphe de dépendances, généré par un analyseur syntaxique. Un tel graphe de dépendances est bien adapté à la représentation de relations, puisqu'il permet, notamment, de repérer aisément les différents arguments d'un verbe dans une phrase. Nous présentons l'implémentation réalisée suivie d'une première phase d'expérimentations.

Mots-clés : Web sémantique, Extraction de connaissances, Apprentissage.

1 Introduction

La masse d'information disponible sur le Web dans des documents essentiellement textuels n'est quasiment plus exploitable sans avoir recours à des aides automatisées. L'extraction de connaissances est centrale dans ce processus d'automatisation. Des travaux visent à analyser des documents dans l'objectif d'extraire des connaissances, et ceci dans un format permettant leur exploitation par des humains ou des machines. L'extraction de connaissances peut servir à construire des ontologies en extrayant des concepts et des relations. La construction et la mise à jour d'ontologies sont des processus essentiels en vue d'une exploitation pérenne du Web. L'approche proposée dans le projet SmartWeb (Schutz & Buitelaar, 2005) pour l'extraction de relations entre concepts s'inscrit dans ce cadre. L'extraction de connaissances peut également être utilisée pour le peuplement (ou instanciation) d'ontologies. Le peuplement d'ontologies à l'aide de méthodes d'extraction de connaissances revient à trouver et associer aux concepts de ces ontologies, des instances décrites dans des textes (Cimiano, 2006). Les différentes ontologies peuplées peuvent alors être utilisées pour annoter les textes qui ont servi au peuplement.

Ces deux utilisations de l'extraction de connaissances – pour construire des ontologies d'une part, pour les peupler en vue d'annoter d'autre part – sont considérées comme faisant partie du Web Sémantique, cadre dans lequel nous plaçons nos travaux.

La particularité des extracteurs de connaissances destinés à une utilisation dans un environnement Web est qu'ils doivent tenir compte de la nature hétérogène des documents mais également, comme la plupart des applications relatives au Web, être en mesure de supporter le passage à l'échelle. L'étude des différentes méthodologies d'extraction actuelles conduit à constater la quasi-absence de démarches indépendantes du domaine d'utilisation présentant de bonnes performances, ou même, à tout le moins, de démarches ne nécessitant pas un gros effort pour leur adaptation à d'autres cas d'utilisation. Ceci est dû à la nécessité d'une grande fiabilité dans le processus d'extraction. Les systèmes développés – souvent des systèmes propriétaires – le sont pour une utilisation donnée ; ils se fondent sur des particularités du type de textes analysés, en élaborant par exemple des règles linguistiques très spécifiques (Aussenac-Gilles & Jacques, 2008). De telles règles sont très liées à la manière dont les informations sont exprimées dans le domaine d'utilisation (Fundel et al., 2007).

L'approche que nous présentons dans cet article a pour but de définir une méthode d'extraction de relations dans des textes, pouvant servir à la fois au peuplement d'ontologies, à la pose d'annotations dans des textes et à la construction d'ontologies. Elle se veut la plus généraliste possible dans le sens où elle doit permettre d'extraire des entités liées sémantiquement dans les textes. Ces entités peuvent être des instances ou des concepts, la seule hypothèse faite étant que les documents sont préalablement annotés par ces entités. Notre méthode d'extraction doit identifier la relation qui lie les entités en question. Notons que, contrairement à l'extraction d'instances de concepts – où les performances des systèmes commencent à être intéressantes, atteignant une précision d'extraction de plus de 90% pour certaines applications – l'extraction d'instances de relations est loin d'atteindre de tels résultats. Pourtant, elle représente un enjeu capital pour l'annotation automatique. Certaines démarches sont néanmoins intéressantes, comme par exemple RelExt (Schutz & Buitelaar, 2005) qui fonde son extraction sur une théorie linguistique centrée sur l'expression des relations à travers la sémantique des verbes et sur une étude statistique sur les occurrences de ces verbes. Cette approche n'est cependant proposée que pour l'enrichissement d'ontologies : les relations extraites se situent uniquement au niveau des concepts. Contrairement à certaines approches (Velardi & al., 2007), nous choisissons de ne pas nous focaliser sur un type de relation donné.

Afin de tenir compte de notre contexte d'utilisation, nous souhaitons que notre extracteur soit :

- adaptatif : l'algorithme devra pouvoir générer de manière automatique des règles d'extraction « adaptées » aux types de textes étudiés et à la nature de la relation à extraire. Il devra pouvoir repérer par quel moyen

la relation s'exprime dans les textes analysés, et ceci à chaque nouvelle utilisation ;

- autonome : l'intervention d'experts dans le processus de génération des règles doit être limitée. L'utilisation de démarches d'apprentissage semble donc opportune ;
- robuste : l'algorithme doit être en mesure de faire face au volume important de ressources à traiter sur la toile.

Notre étude se positionne dans le cadre du projet ANR WebContent qui vise à concevoir une plate-forme de développement d'applications fondées sur les technologies du web sémantique. Le but de notre étude est principalement de fournir à cette plate-forme un extracteur de relations utilisables par les diverses applications construites sur la plate-forme sous forme d'un service Web. Il pourra être utilisé, en le combinant à d'autres services Web, à l'élaboration de tâches plus complexes : annotation de documents, peuplement d'ontologies, construction d'ontologies, etc.

Ce travail est soutenu par l'ANR dans le cadre du projet plate-forme WebContent.

2 L'algorithme LP^2

LP² (Ciravegna, 2001) – pour Learning pattern by language process – est un algorithme adaptatif d'annotation : il génère des règles d'annotation adaptées aux textes traités. Il a été défini pour annoter des documents textuels à partir d'instances de concept. L'algorithme que nous proposons dans cet article est une adaptation de LP^2. Le choix de cet algorithme comme base de notre travail a été motivé par plusieurs de ses caractéristiques :

- son caractère adaptatif : en effet LP² génère des règles d'extraction plus ou moins complexes suivant la richesse du langage utilisé pour exprimer la relation dans le texte ;
- l'approche inductive de l'algorithme, qui permet de générer des règles d'extraction, à l'inverse des démarches transductives peu adaptées à l'extraction de connaissances sur de gros volumes de données ;
- ses performances qui ont été évaluées dans le cadre de différents types d'applications, comme Amilcare (Ciravegna & Wilks, 2003).

2.1 Principe de LP^2

L'algorithme LP^2 effectue sa phase d'apprentissage des règles d'extraction de connaissances à l'aide d'un ensemble d'apprentissage décomposé en deux sous-ensembles : le premier est utilisé pour la génération des règles, le deuxième servant pour sa part à la sélection des règles les plus efficaces parmi les règles générées.

Une règle d'extraction se compose de deux parties : une partie condition, où l'on vérifie l'existence d'un contexte donné – dans l'exemple suivant, nous vérifions l'existence d'un enchaînement de 5 mots – et d'une partie action qui, dans ce cas,

insère un tag de type stime – starting time – entre le 3ème et le 4ème mot, signifiant en cela qu'une instance du concept stime (heure de début) a été identifiée.

Condition	Action
Word=?	Insert Tag
the	
seminar	
at	stime
4	
pm	

Fig. 1 – Exemple de règle d'extraction générée par LP²

LP² est un algorithme inductif. Cela veut dire que les règles qu'il utilise sont induites à partir d'exemples, de la manière suivante :

- une fenêtre de mots appartenant à une phrase de l'ensemble destiné à la génération est considérée ; une première règle fondée sur cette fenêtre est générée (la figure 1 est l'illustration d'une telle règle) ;
- des propriétés – telles que la catégorie grammaticale du mot, le lemme, la casse, la catégorie sémantique identifiée à partir d'une ontologie, etc. – sont ajoutées pour chacun des mots de cette fenêtre (la règle présentée dans la partie gauche de la figure 2 est l'illustration d'une telle règle) ;
- La règle est ensuite généralisée de différentes manières, en relâchant les contraintes portant sur les propriétés ajoutées à l'étape précédente. Les règles généralisées ainsi produites sont testées sur l'ensemble de filtrage, de manière à sélectionner les règles présentant les meilleurs taux de rappel et de précision (la partie droite de la figure 2 présente une règle généralisée induite).

word index	Condition	Additional Knowledge				Action
	Word	Lemma	LexCat	case	SemCat	Tag
1	The	the	Art	low		
2	Seminar	Seminar	Noun	low		
3	at	at	Prep	low		stime
4	4	4	Digit	low		
5	pm	pm	Other	low	timeid	
6	will	will	Verb	low		

Word index	Condition					Action
	Word	Lemma	LexCat	Case	SemCat	Tag
3		at				Stime
4			Digit			
5					timeid	

Fig. 2 – A gauche exemple d'une règle étendue, à droite exemple d'une règle induite à partir de cette règle étendue

2.2 Limitation de LP^2 pour l'extraction de relations entre entités

Etant un algorithme destiné à la détection d'instances de concepts, LP² nécessite une adaptation afin d'autoriser l'extraction de relations entre deux entités. Une adaptation naïve pourrait consister à augmenter la fenêtre prise en compte dans la partie condition de la règle, en la faisant porter sur les contextes des occurrences des deux entités liées par la relation. La taille du contexte serait donc multipliée par deux – mots voisins des deux arguments de la relation. L'induction des propriétés d'extraction paraît alors peu adaptée, et ceci en raison :

- de l'augmentation importante du temps de calcul nécessaire à la généralisation d'un contexte trop grand ;
- de l'absence de certitude que les éléments à travers lesquels la relation est exprimée soient aux alentours des entités – un verbe éloigné, une référence qui se situe en dehors du contexte étudié, etc.

Nous proposons donc de modifier LP² en définissant une nouvelle notion de contexte, mieux adaptée à l'extraction de relations.

3 Approche proposée

3.1 Modifications de l'algorithme LP^2

Afin d'extraire des relations, nous proposons de modifier l'algorithme LP^2 en redéfinissant la notion de contexte utilisée pour l'expression des conditions des règles d'extraction. Nous proposons de repérer les éléments exprimant la relation en nous fondant sur la structure grammaticale de la phrase, grâce à une représentation sous forme de graphes de dépendances. En effet, de tels graphes permettent de repérer des relations entre des entités assez éloignées dans la phrase d'origine (Fundel et al., 2007), alors qu'elles sont proches dans le graphe. D'autre part, nous avons à notre disposition des outils d'analyse de texte permettant de générer de tels graphes automatiquement, comme par exemple (Stanford, 2008). Un graphe de dépendances a une structure d'arbre. Les nœuds fils d'un nœud donné sont appelés « expansion » du nœud, le nœud père étant appelé la « tête ». Il existe une relation de dépendance syntaxique entre une tête et ses expansions. Considérons par exemple la phrase « l'ingénieur d'Airbus a réparé l'A380 en panne », et le graphe de dépendances lui correspondant, représenté dans la figure 3.

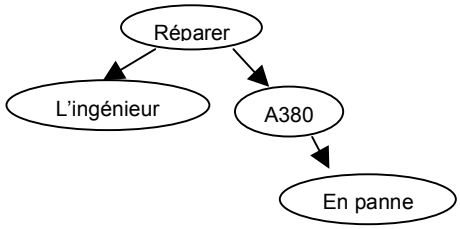


Fig. 3 – Exemple d'un graphe de dépendances

On constate sur cet exemple que la tête d'un nœud permet de décrire la structure narrative dans laquelle il intervient – l'ingénieur intervient dans une réparation, la panne est dans l'A380, etc. De même, si l'on se place du point de vue des expansions, ces dernières peuvent être vues comme des modalités modifiant le sens de leurs pères – l'A380 est qualifié comme étant en panne, etc. De plus, dans plusieurs théories linguistiques (Schutz & Buitelaar, 2005) et (Fundel et al., 2007), le verbe central – le nœud racine de l'arbre – peut être perçu comme le représentant de l'action ou de la situation décrites par la phrase – ici, le nœud central « réparer » précise qu'il s'agit d'une réparation.

Etant donné un graphe de dépendances dans lequel nous définissons deux nœuds comme étant les arguments potentiels d'une relation (arg1 et arg2), le contexte étudié pour savoir si une relation est exprimée ou non entre les arguments sera l'ensemble des nœuds têtes et expansions reliés aux deux arguments et le nœud racine de l'arbre. La figure 4 schématise le graphe de dépendances considéré.

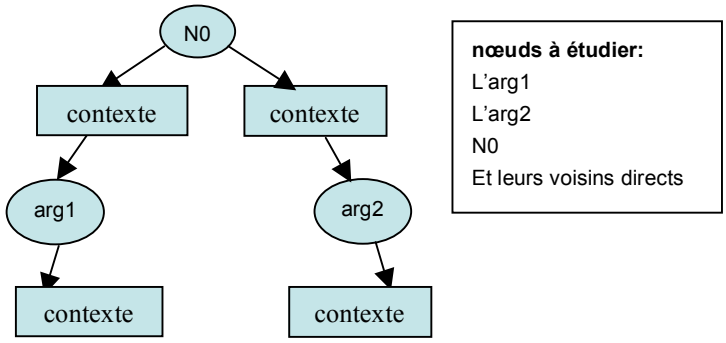


Fig. 4 – Contexte à étudier

L'essentiel des modifications que nous apportons à l'algorithme LP² peut se résumer comme suit :

- l'induction de règles d'extraction se fait maintenant sur les contextes des arguments dans un graphe de dépendances ;
- l'ajout de nouvelles propriétés sur lesquelles se fait l'induction, telle que la position du mot par rapport aux arguments (avant, après ou entre les arguments).

L'exemple suivant illustre la forme des règles que l'algorithme d'induction modifié renvoie. Considérons la phrase suivante de l'ensemble d'apprentissage, préalablement annotée : « <arg1>The 50 planes </arg1> will be <relation deliver>delivered </relation deliver> to <arg2>the airline</arg2> from 2009 to 2010. » La figure 5 représente le graphe de dépendances généré automatiquement, sur lequel nous allons fonder la partie condition de la règle induite.

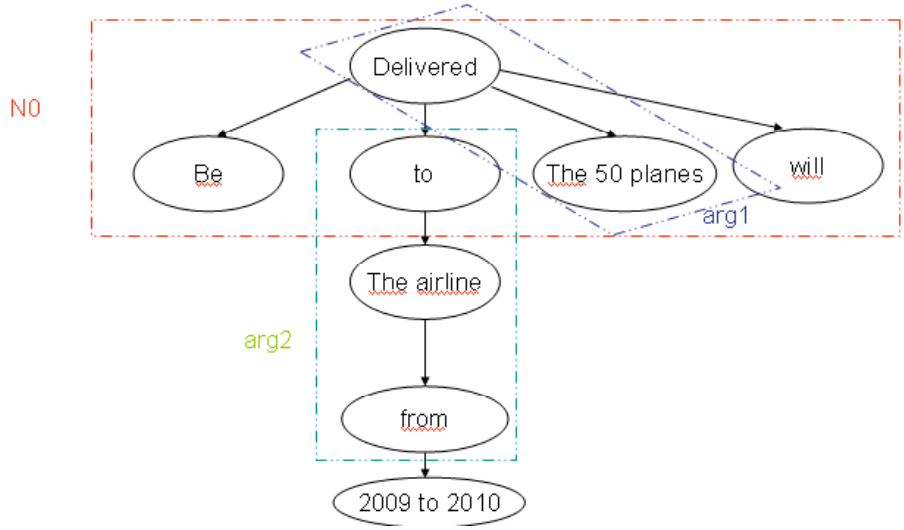


Fig. 5 – Graphe de dépendances généré à partir de la phrase extraite de l'ensemble d'apprentissage

A partir du graphe de dépendances, nous générons la règle suivante, signifiant que si la racine de l'arbre correspond à un verbe identifié comme appartenant à la catégorie sémantique « relation deliver », que son premier argument appartient à la catégorie sémantique « airline » et que son deuxième argument appartient à la catégorie sémantique « plane », alors il existe une relation « deliver » entre ces deux arguments.

Si (N0 semC "relation deliver") ➔(label:"to) ➔(arg1 Csem "airline")
➔arg2 (Csem "plane")

Alors (arg1, arg2) appartient à la relation deliver

3.2 Implémentation

Pour que notre implémentation soit aussi générale que possible, et permette la prise en compte de propriétés variées sur les nœuds du graphe de dépendances, nous

avons utilisé, pour la représentation du contexte, une structure en liste regroupant des éléments de type (nœud, propriété, valeur) sans pour autant spécifier la nature de la propriété utilisée ou les valeurs qu'elle pourrait prendre. Cette façon de procéder nous permet de laisser une grande liberté, en fonction de l'utilisation que l'on souhaite faire de l'algorithme. Par exemple, si l'on souhaite extraire des relations entre concepts plutôt qu'entre instances, les propriétés considérées pourront concerner la définition des concepts dans l'ontologie.

Le processus de génération des règles (qui seront utilisées dans le processus d'extraction) est sensiblement le même que celui adopté par Ciravegna dans (Ciravegna, 2001), mis à part le fait que la génération de règles se fait à présent sur la notion de contexte que nous proposons et qui repose sur les graphes de dépendances. La première étape du processus de génération consiste à créer une règle élémentaire pour chacun des nœuds de l'arbre enrichi par ses propriétés (catégorie sémantique obtenue à partir de l'ontologie, lemme, position dans le graphe, etc.). Ces règles élémentaires sont ensuite combinées, et les différentes combinaisons sont évaluées sur l'ensemble de filtrage des règles générées.

L'implémentation a été réalisée en java 1.5, en utilisant l'environnement de développement Eclipse. Nous avons développé environ 1000 lignes de code réparties sur 9 classes.

4 Expérimentations

Le corpus que nous avons utilisé pour l'évaluation de l'algorithme est celui proposé dans la campagne LLL05 challenge – learning language in logic 2005 (LLL, 2005). Le but de la campagne est principalement d'évaluer les performances d'algorithmes d'apprentissage pour l'extraction de règles décrivant des interactions entre des protéines et des gènes dans des résumés d'articles du domaine biomédical.
Nous considérons pour l'évaluation :

- **un corpus** composé de 55 phrases offrant 103 exemples positifs (interactions exprimées) et 27 exemples négatifs ;
- **un dictionnaire d'entités nommées** incluant les variantes et les synonymes des différents gènes et protéines rencontrés dans les corpus.

Démarche d'évaluation

Pour évaluer notre algorithme, nous avons découpé *aléatoirement* le corpus en trois ensembles : 2 ensembles pour l'apprentissage et 1 ensemble pour l'évaluation de notre approche.

Pour l'apprentissage, nous considérons :

- **un ensemble pour la production des règles** représentant l'ensemble des exemples positifs sur lesquels se fera l'induction des règles d'extraction ;

- **un ensemble de filtrage pour la sélection des règles produites** représentant l'ensemble des exemples positifs et négatifs, sur lesquels nous nous appuierons pour décider des meilleures règles induites à conserver.

Pour l'évaluation de notre approche, nous définissons :

- **un ensemble de test** constitué des exemples positifs et négatifs qui restent, et sur lesquels seront appliquées les règles générées sur l'ensemble d'apprentissage.

L'objectif de l'évaluation est d'analyser les performances de notre algorithme en faisant varier les données prises en compte pour l'apprentissage. Nous souhaitons ainsi évaluer la robustesse de notre algorithme et analyser sa pertinence pour l'extraction de relations dans le cadre du web sémantique.

Notre évaluation s'appuie sur l'analyse des trois mesures suivantes : précision (P), rappel (R) et F-Mesure (F).

Test 1 : test de stabilité : influence de la taille de l'ensemble d'apprentissage sur les résultats

Ce test fait varier la taille de l'ensemble d'apprentissage de façon à évaluer si les performances de notre algorithme sont fortement liées à la quantité d'exemples utilisés pour l'induction et le filtrage des règles.

La taille de l'ensemble d'apprentissage varie de 110 à 80 exemples. La figure 6 représente les résultats obtenus sur cet ensemble de test.

Nombre d'exemples considérés pour l'apprentissage	110	100	90	80
précision	0,53	0,56	0,51	0,61
rappel	0,30	0,29	0,33	0,33
F-mesure	0,36	0,37	0,37	0,39

Fig. 6 – Résultats du test 1

Cette évaluation montre que notre approche ne nécessite pas forcément un ensemble d'apprentissage important. La F-Mesure est plus élevée lorsque seulement 80 exemples sont utilisés pour générer les règles. Les tests suivants visent à analyser les critères permettant de sélectionner les règles capturant le plus d'exemple tout en générant le moins de bruit.

Test 2 : variations sur l'ensemble d'exemples à généraliser : influence du nombre d'exemples considérés pour la génération des règles

Pour ce test, nous faisons varier le nombre d'exemples considérés pour l'induction des règles d'extraction d'une relation. Nous considérons, pour chaque relation, 1 à 4 exemples choisis aléatoirement pour induire les règles. Nous analysons également les performances obtenues pour les relations pour lesquelles 11 exemples sont disponibles, ce qui correspond au maximum disponible dans notre corpus de test.

nombre d'exemples utilisés pour induire les règles de chaque relation	1	2	3	4	11
P	0,55	0,55	0,54	0,49	0,54
R	0,22	0,32	0,37	0,35	0,64
F	0,29	0,39	0,43	0,39	0,58

Fig. 7 – Résultats du test 2

La diversification des règles induites par l'augmentation du nombre d'exemples que l'on généralise influence directement le rappel. Ceci est essentiellement dû au fait que diversifier les règles d'extraction permet d'englober un panel plus large d'exemples positifs. Les taux de rappel et de précision sont les plus importants pour les relations pour lesquelles 11 exemples sont disponibles. Ces 11 exemples permettent de capturer un ensemble représentatif des différentes manières selon lesquelles une relation est exprimée. Cependant, autant d'exemples sont rarement disponibles.

Test 3 : variations sur l'ensemble de filtrage des règles : sélection des règles à conserver en fonction du nombre d'exemples qu'elles capturent

Ce test vise à évaluer un critère de sélection des règles générées sur l'ensemble de production des règles. Nous proposons ici de sélectionner les règles en fonction du nombre d'exemples qu'elles capturent sur l'ensemble de filtrage. Les performances de ces règles sur l'ensemble de test sont représentées dans la figure 8.

Nombre minimum d'exemples de l'ensemble de filtrage capturés par les règles	2	3	4	5
P	0,48	0,54	0,59	0,5
R	0,4	0,39	0,41	0,47
F	0,4	0,42	0,48	0,46

Fig. 8 – Résultats du test 3

En faisant varier le minimum de 2 à 5 exemples capturés, nous pouvons constater une hausse des performances du système – avec une F-mesure variant de 0,40 à 0,46. Les règles englobant plus d'exemples sont en effet moins contraignantes, ce qui permet d'augmenter le rappel. Ce test montre que ce critère de sélection garantit une certaine qualité des règles.

Test 4 : variations sur l'ensemble de filtrage des règles : sélection des n « meilleures » règles

Dans ce test, nous évaluons l'impact du nombre de règles sélectionnées sur les performances du système. Les règles induites sont classées par rapport au nombre d'exemples capturés sur l'ensemble de filtrage et nous en sélectionnons les n premières (n variant de 5 à 20). Les performances obtenues sur l'ensemble de test sont présentées dans la figure 9.

Nombre de règles sélectionnées	5	10	15	20
P	0,55	0,53	0,48	0,52
R	0,24	0,45	0,44	0,52
F	0,3	0,47	0,48	0,51

Fig. 9 – Résultats du test 4

Nous remarquons que l'augmentation du nombre de règles diminue la précision – plus de règles impliquent plus d'exemples négatifs capturés – mais de son côté, le rappel ne cesse d'augmenter – de 0,23 pour 5 règles à 0,52 pour 20 règles.

En résumé, les performances de notre algorithme sont intéressantes car il permet en moyenne d'identifier des relations avec une F-mesure proche de 0,50 lorsque pour chaque relation, au moins 4 exemples sont disponibles pour générer les règles et 6 exemples sont disponibles pour filtrer les règles générées qui sont en moyenne au nombre de 20. Ceci permet d'envisager l'utilisation de notre approche pour l'extraction de relations entre instances ou de relations entre concepts. En effet pour ce genre d'application, il est raisonnable d'envisager l'existence d'un corpus de cette taille pour l'apprentissage.

5 Conclusion

Nous avons présenté dans cet article une approche permettant l'extraction de relations entre entités à partir de textes. Cette approche ne fait pas d'hypothèse sur les entités, de manière à la rendre aussi générique que possible, et d'autoriser par exemple l'extraction de relations entre concepts aussi bien que l'extraction de relations entre instances de concepts. Pour atteindre cet objectif, nous nous sommes fondés sur l'algorithme LP^2, qui a déjà fait ses preuves dans le cadre de l'extraction d'instances de concepts. Cet algorithme fonctionne en analysant le « contexte » dans lequel les instances de concepts sont recherchées. Afin d'adapter cet algorithme à l'extraction de relations, nous avons modifié cette notion de contexte, qui ne repose plus sur une fenêtre de mots mais sur un graphe de dépendances, généré par un analyseur syntaxique. Un tel graphe de dépendances est bien adapté à la représentation de relations, puisqu'il permet, notamment, de repérer aisément les différents arguments d'un verbe dans une phrase.

Nous avons mené une première phase d'expérimentations, en utilisant un corpus de taille modeste. Les premiers résultats nous paraissent encourageants et semblent valider notre approche. Il reste néanmoins à passer à l'échelle du Web. Pour cela, nous travaillons actuellement à la transformation de notre première implémentation en un service Web, qui sera intégré à la plate-forme nationale WebContent. Dès que cette intégration aura eu lieu, nous pourrons tester notre approche dans un contexte applicatif réel, tant pour la construction d'ontologie que pour l'annotation

automatique de documents, et ce sur un ou plusieurs des domaines d'applications retenus dans le projet WebContent.

Plusieurs pistes de recherche s'offrent à nous dans la suite de ce travail :

- nous devons procéder à une meilleure évaluation des performances de notre algorithme, et le comparer à d'autres systèmes d'extraction prenant en compte des textes hétérogènes, ce qui pourrait déterminer avec plus de précision la relation qui existe entre les performances de notre approche et la nature des textes analysés et des relations extraites ;
- l'amélioration de la notion de contexte extrait à partir de graphes de dépendances, par exemple en prenant en compte les chemins, en étudiant plusieurs niveaux de voisins, etc.
- l'étude des performances du système suivant l'ensemble d'exemples choisi pour la génération des règles d'extraction. En utilisant des exemples qui exprimeraient la relation à l'aide de différents procédés narratifs, il serait vraisemblablement possible de générer des règles couvrant un maximum d'exemples.

Références

Aussenac-Gilles N. & Jacques M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with cam_el_eon. *Terminology, special issue on Pattern-Based approaches to Semantic Relations*. 14(1):45-73.

Cimiano P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag.

Ciravegna F. (2001). (LP)2, an adaptive algorithm for information extraction from web-related texts. *Proceedings of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining held in conjunction with the 17th International Joint Conference on Artificial Intelligence*.

Ciravegna F. & Wilks Y. (2003). Designing Adaptive Information Extraction for the Semantic Web in Amilcare. In Handschuh, S. and Staab S., Eds. *Annotation for the Semantic Web*. IOS Press.

Fundel K., Zimmer R. & Küffner R. (2007). RelEx---Relation extraction using dependency parse trees. *Bioinformatics*, 23(3). pp. 365-371. Oxford University Press.

Iria J. (2005). T-Rex: A Flexible Relation Extraction Framework. *8th Annual CLUK Research Colloquium*. Manchester, UK.

LLL (2005). www.cs.york.ac.uk/aig/lll/lll05/. Proceedings of the Workshop Learning Language in Logic (LLL05). Bonn, Germany.

Schutz A. & Buitelaar P. (2005). RelExt: A Tool for Relation Extraction in Ontology Extension. *Proceedings of the 4th International Semantic Web Conference*. Galway, Ireland. LNCS #3729, pp. 593-606, Springer.

Stanford (2008). nlp.stanford.edu/software/lex-parser.shtml. The Stanford Parser: A statistical parser. *The Stanford Natural Language Processing Group*.

Velardi P., Cucchiarelli A., Petit M. (2007). A taxonomy learning method and its application to characterize a scientic web community. *IEEE Transactions on Knowledge and Data Engineering*. 19(2):180-191.

Ontologies pour l'aide à la décision publique et prise en compte des doxas

Maryse SALLES

IRIT/SIG/RI-EVI, Université de Toulouse (Toulouse 1)
Maryse.Salles@univ-tlse1.fr

Résumé : Ce papier présente un travail de conception d'ontologie, réalisé dans le cadre d'un projet de recherche centré sur l'aide à la décision publique en matière de développement économique territorial. Une analyse des usages de l'ontologie, ainsi que des spécificités du domaine et du terrain sont proposées. Ces spécificités conduisent à considérer le travail d'explicitation des visions du monde (*doxas*) en présence comme un préalable indispensable à la conception de l'ontologie. La coexistence de plusieurs de ces doxas dans le domaine nous amène à la construction d'une *ontologie polydoxique*, c'est-à-dire intégrant plusieurs doxas distinctes. Ce choix soulève des problèmes méthodologiques spécifiques, dont il est donné un aperçu. Une illustration est fournie au travers d'un extrait de l'ontologie intégrant deux doxas sur le concept de "territoire".

Mots-clés : Ontologie, Doxa, Ontologie polydoxique, Aide à la décision, Développement économique territorial.

Introduction

Ce papier présente un travail de construction d'ontologies pour l'aide à la décision en matière de développement économique territorial, travail réalisé au sein du projet CAVALA[1]. Ce projet, financé par la Région Midi-Pyrénées, est centré sur l'aide à la définition d'indicateurs pour évaluer les politiques de développement économique, et plus spécialement les aides financières aux entreprises. L'objectif de CAVALA est de concevoir une méthode coopérative pour construire ces indicateurs. La démarche du projet inclut comme élément central la construction d'une ontologie.

Le contexte du projet est marqué par la jeunesse de l'institution (la Région), renforcée par une évolution très rapide de l'étendue de ses missions, en particulier dans le domaine du développement économique territorial. La première conséquence en est une expertise encore réduite. Le projet Cavala se situe donc dans une perspective d'apprentissage des équipes de la Région, et plus globalement de gestion du changement.

[1] méthode CoopérAtive de suivi et d'éVAluation des poLitiques régionAles de développement économique ; le projet CAVALA est un projet pluridisciplinaire réalisé en coopération entre l'IRIT (Institut de Recherche en Informatique de Toulouse) et le LEREPS (Laboratoire d'Étude et de Recherche sur l'Économie, les Politiques et les Systèmes sociaux).

La partie 1 de ce papier présente succinctement les premiers usages de l'ontologie tels qu'ils sont aujourd'hui programmés. Au-delà de la jeunesse de l'institution, la décision publique territoriale, comme le domaine du développement économique, présentent un ensemble de caractéristiques propres qui influent puissamment sur la conception d'outils d'aide à la décision, et, dans notre cas, sur la construction d'ontologies. La partie 2 décrit ces particularités. Parmi celles-ci, il convient de souligner la présence simultanée de plusieurs visions du monde, dans les textes, mais aussi dans les pratiques. Ces visions du monde, ou plutôt ces doxas[2], semblent, paradoxalement, d'autant plus prégnantes qu'elles ne sont pas explicitées.

Le constat de cette *polydoxie*[3] nous a conduit à des choix méthodologiques qui sont exposés en partie 3. La démarche de construction de l'ontologie intègre ainsi la nécessité de prendre en compte les différentes doxas qui coexistent au sein du domaine. L'hypothèse clé est ici qu'une même ontologie peut (et dans certains contextes *doit*) intégrer plusieurs doxas.

Pour illustrer notre démarche, un extrait de l'ontologie est proposé en partie 4.

1. Usages de l'ontologie

A l'issue des premières semaines de CAVALA, les usages de l'ontologie tels qu'ils avaient été projetés dans le projet initial ont sensiblement évolué. L'analyse du besoin, et, plus largement, la compréhension de la situation de la Région nous ont fait privilégier une approche progressive, favorisant l'apprentissage des acteurs.

Nous n'évoquerons pas ici le travail sur les indicateurs : définition d'une typologie, puis d'un ensemble d'indicateurs, qu'il a été prévu de mettre en place progressivement. Notons seulement que compte tenu de la nouveauté du principe d'une évaluation régulière, il a été décidé de commencer par les indicateurs d'*effort*, c'est-à-dire ceux qui mesurent les ressources consacrées aux actions concernées (aides financières aux entreprises). Il s'agit par exemple de calculer le total des sommes versées aux entreprises, le nombre de dossiers traités, le nombre d'aides accordées, etc., en ventilant les chiffres selon un ensemble de dimensions significatives (par type de territoire, par type de PME, par type d'activité, par type de projet, etc.).

L'intérêt de ces indicateurs ainsi ventilés est qu'ils permettent d'avoir une vision des actions *effectivement* mises en œuvre, puis, dans un second temps, d'évaluer la concordance de ces dernières avec la politique annoncée. Ces indicateurs sont donc des outils d'évaluation, mais aussi d'aide à l'apprentissage pour le système de décision de la Région. Le premier usage de l'ontologie est l'aide à l'indexation des dossiers de demandes d'aide financière. Cette indexation se fait selon quelques dimensions

[2] Du grec *δόξα* : opinion. Ce terme relativement large couvre les notions de sens commun, d'opinion, de représentation sociale, voire d'idéologie.

[3] On peut avancer que la Région est polydoxique au sens de (Monteil & al., 1986). Elle mobilise en effet "un ensemble de croyances multiples à l'égard d'un même objet, [croyances] qui coexistent à l'état latent et sont extériorisables isolément (…)".

principales, dans le but de permettre le calcul des indicateurs d'effort. L'indexation sera réalisée au final par les personnels du service Industrie, en charge de l'instruction des dossiers de demande d'aide constitués par les entreprises. Les tableaux d'indicateurs d'effort sont destinés principalement aux élus et aux cadres de la direction à laquelle appartient le service Industrie. Un deuxième usage, lié au premier, est l'aide à l'apprentissage des notions incontournables en politique de développement économique territorial, c'est-à-dire la compréhension ou la prise de conscience de leur(s) signification(s), mais aussi des *parcours conceptuels* dans lesquels elles s'inscrivent ("branche" de rattachement, relations hiérarchiques…).

2. Spécificités du terrain

2.1. Experts professionnels, futurs utilisateurs

Dans le cas du projet CAVALA, une double expertise est attendue : celle concernant la décision publique, celle concernant le domaine du développement économique territorial. Dans le cadre du projet, les deux types d'expertises se sont avérées relativement limitées.

La culture du pilotage reste encore assez faible dans les Régions françaises. Tous les aspects du système de décision sont impactés : la formation des politiques (définition des orientations stratégiques, déclinaison en axes, puis en objectifs quantifiés, construction des jeux d'indicateurs), leur mise en œuvre sur le terrain, et enfin leur évaluation. De façon corollaire on observe, au sein de la Région, l'absence d'experts professionnels en matière de décision. La même situation prévaut en matière de développement économique territorial. L'expertise de la Région s'exprime avec un degré de précision restreint, un faible nombre de notions (y compris au niveau des instances), et une profondeur hiérarchique très réduite (deux à trois niveaux). Les pratiques sont par ailleurs marquées par un nombre important de normes implicites et par l'influence de différentes doxas (non spontanément convergentes), qui donnent lieu à des choix paradigmatiques en général non explicités.

Le constat du double manque d'expertise (sur la décision, sur le domaine), ainsi que celui de l'existence de doxas distinctes coexistant dans les pratiques, a conduit à certains choix méthodologiques (voir partie 3).

2.2. Le domaine du développement économique territorial

Le domaine du développement économique territorial ne semble pas stabilisé. Dans une perspective de construction d'ontologies, les problèmes liés à ce domaine peuvent être rapprochés de ceux que l'on rencontre dans le domaine juridique. (Bourcier *et al.*, 2004) évoquent ainsi "l'imprécision des définitions", l'existence de "catégories en attente de définition opérationnelle", et aussi l'écart entre la vision des chercheurs spécialisés et celle des acteurs de terrain, etc. Ce sont autant de problèmes que l'on rencontre à grande échelle dans le domaine du développement économique

territorial. Du point de vue scientifique, ce domaine est traversé d'écoles de pensées concurrentes, qui déterminent des représentations divergentes des principaux objets. Nous assimilerons ces écoles à des doxas. Enfin, le développement économique territorial est un domaine étroitement lié à la décision politique, et, en ce sens, les concepts utilisés et leur organisation sont susceptibles de traduire des choix amont de type politique (souvent implicites), qui sont une expression particulière des doxas. L'explicitation des doxas en jeu, autant qu'elle est possible, nous est apparue comme un enjeu essentiel d'un travail de construction d'ontologies dans ce domaine.

2.3. Corpus accessible

La Région Midi-Pyrénées (comme l'ensemble des Régions de France) a produit un Schéma Régional de Développement Économique (SRDE), qui expose sa politique économique pour les années à venir. Ce document public représente une source précieuse, bien qu'il reste exprimé à un niveau assez général. Nous avons pu également disposer d'un ensemble de textes non publics liés aux aides aux entreprises ou à la politique économique de la Région. Des entretiens ont été menés auprès d'élus régionaux en charge du développement économique, des membres et de la direction du service chargé à la Région d'évaluer les demandes d'aides des entreprises, ainsi que de divers acteurs régionaux périphériques impliqués dans le support financier aux entreprises. Les définitions produites par l'INSEE sur les catégories utilisées par la Région ont été également prises en compte. Notons qu'il s'agit là de la seule ressource de type terminologique disponible, les documents de la Région, SRDE inclus, ne comportant jamais de glossaire. En ce qui concerne le corpus scientifique, le choix a été fait de se limiter principalement aux productions des chercheurs de l'école de la proximité, dont est proche le LEREPS (laboratoire d'économie partenaire du projet CAVALA). Au total le corpus comporte très peu de textes d'aides à la mise en œuvre des actions ou plus largement d'ingénierie des politiques, ce qui est cohérent avec une culture encore à construire en matière de système de pilotage et de prise de décision.

2.4. Une forte majorité d'objets non physiques

Une caractéristique forte du domaine est que les notions utilisées par les acteurs du développement économique régional renvoient pour leur grande majorité à des objets non physiques : catégories, classifications, regroupements, etc.. C'est le cas des indicateurs, mais également des groupements d'activités (secteurs, filières…), des types de ressources du territoire, des catégories d'entreprises, etc. Or, en ingénierie des connaissances et dans d'autres disciplines, des auteurs ont montré qu'il n'existe pas de classifications *naturelles*, et que les classifications rencontrées dans les textes ont été construites à des fins particulières et/ou produites par des effets de doxa. A titre d'exemple, on peut citer (Hacking, 2006) pour les classifications scientifiques, (Boltanski, 1982) pour la catégorie des cadres, (Desrosières, 2003) pour les objets de la comptabilité publique, ou encore, plus globalement, (Rastier, 2004) sur "l'incidence des normes de la doxa" sur le discours. En IC, pour ne citer qu'un seul exemple, (Masolo *et al.*, 2003), s'intéressant aux objets physiques comme non physiques

précisent que "the categories refer to cognitive artifacts more or less depending on human perception, cultural imprints and social conventions". Si l'on admet avec (Charlet, 2003) que construire une ontologie, c'est "décider de quels objets existent", de la façon de les décrire, de les classer, et que la conception d'ontologies s'appuie sur une approche "constructiviste" (Aussenac-Gilles, 2006, Masolo *et al.*, 2004), il apparaît nécessaire, pour le terrain étudié, et du fait de l'imprécision des notions, de porter une attention toute particulière aux conventions sociales et politiques, empreintes culturelles (l'ensemble étant rassemblé ici sous le terme "doxa") en jeu dans les textes. Ne pas le faire entraînerait deux types de risques. Le premier risque serait de fixer dans l'ontologie un ensemble partiellement ou totalement incohérent du fait de visions du monde contradictoires non dévoilées. Le second risque serait que l'ontologie ne reflète qu'une seule de ces doxas, qui prendrait alors potentiellement valeur de norme. Ce second risque est bien entendu aggravé si l'ontologie n'exprime pas clairement qu'elle ne reflète qu'une seule doxa et n'explicite a fortiori pas laquelle.

2.5. Des objets à caractère prescriptif

Une autre caractéristique importante du terrain concerné est que certaines notions y ont un statut quasi prescriptif, dans un sens approchant celui de (Bachimont, 2004).

La désignation d'un secteur donné, par exemple le secteur "aéronautique", va rendre envisageables des actions centrées sur les entreprises réputées appartenir à ce secteur. A l'inverse, l'absence de désignation d'un secteur va le rendre *invisible* aux décideurs publics. C'est par exemple le cas de l'électronique embarquée, qui n'existe pas dans les nomenclatures utilisées par la Région, et ne fait donc l'objet d'aucune action spécifique, à la grande irritation des industriels concernés. La définition d'un indicateur de "taux d'emploi"[4] en remplacement du taux de chômage[5] (*cf.* "stratégie de Lisbonne") va orienter les actions vers l'amélioration de ce premier taux et non pas du second, l'évolution du taux mesuré renforçant à son tour l'orientation des actions. On retrouve ici l'aspect *performatif* des normes évoqué par (Bourcier *et al.*, 2006), ou celui des nomenclatures dont (Boydens, 1999) rappelle qu'elles sont toujours "historiquement et socialement situées", et ont des "effets performatifs [qui] s'inscrivent sur le réel ainsi normé". Le caractère performatif de certains concepts doit conduire à prendre des précautions particulières dans la construction de l'ontologie.

Le cas particulier des catégories importées

Dans les textes et les actions de la Région, on trouve beaucoup de catégories provenant de l'extérieur de la Région. C'est en particulier le cas des groupements d'activités (secteurs) ou des découpages de l'espace régional produits par l'INSEE, des découpages administratifs nationaux ou européens... Compte tenu du caractère performatif que l'on vient d'évoquer, et dans une perspective de cohérence de l'ontologie, il sera utile de vérifier quelles conventions ou quelles doxas fondent ces catégories "exogènes".

[4] Une personne est considérée comme ayant un emploi si elle occupe un ou plusieurs emplois, indépendamment du temps consacré à cet emploi.

[5] Un chômeur est une personne sans emploi.

L'ensemble de ces caractéristiques a rendu très problématique l'utilisation des outils méthodologiques "classiques" en conception d'ontologies, et nous a conduits à développer des outils spécifiques ainsi qu'à apporter des aménagements à des méthodologies existantes.

3. Méthodologie

3.1. Type d'ontologie

Compte tenu du caractère exploratoire du projet CAVALA, de la complexité du contexte, et surtout des usages prévus, il a été décidé de construire une ontologie légère, formalisant des concepts liés par des relations de subsomption (Est-un) et, marginalement, de méronymie (Est-une-partie-de) et de quelques relations ad hoc (Est-localisé-dans). Cette ontologie comprend pour sa première version un nombre limité de termes. Certains concepts "parataxiques" au sens de (Bachimont, 2004) n'ont pas été inclus dans leur totale extension possible (ex. : la liste des secteurs, ou celle de certains territoires, ont été limitées à quelques représentants significatifs). Cet ensemble de concepts suffit largement pour les premiers usages de l'ontologie tels qu'ils ont été évoqués en partie 1.

3.2. Les trois niveaux de l'expression de la politique économique : normes, principes, doxas

La première étape du projet, l'analyse du besoin, a révélé la nécessité d'une clarification générale de l'expression de la politique économique (grandes orientations, déclinaison en objectifs, actions…). La définition d'indicateurs d'évaluation des actions est en effet tributaire de la précision et de la cohérence des objectifs qu'ils doivent évaluer. Un premier travail a donc été mené sur les textes de la Région présentant sa politique économique, et en particulier sur le SRDE. Compte tenu des spécificités évoquées en partie 2, les outils méthodologiques dont nous avions connaissance nous ont paru peu adaptés pour ce premier traitement. Une grille d'analyse ad hoc a donc été utilisée (Salles, 2007, Salles & Colletis, 2007), qui distingue trois niveaux principaux dans l'expression de la politique économique.

1) Le niveau des *normes* : pour mettre en œuvre concrètement les décisions, les acteurs opérationnels de la Région doivent disposer de procédures, de normes ou de référentiels précis. Un membre du service qui instruit un dossier d'aide pour une entreprise doit, par exemple, pouvoir suivre une procédure d'instruction précise, se référer à une liste de secteurs industriels prioritaires, à un ensemble de types de projets à soutenir, à un jeu de critères d'éligibilité, etc. C'est au niveau des normes que se situent les indicateurs d'évaluation. Ce niveau est le plus concret et aide en général à produire le dernier ou les deux derniers niveaux de l'ontologie. Les normes ne sont cependant en général pas intégrables telles quelles dans l'ontologie.

2) Le niveau des *principes* : il exprime le cadre conceptuel des normes. A ce niveau sont exprimées des typologies, des catégories, des logiques d'action, etc. Dans le cas des politiques économiques, ces principes vont par exemple concerner les principes de regroupement des activités (ex. : méthode pour définir un secteur), ceux des découpages infrarégionaux du territoire, ou encore la logique générale du système d'aides aux entreprises (simple soutien ou à l'inverse incitation[6]), etc. Ce niveau aide à construire les niveaux médians de l'ontologie. Certains principes peuvent être directement traduits en logique pour définir les hiérarchies, mais d'autres, plus nombreux, demandent un traitement pour être intégrés dans l'ontologie, sous forme de logiques, de concepts, de relations…

3) Le niveau des *doxas* : un troisième niveau s'est avéré nécessaire pour rendre compte de la structure de la connaissance exprimée dans les textes de la Région. En effet, si les *normes* découlent assez naturellement des *principes*, ces derniers sont potentiellement en nombre élevé. La question se pose alors de ce qui détermine le choix de tel principe plutôt que de tel autre. En matière de politique économique, il n'y a pas de corps de connaissance qui offrirait les concepts de haut niveau totalement consensuels à partir desquels inférer les catégories intermédiaires. On peut d'ailleurs douter qu'une telle situation existe pour un quelconque domaine. Le niveau des *doxas* correspond aux représentations ou visions du monde sous-jacentes aux principes (et donc également aux normes), aux choix paradigmatiques, et, dans le contexte du projet, aux grandes options politiques. Dans le cas des politiques économiques territoriales, les choix faits à ce niveau vont prescrire la totalité des niveaux inférieurs (objectifs, actions…, indicateurs d'évaluation). Ces choix sont très rarement explicités, et sont le plus souvent à reconstruire à partir des textes. A titre d'exemple de doxas dans le domaine, on peut citer la vision du social et de l'économique (qui les dissocie, ou au contraire les considère indissociables), la vision de ce qu'est un territoire (un espace délimité par des frontières, ou, à l'opposé, un lieu de coordination révélé à la faveur d'un projet), la vision de la technologie (ressource préexistante transférable en l'état, ou à l'inverse résultat d'un processus d'innovation impliquant la coordination de plusieurs types d'acteurs), etc. Le niveau des doxas détermine tous les autres niveaux de l'ontologie (sans qu'il s'agisse d'un rapport de subsomption, *cf.* 3.3.).

3.3. La prise en compte du niveau des doxas dans l'ontologie

Le niveau des doxas n'est en général pas traité dans la conception d'ontologies, si ce n'est sous forme de choix de départ consistant à décider de traiter une école de pensée donnée, à l'exclusion des autres. (Falquet & Mottaz, 2001) expliquent ainsi : "Si plusieurs écoles de pensées (…) s'affrontent à l'intérieur du domaine, nous pensons qu'il est préférable de les traiter comme des domaines séparés" et précisent plus loin qu'ils utilisent "le terme 'domaine' pour désigner une *vision d'un domaine*". Comme indiqué dans l'introduction, notre posture est opposée, en ce que notre hypothèse est qu'une même ontologie peut intégrer plusieurs doxas. Dans les faits, le discours des acteurs de la Région est bien porteur d'une pluralité de doxas. Ce niveau n'est pas l'équivalent d'une

[6] Dans une logique d'incitation, l'aide vise à infléchir le comportement de l'entreprise.

top ontologie, car, comme dans le cas d'une "ontologie catégoriale" (Nadah *et al.*, 2008), les concepts de l'ontologie ne spécialisent pas une doxa, mais s'expriment dans le contexte de la doxa, ou sont produits par elle. Les doxas, selon nous, ne sont pas non plus assimilables, de façon simple, à des "points de vue". En effet, des points de vue multiples sont en général des points de vue sur une *même* entité, qui est supposée dotée d'une existence propre, indépendante des divers points de vue. Dans cette logique, chaque point de vue correspond à "une famille de caractéristiques de l'entité" (souvent une vue métier) (Cahier *et al.*, 2004), ou bien est un composant d'un objet composite (Stefik & Bobrow, 1985[7], cités par Bach, 2006), ou encore organise un même ensemble de concepts d'une façon qui lui est propre. Les doxas interviennent de façon beaucoup plus globale. Un objet reconnu par une doxa peut voir remis en cause par une autre doxa son périmètre, sa nature même, ou encore son existence. Mais, plus largement, une doxa va déterminer les concepts de haut niveau dans l'ontologie et les types de principes pour construite les niveaux inférieurs. (Falquet & Mottaz Jiang, 2001) situent les points de vue dans une hiérarchie à trois niveaux : le domaine, la vision (conceptualisation), et le point de vue (un domaine peut donner lieu à plusieurs visions, une vision à plusieurs points de vue). Dans cette hiérarchie, les doxas se situent bien au niveau des visions.

A titre d'exemple, il existe en économie deux doxas très globales : l'une développe une représentation statique de l'économie (qui induit un univers d'équilibre, de ressources existantes, d'allocation optimale de ressources, de rareté, etc.), l'autre une représentation dynamique (qui induit un univers de développement, de création de ressources, de redéployabilité des ressources, et donc de caractère potentiellement infini de celles-ci, etc.). Ces deux doxas s'expriment au niveau de la conception du territoire. Pour la première doxa le territoire est posé comme existant (postulé), il est vu comme un espace doté, et délimité par des frontières. La seconde considère le territoire comme devant être créé à la faveur d'un projet coordonnant des acteurs proches spatialement, projet et territoire étant alors indissociables (un territoire ne pouvant exister qu'au travers d'une suite de projets). Notons que dans les catégories de DOLCE (Masolo *et al.*, 2003), la première doxa considère le territoire comme un objet physique (un *endurant*), quand la deuxième le voit comme un processus (un *perdurant*). Au niveau des politiques économiques, ces deux visions du territoire induisent des modes d'action totalement différents, voire opposés (Salles et Colletis, 2008). L'action politique (aides notamment) sera centrée sur l'entreprise pour la première doxa, sur les interrelations entre acteurs pour la seconde.

3.3.1. L'identification des doxas

Les doxas ont été identifiées par l'expertise des chercheurs en économie impliqués dans le projet, et par une analyse des textes "manuelle" mais soutenue par l'utilisation de Syntex (Bourigault & Fabre, 2000). Sans entrer dans le détail de la démarche, notons que certaines modes d'expression nous ont semblé caractéristiques, comme, par exemple des affirmations non suivies ni précédées d'argumentation (ex. "Il n'y a

[7] Stefik, M. & Bobrow, D.G. (1985), Object-oriented programming: themes and variations, *The Al Magazine*, 6, 4 40-62.

pas de compétitivité économique s'il n'y a pas d'attractivité des territoires concernés"[8]), mais qui donnent lieu à toute une série de choix politiques exposés dans la suite du texte. Les doxas n'ont cependant pas été extraites à l'aide de méthodes linguistiques du type de celles décrites par (Rastier, 2004) ou a fortiori (Malrieu, 1995), même si le support de Syntex a été précieux.

L'organisation de la démarche s'est inspirée de la méthode Archonte proposée par (Bachimont, 2000), qui distingue quatre étapes dans la construction d'une ontologie. La première étape d'identification des termes au sein des textes permet de recueillir ceux-ci dans leur contexte, le sens des termes est donc contextuel. La deuxième étape dite de normalisation sémantique, consiste à passer d'un sens contextuel à un sens a-contextuel en produisant une "ontologie différentielle". Cette deuxième étape est conduite en construisant une hiérarchie des concepts qui respecte quatre "principes différentiels" : *i)* le principe de propriété commune avec le parent, *ii)* le principe de différence avec le parent, *iii)* le principe de propriété commune entre les, *iv)* le principe de différence entre les frères. La troisième étape formalise l'ontologie différentielle et produit une "ontologie référentielle", qui, lors de la dernière étape, est codée dans un langage et donne lieu à une "ontologie computationnelle".

Dans le cadre du projet CAVALA, la phase de normalisation sémantique a été précédée d'une étape d'identification des doxas, pendant laquelle s'est également faite, de façon itérative, l'extraction des unités linguistiques. Ce dernier travail a été mené à l'aide principalement de Syntex et Terminae (Szulman *et al.*, 2002).

3.3.2. La normalisation sémantique

La prise en compte des doxas permet de préserver une part du sens contextuel qui est par définition perdu lors de la normalisation sémantique. Il a été nécessaire de traiter quelques problèmes liés à la présence simultanée de plusieurs doxas La normalisation sémantique doit donc être précédée d'une étape de normalisation inter-doxas, qui va déterminer certains choix de structuration. Cette normalisation doit s'appuyer sur une typologie des différentes configurations inter-doxas et des modes de traitements correspondants. Les doxas peuvent par exemple être totalement séparables (le seul lien est alors un concept générique, et éventuellement des relations ad hoc entre certains concepts), partager certains concepts (avec des définitions identiques ou non, avec des rattachements différents ou non, etc.), avoir des fragments de hiérarchies en commun, etc. Le projet a permis d'identifier quelques cas typiques de relations entre doxas, mais le travail reste pour l'essentiel à mener.

Le travail de normalisation sémantique (au sens strict d'Archonte) a été mené en respectant autant qu'il était possible les quatre principes différentiels. Le travail sur les libellés (noms) des concepts reste à affiner, de trop nombreux concepts étant exprimés par des expressions relativement longues. La phase de formalisation est en cours. Compte tenu des usages prévus et de la progressivité de la mise en place des indicateurs, elle a vocation à être très simplifiée dans cette première version de l'ontologie.

[8] Cette affirmation (extraite du SRDE), non argumentée, et qui donne une place centrale à l'attractivité, est réfutée par certains chercheurs.

4. Illustrations

Nous présentons dans cette section des extraits simplifiés de la partie de l'ontologie consacrée à la notion de "territoire", selon trois étapes de sa construction. Cet extrait renvoie à une configuration assez simple, dans le sens où le travail de normalisation inter-doxas a permis relativement aisément de constituer au final deux branches distinctes, chacune correspondant à une doxa. La figure 1 montre un extrait construit à partir des concepts issus des textes et des entretiens (première étape), qui sont marquées par la doxa qui considère le territoire comme un espace délimité par des frontières (doxa 1). Cette vision du territoire est dominante dans les catégories importées (*cf.* partie 2), notamment de l'INSEE. On note cependant la présence de la notion de "territoire en émergence" qui se distingue de l'ensemble. Cette notion est ressentie par les acteurs comme signifiante et importante, mais sans qu'une définition ni une liste de tels territoire aient été produites. En s'inspirant de la typologie de (Biebow et Szulman, 2000) on pourrait dire qu'il s'agit en quelque sorte d'un concept pré-préterminologique.

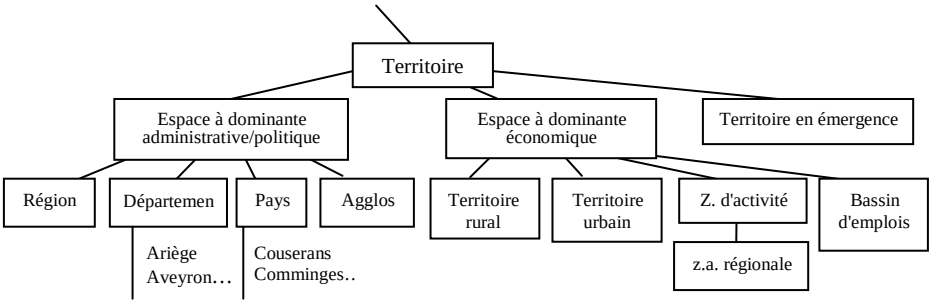


Fig. 1 - Concepts issues des textes et des entretiens (extrait)

Nous ne présentons pas ici d'extrait concernant les concepts issus de la recherche (école de la proximité, voir notamment Colletis, 2009), qui proviennent de la doxa qui voit le territoire comme un processus de concentration d'acteurs en interrelations (2ème étape). La figure 2 intègre ces concepts.

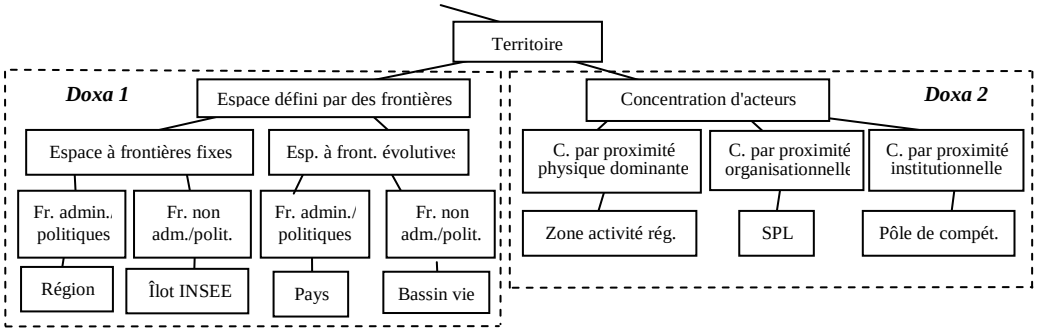


Fig. 2 - Ontologie intégrant les deux doxas (extrait)

La figure 2 montre un extrait de l'ontologie après la 3ème étape, qui consiste en un double travail d'une part d'intégration des doxas 1 et 2, d'autre part de normalisation sémantique. Les notions de territoires urbain ou rural sont abandonnées comme telles. Le caractère urbain ou rural devient une propriété. La notion de "territoire en émergence" est de fait intégrée dans la doxa 2, car pour cette doxa, rappelons que le territoire est une construction.

Conclusion

Dans ce papier, nous avons présenté de premiers résultats du projet CAVALA, d'aide à la décision publique en matière de développement économique territorial. Ce projet a un double caractère, opérationnel et aussi exploratoire. Notre travail montre que dans certaines configurations applicatives (et sans doute plus largement dans certains domaines), le niveau des doxas (représentations, visions du monde, choix paradigmatiques…) doit être explicité comme tel et traité spécifiquement. Dans la poursuite de son objectif d'aide à la décision, le projet CAVALA a conduit à la conclusion que la construction de l'ontologie devait intégrer plusieurs doxas, et mener ainsi à une ontologie polydoxique. Un extrait de cette ontologie a été proposé, centré sur le concept de "territoire". La conception d'ontologies polydoxiques soulève des problèmes méthodologiques spécifiques, dont nous avons donné un aperçu ici. Le travail se poursuit sur ces questions, notamment sur les conditions et modes d'identification des doxas, sur les types de liens entre les trois niveaux des normes, des principes et des doxas, ainsi que sur une typologie des différentes configurations inter-doxas et les modes de traitements associés.

Références

AUSSENAC-GILLES N. (2006), *Méthodes ascendantes pour l'ingénierie des connaissances*, Mémoire d'Habilitation à diriger des recherches, Université Paul Sabatier (Toulouse III).

BACH T. L. (2006)**,** *Web sémantique multi points de vue*, Thèse, Université de Nice-Sophia Antipolis.

BACHIMONT B. (2004), *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*, Mémoire d'Habilitation à diriger des recherches, Université de Compiègne.

BACHIMONT B. (2000), Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In *Ingénierie des Connaissances. Évolutions récentes et nouveaux défis*, Charlet J., Zacklad M., Kassel G. & Bourigault D. (éds.), Eyrolles, Paris.

BIÉBOW B., SZULMAN S. (2000), Une approche terminologique pour catégoriser les concepts d'une ontologie, In *Ingénierie des connaissances, évolutions récentes et nouveaux défis*, Charlet J., Zacklad M., Kassel G. & Bourigault D. (éds.), Eyrolles, Paris.

BOELLA G., van der TORRE L.W.N. (2006), A foundational ontology of organizations and roles. In Baldoni, M., Endriss, U., (eds.): *DALT. Volume 4327 of Lecture Notes in Computer Science*., Springer, pp. 78-88.

BOLTANSKI L. (1982), *Les cadres. La formation d'un groupe social*, Paris, Minuit.

BOURCIER D., DULONG DE ROSNAY M., LEGRAND J. (2006), Susciter la construction interdisciplinaire d'ontologies juridiques : bilan d'une expérience, *Semaine de la connaissance (SDC 2006)*, Nantes, 26 au 30 juin.

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151 Université Toulouse le Mirail

BOYDENS I. (1999), *Informatique, normes et temps*, Bruylant, Bruxelles.

CAHIER J.-P., ZAHER L.H., LEBOEUF J.P., PÉTARD X., GUITTARD C., Une expérience de co-construction de "carte de thèmes" dans le domaine des logiciels libres, *Colloque En route vers Lisbonne*, Luxembourg 12-13 octobre 2004.

CHARLET J., *L'ingénierie des connaissances. Développements, résultats et perspectives pour la gestion des connaissances médicales*, Mémoire d'Habilitation à diriger des recherches, Université Pierre et Marie Curie, version complétée, 2003.

COLLETIS G., Local Development, Proximities & Productive Encounters: The Case of Development Dynamics in the Region of Toulouse, *Canadian Journal of Regional Science*, n°32, 2009 (à paraître).

DESROSIÈRES A. (2003), Du réalisme des objets de la comptabilité nationale, *Congrès de l'Association Française de Sciences Économiques*, Paris, septembre.

FALQUET G., MOTTAZ JIANG C.-L. (2001), Navigation hypertexte dans une ontologie multi-points de vue, in *Proc. NimesTIC'01 conference*, Nîmes, 12-14 décembre.

HACKING I. (2006), *Cours "B" : Les choses, les gens et la raison*, Collège de France, Mai.

MALRIEU J.-P. (1995), La cohérence idéologique du discours, une méthode d'estimation, *Intellectica*, 1, 20, pp.185-215.

MASOLO C., VIEU L., BOTTAZZI E., CATENACCI C., FERRARIO R., GANGEMI A., GUARINO N. (2004), Social roles and their descriptions. In D. Dubois, C. Welty, & M.-A. Williams (Eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004)*, Menlo Park: AAAI Press, pp. 267-277.

MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. (2003), *WonderWeb Deliverable D18, Ontology Library (final)*, IST Project 2001-33052 WonderWeb, December.

MONTEIL J.-M., BAVENT L. et LACASSAGNE M.-F. (1986), Attribution et mobilisation d'une appartenance idéologique : un effet polydoxique, *Psychologie française*, 1986, 31.

NADAH N., CHARLET J., BANEYX A., BACHIMONT B. (2008), Ontologies catégoriales : motivations et usages, *19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, Session posters.

RASTIER F. (2004), Doxa et lexique en corpus - pour une sémantique des idéologies, *Texto !*, Décembre.

SALLES M., COLLETIS G. (2008), How to deal with the conflicting views of the world expressed in regional economic development policies? *International Conference of Territorial Intelligence*, Besançon (France), 16-17 Octobre.

SALLES M. (2007), Présentation du dossier Représentations, modèles et normes pour l'entreprise, *Droit et Société*, n° 65.

SALLES M., COLLETIS G. (2007), Représentations de l'entreprise dans les systèmes d'information statistique et décision dans les collectivités territoriales, *Droit et Société*, n° 65.

SZULMAN S., BIÉBOW B., AUSSENAC-GILLES N., Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE, *Revue TAL (Traitement Automatiques des Langues)*, Volume 43, n°1/2002.

Vers une ontologie formelle des artefacts[†]

Gilles Kassel

Laboratoire MIS, Université de Picardie Jules Verne
gilles.kassel@u-picardie.fr

Résumé : Dans cet article, nous jetons les bases d'une ontologie formelle permettant de rendre compte de la nature générale des artefacts. L'objectif visé par une telle ontologie est d'aider à structurer des ontologies d'application dans des domaines où des artefacts spécifiques sont présents, autrement dit pratiquement tout domaine d'activité ! La conceptualisation s'appuie sur une littérature philosophique récente consacrée aux artefacts, que nous exploitons pour éclairer les choix de modélisation. L'ontologie étend par ailleurs l'ontologie formelle DOLCE, en venant compléter son axiomatisation. Les primitives conceptuelles introduites sont celles d'*entité artificielle*, de *production intentionnelle d'objets*, de *capacité* à exercer un *rôle* dans des actions d'un type donné, de *fonction* et d'*entité fonctionnelle*. Ces primitives permettent de caractériser les artefacts comme des entités intentionnellement produites, auxquelles une fonction est attribuée.

Type de la communication : Recherche
Thèmes : Ontologie d'artefacts, Ontologies formelles, Construction d'ontologies d'application.

1 Motivations

Dans le cadre du projet NeuroLOG[1], une ontologie est développée come composant d'une plate forme logicielle destinée à permettre à une communauté de chercheurs en neuroimagerie de mutualiser des ressources, à savoir des images et des logiciels de traitement d'images (Temal *et al.*, 2006). L'approche suivie pour la construction de cette ontologie consiste à la structurer en sous-ontologies situées à différents niveaux d'abstraction. Schématiquement, trois niveaux sont distingués : au niveau le plus abstrait, l'ontologie formelle[1] DOLCE[2] (Masolo *et al.*, 2003) apporte

[†] Ce travail est en partie financé dans le cadre du projet NeuroLOG (ANR-06-TLOG-024) du programme Technologies Logicielles de l'Agence Nationale de la Recherche : http://neurolo.polytech.unice.fr.

[1] Le terme « formel » est à entendre ici dans le sens opposé à « régional ». Il indique que la conceptualisation est abstraite et qu'elle permet de rendre compte de la nature d'objets particuliers existant dans les différents domaines d'activité. Dans un sens complémentaire, il indique que l'ontologie est spécifiée dans un langage logique muni d'une sémantique de type théorie des modèles, ce qui est le cas de DOLCE et de notre ontologie des artefacts.

[2] http://www.loa-cnr.it/DOLCE.html

un ensemble de concepts et de relations abstraits censés permettre de structurer, par spécialisation, la conceptualisation de n'importe quel domaine ; à un niveau médian, des ontologies « noyaux » de domaines définissent des concepts génériques et centraux dans différents domaines, notamment le domaine des images médicales (Temal *et al.*, 2008) et le domaine des programmes et des logiciels (Lando *et al.*, 2007) ; enfin, au niveau le plus spécifique, les ontologies précédentes sont à leur tour spécialisées pour définir des concepts plus concrets, respectivement dans le domaine de la neuroimagerie et celui des outils de traitement d'images.

L'expérience acquise par la construction de cette ontologie montre que les concepts très abstraits apportés par une ontologie comme DOLCE ne sont pas directement utilisables pour définir des concepts génériques de domaines comme ceux d'*image médicale* ou de *programme informatique*. En l'occurrence, les images médicales, tout comme les programmes informatiques, sont des artefacts ayant un auteur et une fonction, or les notions d'*artefact*, d'*auteur* et de *fonction* sont absents de DOLCE. Élaborer une telle ontologie d'application nécessite donc d'introduire ces concepts d'un niveau intermédiaire.

Une analyse de l'existant concernant le traitement des artefacts dans les ontologies montre que des concepts généraux de ce domaine sont présents dans des ontologies de haut niveau comme OpenCyc[3] ou SUMO[4], cependant les principes sous-tendant leur structuration ne sont pas explicités et les concepts sont définis sans référence à la littérature sur ce domaine. Une telle situation nous a motivé à définir une ontologie formelle d'artefacts qui spécialise l'ontologie DOLCE et qui soit fondée sur la littérature récente sur le domaine[5].

L'article suit le plan suivant : nous partons de la notion philosophique courante d'artefact et la confrontons à la littérature récente dans le domaine, ce qui nous conduit à définir les artefacts comme étant à la fois des entités artificielles intentionnellement produites et des entités auxquelles une fonction est attribuée ; nous présentons ensuite une vue d'ensemble de notre cadre ontologique de référence (DOLCE) et montrons comment nous l'avons étendu pour définir une ontologie formelle d'artefacts ; nous comparons alors notre proposition à un travail en cours similaire, celui de Borgo & Vieu (2006, 2008) ; enfin, nous concluons en dégageant des pistes de travail pour évaluer et étoffer notre ontologie.

2 Vers une notion d'artefact

Les dictionnaires courants recensent deux notions principales pour le terme « artefact » :

- Un objet fait par l'homme (ex : une arme, un ornement),

[3] http://www.opencyc.org

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

[5] Les premières pierres à cet édifice ont été posées dans l'article (Kassel *et al.*, 2007), montrant l'apport des notions introduites pour définir le concept de *programme informatique*. L'objet du présent article est de présenter une version plus élaborée de cette ontologie.

- Un résultat expérimental étranger au phénomène naturel étudié et qui est dû au cadre expérimental même (ex : une ombre sur une image de poumons se révélant être due à la technique d'imagerie médicale utilisée).

En philosophie, un artefact est communément défini comme une « entité intentionnellement faite ou produite pour une certaine raison » (Hilpinen, 2004). La notion philosophique se veut plus précise que la notion ordinaire des dictionnaires, en mettant en avant deux propriétés nécessairement vérifiées par tout artefact : ils sont « intentionnellement produits » et « pour une certaine raison ». Dans la section suivante, nous commençons par analyser la première propriété.

2.1 Entités artificielles et intentions

Les notions liées d'*action* et d'*intention* ont fait l'objet de nombreux travaux en philosophie depuis le début des années 70. À la suite de John Searle, les philosophes contemporains distinguent schématiquement deux types d'intention, que (Pacherie, 2000), reprenant la terminologie de Searle, nomme « intention préalable » et « intention en action ». L'*intention préalable* suppose une planification de l'action et une représentation de son but, tandis que l'*intention en action* relève du guidage et du contrôle de l'action tout au long de son exécution (Pacherie, 2000). Dans notre contexte, le terme « intentionnel » est à prendre au sens d'une intention préalable. Il implique que l'artefact corresponde à un résultat visé par son créateur. En conséquence, tout artefact possède un auteur. La notion philosophique épouse ainsi, en l'étendant, la première notion ordinaire d'*artefact visé*.

La seconde notion recensée par les dictionnaires, celle d'*artefact expérimental*, fait référence, a contrario, à une entité créée non intentionnellement, autrement dit qui n'est pas visée. L'existence de telles entités amène à considérer une classe des entités *artificielles*, autrement dit des entités produites comme conséquence d'une activité humaine (par opposition aux entités *naturelles*), qui est plus large que la classe des artefacts. Parmi celles-ci figurent des entités correspondant à des « effets de bord » non désirés d'actions intentionnelles (ex : les artefacts expérimentaux, de la sciure de bois, des cheveux coupés, de l'herbe tondue) ou aux effets d'une série d'actes intentionnels non coordonnés et ne pouvant compter pour une intention collective (ex : suivant l'analyse de (Hilpinen, 1992), reprise par (Thomasson, 2003), un chemin qui résulterait d'une série d'actes intentionnels consistant à emprunter le plus court chemin à travers champ entre deux emplacements ne peut être considéré comme un artefact).

Suivant une définition stricte de l'*intention*, les artefacts sont donc à considérer comme des entités artificielles (préalablement) intentionnellement produites et possédant un auteur. Nous en venons à la raison pour laquelle l'artefact est produit.

2.2 Compétences et fonctions

Un artefact est produit pour permettre à son auteur, ou une autre entité, d'effectuer quelque chose, autrement dit de réaliser une *action* : c'est là sa *fonction*. Plusieurs théories de fonctions ont été proposées dans la littérature, notamment en

philosophie et en ingénierie, cependant aucune n'est largement admise et la plupart (en philosophie) concernent la fonction naturelle d'entités biologiques et de leurs parties. Nous préférons donc proposer une notion de fonction qui soit adaptée à la définition des artefacts. Comme nous venons de convoquer la notion d'*action*, nous continuons à nous placer dans une théorie de l'action. Pour souligner la nature de *capacité* que nous conférons à la fonction, nous définissons celle-ci en dressant un parallèle avec la notion de *compétence* et celle d'*agent*, couramment définies en Intelligence Artificielle comme suit :

- Une *compétence* est la capacité à réaliser une action,
- Un *agent* est une entité ayant la capacité à réaliser une action.

Le terme « capacité » indique un potentiel qui est exploité lors de la réalisation d'une action et rend cette dernière possible. En tant que potentiel, la compétence est à distinguer de l'action à laquelle elle se rapporte et du fait que cette action réussisse ou échoue, donc du fait que le résultat visé existe ou pas. À noter que l'action à laquelle il est fait référence correspond à un type d'action plutôt qu'à une action individuelle, la capacité à réaliser une action supposant la capacité à répéter cette action. La compétence est attribuée à une entité, l'*agent*. Ce potentiel pouvant être acquis puis perdu, il doit être distingué d'une *qualité* qui, à l'instar de la masse ou de la couleur d'un objet physique, est inhérente à l'entité.

Concernant la notion d'*agent* : le terme « agent » est couramment utilisé selon deux sens principaux, celui d'un rôle temporel joué par une entité lors d'une action individuelle – ce rôle est appelé dans la littérature « rôle de participation » ou « rôle thématique » (Sowa, 2000) -, et celui, que nous venons d'introduire, d'une entité à laquelle est attribuée une compétence. Pour distinguer ces deux notions, nous emploierons dorénavant les termes « agent » (pour le rôle) et « agentive » (pour le statut). Ces deux notions sont liées : un *agentive* est une entité à laquelle quelqu'un attribue un statut, la capacité à pouvoir jouer le rôle d'*agent* dans des actions.

Par analogie, nous proposons pour les notions de *fonction* et d'*entité fonctionnelle*, d'adopter les définitions suivantes :

- Une *fonction* est la capacité à permettre de réaliser une action,
- Une *entité fonctionnelle* est une entité ayant la capacité à permettre de réaliser une action.

Cette notion de fonction rejoint celle proposée par les théories de fonctions que Kroes et Meijers (2006) qualifient d'intentionnelles :

> "Intentional theories take as their starting point that agents ascribe functions to artefacts by embedding them in means-ends relations. Objects and their parts 'have' functions only insofar as they contribute to the realization of a goal".[6]

Nous noterons également la proximité de notre notion de fonction avec celle de *function as effect* proposée en ingénierie par (Chandrasekaran & Josephson, 2000) :

[6] On peut toutefois noter que notre définition est plus stricte, dans la mesure où nous subordonnons la fonction à la réalisation d'une action : nous assimilons en effet le « but » à celui des actions que l'artefact permet à son utilisateur de réaliser.

> "All the meanings for the term "function" arise from the idea of a machine, system or a person *doing* something or having a property that is *intended* or *desired* by someone, or deemed as appropriate from someone's point of view. Thus the ontology of function starts with the ontology of behavior, but it is distinguished by the fact that some agent regards it as desirable or intends the behaviour. All the other terms – structure, behaviour, causal models – are neutral with respect to intent […] Thus a central meaning of function is *function as (desired) effect*."

Ces auteurs considèrent la notion de *comportement* comme première. Bien que notre ontologie ne fasse pas de place à cette notion, nous retrouvons toutefois une notion proche avec les couples (rôle, action) représentant des *manières de participer à des actions*. En assimilant un comportement à un tel couple, nous pouvons voir que la distinction entre nos notions de *fonction* et de *compétence* repose sur une distinction entre deux comportements : capacité à participer en tant qu'*agent* pour la *compétence* ; capacité à participer en tant qu'*instrument* pour la *fonction*. Le rôle d'instrument, que traduit la présence du verbe « permettre » dans la définition, est à prendre au sens large d'*apporter une aide à un agent pour la réalisation d'une action*.

Concernant finalement la notion d'*entité fonctionnelle* : de même que nous avons distingué supra le rôle temporel de participation d'*agent* et le statut d'*agentive* attribué à une entité, nous retrouvons le rôle d'*instrument* et le statut d'*entité fonctionnelle*. Munis des notions de *fonction* et d'*entité fonctionnelle*, nous pouvons à présent compléter notre notion d'*artefact*.

2.3 Entités fonctionnelles et artefacts

Nous en étions restés à caractériser un artefact comme une entité intentionnellement produite. Considérant qu'un artefact est produit pour une certaine raison, nous ajoutons qu'un artefact est nécessairement une entité fonctionnelle.

Cet ajout revient à considérer que la production intentionnelle d'un artefact s'accompagne d'un acte mental d'attribution d'une fonction à cet artefact. Suivant la théorie « intentionnelle-historique » des concepts d'artefacts, proposée par le psychologue Paul Bloom (1996), un artefact est toujours conçu par rapport à un type existant[7]. En conséquence, il « hérite » de la fonction communément attribuée à ce type d'artefact, qu'il conserve aussi longtemps qu'il existe en tant que membre du type d'artefact. En ce sens, il est possible d'affirmer, à l'instar de Lynne Rudder Baker (2004), que les artefacts sont *essentiellement* fonctionnels. En ce sens également, on peut considérer que les artefacts ont une nature duale étant à la fois des objets physiques et des objets fonctionnels (Kroes & Meijers, 2006).

À noter, pour distinguer nos notions d'*entité fonctionnelle* et d'*artefact*, qu'à l'inverse, une entité fonctionnelle n'est pas nécessairement un artefact, ni même une entité artificielle. Comme l'indiquent Borgo & Vieu (2006), un objet naturel comme un galet peut très bien être considéré comme un presse-papier. De même, un artefact

[7] (Bloom, 1996, p. 10) : "We construe the extension of artifact kind *X* to be those entities that have been successfully created with the intention that they belong to the same kind as current and previous *X*s. "

déjà conçu, par exemple une agrafeuse, peut se voir attribuer une fonction différente de celle de son type, à savoir celle de presse-papier (faute de mieux !).

En résumé, nous assimilons un artefact à une entité produite intentionnellement comme faisant partie d'un certain type d'artefact et se voyant par là même attribuer la (ou les) fonction(s) communément attachée(s) à ce type d'artefact. Par ailleurs, comme tout type d'entité (naturelle ou artificielle), d'autres fonctions peuvent lui être attribuées. Nous allons par la suite proposer une formalisation logique pour cette conceptualisation. Auparavant, nous présentons notre cadre conceptuel de référence en introduisant les primitives de l'ontologie DOLCE.

3 Cadre ontologique de référence

Suivant des principes philosophiquement fondés, le domaine de DOLCE (Masolo *et al.*, 2003) – les *Particulars* (PT)[8] – est partitionné en quatre sous-domaines :

- Les *Endurants* (ED) sont des entités « endurant dans le temps » (ex. : le présent article). Parmi les *Endurants* sont distingués les *Physical Objects* (POB) et les *Non-Physical Objects* (NPOB), les premiers étant les seuls à posséder des qualités spatiales directes. Le domaine des *Non-Physical Objects* recouvre le domaine des entités sociales (ex. : la communauté francophone des chercheurs en Ingénierie Ontologique) et des entités cognitives parmi lesquelles figurent les *Concepts* (CPT) (ex. : votre notion d'ontologie) réifiant des types d'instances.
- Les *Perdurants* (PD) sont des entités se « déroulant dans le temps » (ex. : votre lecture de l'article). Parmi les *Perdurants* sont distingués, suivant un principe de cumulativité, les *Statives* et les *Events*[9]. Parmi ces derniers, suivant qu'ils sont atomiques ou non, les *Achievements* sont distingués des *Accomplishments*. Finalement, au sein des *Accomplishments*, les *Actions* (ACT) « exemplify the intentionality of an agent » : il s'agit ainsi d'*Accomplishments* qui sont contrôlés par un agent[10].
- *Endurants* et *Perdurants* ont des *Qualities*, que nous percevons et/ou mesurons (ex. : le poids de la copie papier de l'article entre vos mains, la durée de votre lecture de l'article).
- Ces *Qualities* prennent une valeur (*Quale*) dans des régions de valeurs qui sont des *Abstracts* (ex. : 25 grammes, 20 minutes).

[8] Les ontologies présentées dans l'article n'existant qu'en langue anglaise, nous conservons les étiquettes anglaises pour nommer les entités conceptuelles. Ces étiquettes sont écrites en *Italique* (avec une première majuscule), pour les concepts, et dans une *notationALaJAVA*, pour les relations. Nous associons également à l'entité un nom abrégé, qui est utilisé pour la formalisation logique.

[9] La somme méréologique de deux *Statives* (ex. : être assis) est un *Stative* du même type, cette propriété n'étant pas valable pour des *Events* (ex : un match de football).

[10] La notion d'*Action* figure en réalité dans DOLCE-Lite+, une extension « brouillon » de DOLCE, sans être formellement définie. Nous introduirons la relation *contrôle* et le rôle *Agent* dans la section suivante.

On notera également la principale relation entre *Endurants* et *Perdurants*, la relation ternaire de participation temporelle (PC) signifiant que : *un Endurant participe à un Perdurant durant un Time Interval*.

4 Formalisation de notre notion d'artefact

Dans cette section, nous proposons une formalisation de notre notion d'artefact. Suivant le mode originel de spécification de DOLCE, nous élaborons une théorie logique du 1er ordre comportant différents types de formules : des Axiomes (A), des Définitions (D), des Théorèmes (T) et des Faits (F). Les variables apparaissant libres dans les formules doivent être considérées comme universellement quantifiées.

4.1 Rôles de participation

Comme nous l'avons vu dans l'analyse informelle menée en introduction, notre notion d'*artefact* repose sur des notions comme celles d'*agent* et d'*instrument* qui s'apparentent à des rôles de participation, autrement dit à des manières, pour des *Endurants*, de participer temporellement à des *Perdurants*. Pour modéliser de tels rôles, nous introduisons des relations primitives spécialisant la relation *PC* de participation temporelle. Nous définissons ainsi les relations *isAgentOfAt* (A1), *isInstrumentOfAt* (A2) et *isResultOfAt* (A3) signifiant respectivement que : i) *une entité contrôle l'action à laquelle elle participe* ; ii) *une entité est utilisée (par l'entité contrôlant l'action) pour aider à réaliser cette action* ; et iii) *une entité actualise (pour l'entité contrôlant l'action) le but visé par cette action*. On peut noter que ces rôles ne sont définis que vis-à-vis d'*Actions*. L'entité participante est pour sa part nécessairement un *Endurant* (T1)(T2)(T3).

(A1) $\text{isAgentOfAt}(x,y,t) \rightarrow \text{PC}(x,y,t) \wedge \text{ACT}(y)$
(T1) $\text{isAgentOfAt}(x,y,t) \rightarrow \text{ED}(x)$
(A2) $\text{isInstrumentOfAt}(x,y,t) \rightarrow \text{PC}(x,y,t) \wedge \text{ACT}(y)$
(T2) $\text{isInstrumentOfAt}(x,y,t) \rightarrow \text{ED}(x)$
(A3) $\text{isResultOfAt}(x,y,t) \rightarrow \text{PC}(x,y,t) \wedge \text{ACT}(y)$
(T3) $\text{isResultOfAt}(x,y,t) \rightarrow \text{ED}(x)$

Ces relations permettent de définir des classes d'*Endurants* suivant leur manière de participer à une *Action* (D1)(D2)(D3) et de spécialiser à leur tour ces classes suivant le type d'*Action* à laquelle l'*Endurant* participe (D4)(D5)(D6).

(D1) $\text{Agent}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y,t(\text{isAgentOfAt}(x,y,t))$
(D2) $\text{Instrument}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y,t(\text{isInstrumentOfAt}(x,y,t))$
(D3) $\text{Result}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y,t(\text{isResultOfAt}(x,y,t))$
(D4) $\text{AgentOfWritting}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y,t(\text{Writting}(y) \wedge \text{isAgentOfAt}(x,y,t))$
(T4) $\text{AgentOfWritting}(x) \rightarrow \text{Agent}(x)$
(D5) $\text{InstrumentOfPaperKeeping}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y,t(\text{PaperKeeping}(y) \wedge \text{isInstrumentOfAt}(x,y,t))$
(T5) $\text{InstrumentOfPaperKeeping}(x) \rightarrow \text{Instrument}(x)$

(D6) ResultOfDiagnosing(x) $=_{def}$ ED(x) $\wedge$ $\exists$y,t(Diagnosing(y) $\wedge$ isResultOfAt(x,y,t))
(T6) ResultOfDiagnosing(x) $\rightarrow$ Result(x)

Une telle modélisation conduit à considérer une taxinomie de rôles de participation sous le concept *Endurant* (cf. Fig. 1). Cette figure illustre surtout, avec le rôle *Instrument*, le choix de modélisation retenu pour les instances de ces rôles. Ainsi, pour prendre l'exemple d'un objet physique comme un galet jouant temporellement le rôle d'aider à maintenir des papiers, nous ne considérons dans notre modèle qu'une seule entité sur laquelle portent deux points de vue différents (F1)(F2). Il convient de noter que cette conceptualisation est conforme au paradigme courant de modélisation des rôles (Steimann, 2000).
(F1) Pebble(Pebble#i)
(F2) InstrumentOfPaperKeeping(Pebble#i)

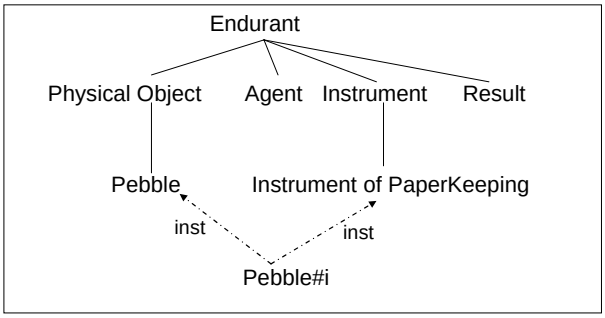


Fig. 1 – Pebble#i *a pour type* Pebble et *joue le rôle* de InstrumentOfPaperKeeping

4.2 Compétences et fonctions

La modélisation des rôles de participation *Agent* et *Instrument* nous permet d'aborder la modélisation des notions de *compétence* et de *fonction*.

Ces dernières notions reposent, comme on l'a vu, sur l'attribution de capacités à des entités. Pour rendre compte de cette attribution, nous introduisons la relation primitive *hasCapacity* (A4) signifiant qu*'une entité (nécessairement un Endurant) a la capacité d'exercer un rôle donné dans une classe d'Actions donnée*. La capacité attribuée correspondant à différentes façons d'exercer ce rôle dans des actions, certes d'un même type, mais diverses, nous l'assimilons à l'idée générale (un *Concept*) de jouer ce rôle dans ce type d'action. Un *Concept* classe temporellement des entités : la relation *classifiesAt* (A5) signifie qu'*un Concept classe une entité durant un intervalle de temps* ou, formulé en d'autres termes, qu*'une entité vérifie durant un intervalle de temps l'ensemble des propriétés constituant le Concept*. Une *Capacity* est dès lors définie comme un *Concept* de rôles de participation à des *Actions* – un *Role* - (A6) attribué à un *Endurant* (D7). À noter la notation utilisée pour nommer les constantes désignant des *Concepts* : le nom du prédicat est placé entre crochets (ex : [Physical Object]).

(A4) $\text{hasCapacity}(x,y) \rightarrow \text{ED}(x) \wedge \text{Role}(x)$
(A5) $\text{classifiesAt}(x,y,t) \rightarrow \text{ED}(y) \wedge \text{PT}(x) \wedge \text{T}(t)$
(F3) $\text{classifiesAt}([\text{Pebble}],\text{Pebble\#i},t\#j)$
(A6) $\text{Role}(x) \rightarrow \text{CPT}(x) \wedge \forall y,t(\text{classifiesAt}(x,y,t) \rightarrow \text{ED}(y) \wedge \exists z,t'(\text{ACT}(z) \wedge \text{PC}(y,z,t'))$
(D7) $\text{Capacity}(x) =_{\text{def}} \text{Role}(x) \wedge \exists y(\text{hasCapacity}(y,x))$

Suivant le rôle (*Agent* ou *Instrument*) exercé, la *Capacity* devient une *Competence* (D8) ou une *Function* (D10). La *Competence* et la *Function* les plus abstraites correspondent respectivement aux concepts d'*Agent* et d'*Instrument* (F4)(F6), autrement dit aux idées respectives de contrôler et d'aider à réaliser une *Action*. Une *Competence* plus spécifique est le concept d'*AgentOfDiagnosing* (F5) ; une *Function* plus spécifique est celle d'*InstrumentOfPaperKeeping* (F7).

(D8) $\text{Competence}(c) =_{\text{def}} \text{Capacity}(c) \wedge \forall y,t(\text{classifiesAt}(c,y,t) \rightarrow \text{Agent}(y))$
(D9) $\text{hasCompetence}(x,c) =_{\text{def}} \text{hasCapacity}(x,c) \wedge \text{Competence}(c)$
(F4) $\text{Competence}([\text{Agent}])$
(F5) $\text{Competence}([\text{AgentOfDiagnosing}])$
(D10) $\text{Function}(f) =_{\text{def}} \text{Capacity}(f) \wedge \forall y,t(\text{classifiesAt}(f,y,t) \rightarrow \text{Instrument}(y))$
(D11) $\text{hasFunction}(x,f) =_{\text{def}} \text{hasCapacity}(x,f) \wedge \text{Function}(f)$
(F6) $\text{Function}([\text{Instrument}])$
(F7) $\text{Function}([\text{InstrumentOfPaperKeeping}])$

Les primitives que nous venons d'introduire permettent finalement de définir deux catégories d'*Endurants* : des *Agentives*, auxquels une *Competence* est attribuée (D12), et des *FunctionalObjects*, auxquels une *Function* est attribuée (D13).

(D12) $\text{Agentive}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y(\text{hasCompetence}(x,y))$
(D13) $\text{FunctionalObject}(x) =_{\text{def}} \text{ED}(x) \wedge \exists y(\text{hasFunction}(x,y))$

4.3 Artefacts

Pour définir finalement le concept d'*artefact*, il nous reste une étape à franchir : définir la notion d'*artificialité*. Le terme « artificiel » s'oppose couramment au terme « naturel » pour désigner toute entité « découlant » ou « résultant » d'une action humaine. Dans cette phrase, nous avons volontairement accolé les termes « découler » et « résulter » pour souligner leur large synonymie. Sur le plan sémantique, pour notre modélisation, nous avons besoin d'une primitive conceptuelle traduisant une notion de *conséquence* mais n'entretenant aucun lien avec le but visé par l'action (contrairement au concept *Result*). Nous introduisons ainsi la relation *isConsequenceOf* (A7) pour pouvoir désigner *une entité dont, soit l'existence, soit des propriétés, découlent ou « sont la conséquence » d'une Action*, autrement dit *une entité produite ou transformée au cours d'une Action*. Plutôt qu'une relation temporelle de participation à une *Action*, à l'instar de la relation *isResultOfAt*, la relation *isConsequenceOf* introduit une propriété « historique » (le fait d'avoir été créé ou transformé par une *Action*) qui demeure satisfaite par l'entité tant que celle-ci existe, ou tant que les modifications apportées subsistent.

(A7) $\text{isConsequenceOf}(x,y) \rightarrow \text{ED}(x) \wedge \text{ACT}(y)$

La relation *isConsequenceOf* permet de définir le concept *ArtificialObject* (D14) et ce dernier permet finalement de définir notre concept *Artefact* (D15)(T7) : il s'agit d'un *ArtificialObject* résultat (au sens de la relation *isResultOfAt*) de l'*Action* l'ayant produit (ce résultat était donc visé) et auquel une *Function* est attribuée (c'est un *FunctionalObject*). Comme conséquence de cette modélisation, deux catégories de *FunctionalObjects* sont considérées (cf. Fig. 2) : des entités qui, à l'instar des presse-papiers (*Paper-Weight*) (D16)(T8), peuvent être indifféremment naturels ou artificiels, et des *Artefacts* qui, à l'instar des agrafeuses (*Stapler*) (D17), sont nécessairement des *ArtificialObjects*.

(D14) $\text{ArtificialObject}(x) =_{def} \text{ED}(x) \wedge \exists y(\text{isConsequenceOf}(x,y))$

(D15) $\text{Artefact}(x) =_{def} \text{ED}(x) \wedge \exists y,t(\text{isConsequenceOf}(x,y) \wedge \text{isResultOfAt}(x,y,t)) \wedge \text{FunctionalObject}(x)$

(T7) $\text{Artefact}(x) \rightarrow \text{ArtificialObject}(x)$

(D16) $\text{Paper-Weight}(x) = \text{ED}(x) \wedge \text{hasFunction}(x,[\text{InstrumentOfPaperKeeping}])$

(T8) $\text{Paper-Weight}(x) \rightarrow \text{FunctionalObject}(x)$

(D17) $\text{Stapler}(x) =_{def} \text{Artefact}(x) \wedge \text{hasFunction}(x,[\text{InstrumentOfPaperStapling}])$

Comme pour les rôles de participation, ce mode de définition des *Artefacts* revient à spécialiser le concept *Endurant* au moyen d'une nouvelle taxinomie et à considérer des entités ayant tout à la fois un type et remplissant une (ou plusieurs) fonctions(s).

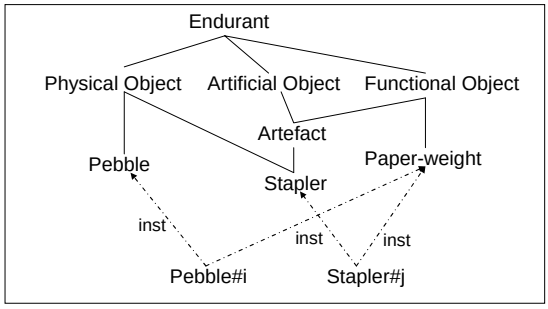


Fig. 2 – Pebble#i *a pour type* Pebble *et remplit la fonction de* Paper-Weight ; Stappler#j *a pour type* Stappler *et remplit la fonction de* Paper-Weight.

5 Travaux liés

Stefano Borgo et Laure Vieu (2006, 2008) ont récemment jeté les bases d'une ontologie formelle d'artefacts physiques spécialisant également l'ontologie DOLCE. Leur proposition consiste à attribuer aux artefacts un statut ontologique à part entière en leur conférant comme propriété essentielle le fait d'avoir été intentionnellement « créés ». L'acte de création en question correspond à la « sélection » d'un objet physique (ex : un galet) pour en faire un artefact (ex : un presse-papier). La sélection

correspond pour sa part à l'attribution de capacités (ex : maintenir des papiers sans les abimer, être aisément saisissable par la main). Ces capacités, sélectionnées parmi des capacités de l'objet physique *constituant* (au sens de DOLCE) l'artefact, sont assimilées à des *qualités* (toujours au sens de DOLCE) particulières prenant temporellement des valeurs dans un espace de capacités. Bien que l'ontologie de Stefano Borgo et Laure Vieu (B&V) prenne également DOLCE comme référence, nous pouvons noter plusieurs différences significatives entre les deux ontologies.

En premier lieu, B&V considèrent trois entités – l'artefact, l'objet physique constituant l'artefact et la quantité de matière constituant l'objet physique – là où nous ne considérons que les deux dernières. Nous adoptons ainsi une stratégie de modélisation plus réductionniste. Ceci revient à admettre qu'un concept d'artefact comme celui de chaise est un concept complexe comportant, d'une part, la description d'une catégorie d'objets physiques et, d'autre part, la propriété d'avoir été créé intentionnellement ainsi qu'une ou plusieurs propriétés fonctionnelles.

En second lieu, on notera que B&V utilisent le terme "artefact" en lieu et place de notre terme "entité fonctionnelle". Comme nous l'avons noté, B&V assimilent la création d'un artefact à l'attribution d'une fonction à un objet, ce qui pour nous correspond à l'attribution d'une propriété à un objet déjà existant. De fait, B&V ne considèrent pas dans leur analyse la création intentionnelle de l'objet physique. Ce faisant ils s'éloignent, nous semble-t-il, de la définition courante d'artefact.

Enfin, une différence importante tient à la nature accordée à la *capacité*. Pour B&V, la capacité est assimilée à une qualité inhérente à une entité, prenant des valeurs dans un espace de capacités. Dans notre approche, au contraire, une capacité (fonction ou compétence) étant assimilée à un concept attribué par un observateur à une entité, nous la considérons comme extrinsèque à l'entité. Notre modélisation de ces capacités passe par l'introduction dans le domaine de propriétés réifiées. Moyennant cette introduction, nous disposons d'un ensemble de concepts et de relations permettant de décrire des artefacts spécifiques en les reliant à des actions spécifiques. À noter toutefois, sur le plan de la représentation des connaissances, que nous sortons des capacités d'inférence des logiques de description.

6 Conclusion

Dans cet article, nous avons proposé un cadre conceptuel pour rendre compte de la nature générale des artefacts, qui reprend la notion philosophique courante d'artefact et la précise en tenant compte de la littérature récente sur ce sujet. Ce cadre pose les bases d'une ontologie d'artefacts que nous espérons suffisamment solides pour permettre d'élaborer une ontologie plus large. En vue d'étendre cette proposition, nous avons inscrit deux objectifs dans notre agenda recherche.

D'une part, notre ontologie étant ancrée dans une ontologie d'actions largement informelle, nous avons le projet de compléter et de formaliser cette dernière. Nous souhaitons notamment disposer de principes pour définir des types d'actions. Nous considérons en effet que ceci constitue un préalable pour définir des types d'entités fonctionnelles et, par là même, des types d'artefacts.

D'autre part, nous comptons évaluer notre ontologie en la confrontant à un domaine spécifique d'artefacts, celui des programmes informatiques. Il s'agit là de poursuivre un travail déjà engagé (Kassel *et al.*, 2007) consistant à évaluer la pertinence de notre cadre conceptuel à rendre compte de notions comme celles de *programme*, *plate-forme (logicielle et/ou matérielle)* ou *service Web*, ces notions désignant des entités à la fois structurelles et fonctionnelles à des degrés divers.

Références

BAKER L.R. (2004). The Ontology of Artifacts. *Philosophical Explorations*, 7(2), p. 99-111.

BLOOM P. (1996). Intention, history, and artifacts concepts, *Cognition*, vol. 60, p. 1-29.

BORGO S. & VIEU L. (2006). From Artefacts to Products. In Proceedings of the *second Workshop: Formal Ontologies meet Industry (FOMI 2006)*, Trento (Italie).

BORGO S. & VIEU L. (2008). Artifacts in formal ontology. In A. MEIJERS (ed.), *Handbook of the Philosophy of the Technical Sciences*, volume 2: Artifact ontology and artefact epistemology, Elsevier.

CHANDRASEKARAN B. & JOSEPHSON J.R. (2000). Function in device representation. *Engineering with Computers*, vol. 16, n° (3/4), p. 162-177.

HILPINEN R. (1992). On Artifacts and Works of Art. *Theoria*, vol. 58, p. 58-82.

HILPINEN R. (2004). Artifact. *Stanford Encyclopedia of Philosophy*, 2004. Disponible à http://plato.stanford.edu/entries/ artifact/.

KASSEL G., LANDO P., LAPUJADE A. & FURST F. (2007). Des Artefacts aux Programmes. In F. GARGOURI *et al.* (eds), Actes de la *1ère édition des Journées Francophones sur les Ontologies (JFO 2007)*, Sousse (Tunisie), p. 281-300.

KROES P. & MEIJERS A. (2006). The dual nature of technical artefacts. *Studies in History and Philosophy of Science*, vol. 37, p. 1-4.

LANDO P., LAPUJADE A., KASSEL G. & FURST F. (2007). Towards a general ontology of computer programs. In Proceedings of the *2nd International Conference on Software and Data Technologies (ICSOFT 2007)*, Conference area: Knowledge Engineering, Barcelonia (Spain).

MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. & SSCHNEIDER L. (2003). The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. *WonderWeb Deliverable D18, Final Report*, vr. 1.0.

PACHERIE E. (2000). The content of intentions. *Mind and Language* 15, p. 400-432.

SOWA J.F. (2000). *Knowledge Representation: logical, philosophical, and computational foundations*, Brooks/Cole.

STEIMANN F. (2000). On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering*, 35, p. 83-106.

TEMAL L., DOJAT M., KASSEL G. & GIBAUD B. (2008). Towards an ontology for sharing medical images and regions of interest. *Journal of Biomedical Informatics*, 41, p. 766-778.

TEMAL L., LANDO P., GIBAUD B., DOJAT M., KASSEL G. & LAPUJADE A. (2006). OntoNeuroBase: a multi-layered application ontology in neuroimaging. In Proceedings of the *2nd Workshop: Formal Ontologies Meet Industry (FOMI 2006)*, Trento (Italy).

THOMASSON A.L. (2003). Realism and Human Kinds. *Philosophy and Phenomenological Research*, vol. LXVII, n°3, p. 580-609.

Alignement entre des ontologies de domaine et la Snomed: trois études de cas

Laurent Mazuel et Jean Charlet

INSERM UMR_S 872, Eq. 20
15, rue de l'École de Médecine, 75006 Paris
{Laurent.Mazuel,Jean.Charlet}@spim.jussieu.fr

Résumé : Les expériences sur les ontologies montrent de plus en plus clairement qu'elles ne représentent correctement et de façon consensuelle que des domaines réduits. Ainsi, les ontologies de domaines sont développées pour des applications particulières alors que des ontologies de référence tendent à être utilisées pour fédérer les résultats des applications spécifiques. Nous présentons dans cet article la construction, l'analyse et la discussion d'un alignement entre trois ontologies de domaine construites à l'INSERM UMR_S 872, Éq. 20 (*i.e.* OntoPneumo, OntoHTA et OntoReaChir) et la classification SNOMED v3.5.

Mots-clés : Alignement d'ontologies, ontologie médicale, SNOMED, terminologie de référence, terminologie d'interface.

1 Introduction

Depuis de nombreuses années, la médecine a produit de nombreuses terminologies pour des applications diverses, de tous types, classification, thésaurus et plus récemment des ontologies.

Les expériences sur les ontologies montrent de plus en plus clairement qu'elles ne représentent correctement et de façon consensuelle que des domaines réduits. En suivant ces travaux (Rosenbloom *et al.*, 2006), on parle de *terminologies de référence* et de *terminologies d'interface* : 1) les terminologies de référence ont des visées de représentations larges et de référentiel pour des futures études épidémiologiques. La plus connue est la SNOMED ; 2) les terminologies d'interface sont développées pour des applications spécifiques. On retrouve cette dichotomie au niveau des ontologies où des ontologies d'interfaces contenant des ontologies de domaines sont développées pour des applications particulières alors que des ontologies de référence – *i.e.*, toujours la SNOMED dans sa version ontologique, la SNOMED-CT– tendent à être utilisées pour fédérer les résultats, souvent de l'indexation au sens large, des applications spécifiques.

À l'INSERM UMR_S 872, Éq. 20, nous avons développé un certain nombre d'ontologies pour des applications, souvent d'aide au codage médical, mais aussi pour des motivations de modélisation liées à des études d'usage (Charlet *et al.*, 2008). Il existe par ailleurs une terminologie de référence, la SNOMED v3.5, dont la France a acquis les

droits d'usage pour tout le territoire et qui a vocation à être la terminologie de référence de la santé en France.

Dans ce contexte, les applications médicales diverses qui, pour la plupart encore une fois, indexent des données médicales liées à des patients, ne peuvent prétendre à l'avenir contribuer à l'indexation de données de santé pour des études épidémiologiques que si elles sont alignées avec la SNOMED v3.5[1]. Cet alignement est donc un passage obligé du développement de ces ontologies. C'est lui que nous allons étudier ici en analysant comment trois ontologies développées à l'INSERM UMR_S 872, Éq. 20 s'alignent avec la SNOMED. Ces trois ontologies sont une ontologie de la réanimation chirurgicale, OntoReaChir, une ontologie de la pneumologie, OntoPneumo, et une ontologie de l'hypertension artérielle, OntoHTA.

Dans la section 2, nous décrivons le contexte et l'objectif de ce travail ; dans la section 3, nous décrivons précisément les ontologies et terminologies impliquées dans ce travail ; dans la section 4 nous décrivons la méthode d'alignement mise en œuvre ; dans la section 5 nous présentons et discutons les résultats ; enfin, nous concluons en 6.

2 Contexte et objectif

Notre but va être ainsi de vérifier les possibilités d'alignement, ensuite à quels niveaux ils se font. Ainsi, une ontologie est maintenant classiquement découpée en 3 niveaux :

1. La *top*-ontologie, que l'on devrait plus précisément appeler « ontologie formelle » pour reprendre l'appellation des philosophes. C'est le niveau le plus abstrait structurant les connaissances de haut niveau avec des catégories dont l'organisation dépend de réflexions philosophiques.
2. La *core*-ontologie, fournissant les concepts structurant du domaine et décrivant les relations entre ces concepts – en médecine, on y trouve des concepts de *diagnostic*, *signe*, *structure anatomique* et des relations comme celles liées à la localisation d'une pathologie sur une structure anatomique.
3. L'ontologie du domaine, c'est-à-dire les concepts du domaine tels qu'ils sont manipulés par les professionnels – ici de santé. Dans notre équipe, le troisième et dernier niveau est celui que l'on construit avec les outils de TAL puisque l'on analyse les documents produits en activité avec ceux-ci.

Ce découpage nous servira de caractérisation des ontologies étudiées dont on peut déjà dire qu'elles sont, comme beaucoup d'autres, faites pour des usages spécifiques.

Par rapport à ce découpage et la structure connue des différentes ontologies, nous énonçons les hypothèses suivantes avant expérimentation :

– les feuilles des hiérarchies des ontologies de spécialité seront plus précises que les concepts de la SNOMED ;
– la SNOMED ne contenant pas de *top*-ontologie, la *top*-ontologie des spécialités, si elle existe, ne devrait pas s'apparier ;
– finalement, les appariements devraient se situer au niveau de la *core*-ontologie.

[1]Dans la suite de l'article, sauf précision particulière, nous utiliserons le vocable SNOMED pour la SNOMED v3.5.

Notre objectif est donc de confirmer ou d'infirmer ces hypothèses et, le cas échéant, de découvrir d'autres caractéristiques de ces ontologies.

3 Matériel

Nous décrivons dans cette section les modèles de connaissances que nous considérons dans nos alignements, à savoir la SNOMED d'un coté et les trois ontologies de spécialités de l'autre.

3.1 La classification SNOMED v3.5

La SNOMED v3.5 est une classification multi-axiale standardisée en français, contenant actuellement 116 000 concepts (pour environ 150 000 termes en comptant les synonymes). Elle est organisée en 11 axes (Terminologie, Diagnostic, etc.), chacun de ces axes étant organisé hiérarchiquement. Il existe de plus des liens non hiérarchiques entre ces axes. Par exemple, le concept « hémorragie ombilicale » de l'axe *Fonction* est associé à « ombilic » (axe *Terminologie*) et « hémorragie » (axe *Morphologie*).

La SNOMED-CT, évolution de la SNOMED v3.5[2], est actuellement la plus grande ontologie médicale, avec près de 344 000 concepts. Néanmoins, elle n'est actuellement définie qu'en anglais et n'est ainsi pas applicable dans le cadre de cet article, étant donné que nos ontologies de spécialité sont définies en français.

3.2 L'ontologie OntoPneumo

Le projet OntoPneumo visait à construire une ontologie de la pneumologie pour l'aide au codage médical (Baneyx, 2007). La construction de cette ontologie a fortement utilisé les ressources terminologiques du domaine à modéliser pour rendre compte, le plus précisément possible, non seulement des pratiques médicales actuelles en pneumologie mais également des vocabulaires utilisés par les médecins.

Cette construction est basée sur l'utilisation du logiciel SYNTEX-UPERY (Bourigault & Fabre, 2000) comme outils d'analyse de corpus de texte et de traitement (automatique) du langage pour obtenir un réseau de candidats termes, leurs proximités contextuelles et leurs liens avec le corpus source. L'éditeur DOE[3] a permis de construire l'ontologie selon la sémantique différentielle, les étapes de formalisation et d'opérationalisation étant réalisées à l'aide de l'éditeur d'ontologies PROTÉGÉ[4]. Par ailleurs, l'ontologie a été complétée par une analyse du thésaurus de spécialité de la pneumologie[5]. Ce thésaurus est conçu comme une sous-partie de la CIM-10 réduite aux pathologies de la pneumologie. Ainsi, par construction, les 337 termes de ce thésaurus sont bien inclus dans OntoPneumo. Pour finir, l'ontologie a été validée par un expert du domaine, médecin en milieu hospitalier dans un service de pneumologie.

[2]Les liens vers les concepts originaux SNOMED v3.5 sont d'ailleurs accessibles.

[3]The Differential Ontology Editor, `http://homepages.cwi.nl/~troncy/DOE/`

[4]`http://protege.stanford.edu/`

[5]Disponible sur le site de la Société de Pneumologie de Langue Française : `http://www.splf.org`

Cette ontologie du domaine compte actuellement 1 114 concepts, mais sans l'utilisation d'une *top*-ontologie (*i.e.* actuellement OntoPneumo est constitué de 25 arbres disjoints hiérarchiquement). En effet, l'intégration de la *top*-ontologie et de la *core*-ontologie du projet MENELAS[6] devrait mieux organiser la hiérarchie et ajouter environ 400 concepts. L'ontologie définit également une hiérarchie de 27 relations.

3.3 L'ontologie de l'hypertension artérielle, OntoHTA

OntoHTA est issue d'un projet de recherche sur les déterminants du raisonnement médical qui a abouti à la construction d'une première ontologie et a déjà eu pour effet de proposer une mise à jour des formulaires d'entrées de données cliniques dans le domaine de l'hypertension artérielle. Cette ontologie est en cours de construction par un médecin spécialiste (Steichen *et al.*, 2007) en tenant en partie compte de la SNOMED-CT, en particulier pour les termes associés aux concepts en anglais.

Comme dans le projet OntoPneumo, les outils de traitement automatique du langage SYNTEX et UPERY ont été choisis pour l'analyse des corpus (commentaires en texte libre et guides de bonne pratique). La modélisation ontologique a été réalisée, concept par concept, dans l'éditeur DOE.

Actuellement cette ontologie est une monohiérarchie strictement taxinomique, dans le respect des principes de la sémantique différentielle, et qui organise 506 concepts. Cette ontologie bénéficie d'une *top*-ontologie articulant l'ensemble des concepts.

3.4 L'ontologie de la réanimation chirurgicale, OntoReaChir

La réanimation chirurgicale est un domaine médical spécialisé dans la prise en charge des complications postopératoires et dans la traumatologie. Comme dans les deux ontologies précédentes, la base de l'ontologie a été construite à partir de corpus (800 comptes rendus hospitaliers) sur le logiciel SYNTEX-UPERY (Le Moigno *et al.*, 2002). Par ailleurs, l'élément de référence utilisé pour l'évaluation de l'ontologie est la version du thésaurus de spécialité — correspondant à peu près au thésaurus de la CIM-10 – émise en 1999.[7]

Ceci a abouti à une ontologie constituée d'une hiérarchie taxinomique de 2 039 concepts, ainsi que d'une hiérarchie de 200 relations. Cette ontologie possède une *top*-ontologie très détaillée ne correspondant pas à une *top*-ontologie spécifiée préalablement mais proche de celle de MENELAS (Charlet *et al.*, 1996). La partie basse est ainsi celle qui correspond le plus au thésaurus initial du domaine. OntoReaChir a été récemment reprise pour être décrite formellement en OWL.

4 Méthode

Cette section décrit la méthode que nous avons utilisée pour produire les alignements entre la SNOMED et nos trois ontologies. Dans une première section, nous justifierons

[6] `http://estime.spim.jussieu.fr/Menelas/`

[7] Version disponible sur le site de la société française d'anesthésie et de réanimation : `www.sfar.org`.

nos choix pour la production de ces alignements. Dans un deuxième temps, nous présenterons en détail notre méthode semi-automatique.

4.1 Méthodes d'alignements applicables pour nos ontologies

Il existe différentes méthodes d'alignement, plus ou moins utilisables en fonction des situations et des formalismes considérés pour les ontologies (Euzenat & Shvaiko, 2007). Dans notre situation, l'absence d'instances aussi bien dans la SNOMED que dans nos trois ontologies supprime les possibilités liées à ce type de méthode (Ichise *et al.*, 2003). De même, les approches demandant une troisième ontologie (avec un rôle de *médiateur* (Aleksovski *et al.*, 2006)) ne sont pas applicables, car la SNOMED est elle-même une ontologie médiatrice (de part son aspect générique et sa grande couverture générale du domaine). Les méthodes d'alignement structurel (Breitman *et al.*, 2005) reposent sur l'idée que si deux noeuds sont alignés, alors leurs ancêtres et leurs enfants doivent s'aligner mutuellement entre eux (cette hypothèse permettant à la fois de tester la cohérence des propositions d'appariements ainsi que d'en proposer de nouveaux). La structure particulière de nos ontologies (*e.g.* l'absence de *top*-ontologie pour OntoPneumo), ainsi que la structure de la SNOMED (qui bien qu'organisée hiérarchiquement n'est pas parfaitement analogue à une hiérarchie de subsomption comme dans les ontologies) font qu'il nous était difficile d'appliquer cette méthode de manière automatique. Néanmoins, nous avons tenu compte manuellement, lorsque cela était possible, de la cohérence structurelle des alignements proposés.

Finalement, nous ne pouvons utiliser que les méthodes morpho-syntaxiques pour produire automatiquement notre alignement. L'alignement morpho-syntaxique consiste à chercher un alignement entre concepts en ne se basant que sur les labels des termes. Cette étape utilise différents outils de TAL pour proposer des alignements basés sur la correspondance entre les chaînes de caractères représentant les concepts. Toute méthode d'alignement commence toujours par une première étape de ce type, afin de fournir un alignement initial de travail. Ainsi, notre calcul initial réside aussi sur ce type d'algorithmes. D'autre part, la complétude et la vérification de l'alignement sont effectuées manuellement. La section suivante décrit en détail ces processus.

4.2 Méthode d'alignement utilisé

Cette section décrit la méthode que nous avons utilisée pour calculer notre alignement. Dans un premier temps, nous présenterons les méthodes automatiques mises en place pour produire une proposition d'alignement initial, puis les méthodes de vérification et complétion manuelles.

4.2.1 Description de la fonction d'appariement morpho-syntaxique

Notre approche morpho-syntaxique se découpe en deux parties. Tout d'abord, nous normalisons et simplifions les chaînes de caractères représentant les concepts pour les ramener à des formats équivalents. Ensuite, nous utilisons la distance de Levenshtein et

la distance de Stoilos *et al.* (2005)[8] pour augmenter la portée des propositions d'alignements.

Dans la partie sur la normalisation syntaxique, nous effectuons plusieurs opérations successives :

- retirer les diacritiques (*i.e.* les accents, cédilles, etc.). Ceci peut poser quelques problèmes d'ambiguïtés que nous résolverons dans la partie manuelle (*e.g.* les termes « côte - coté » ou « aine - aîné » ne sont plus distinguables sans diacritiques) ;
- réduire l'ensemble de la casse aux lettres minuscules ;
- normaliser le nombre d'espaces et les remplacer par le caractère « _ » ;
- supprimer les mots de liaisons pour ne garder que les noms (*e.g.* suppression de « l' », « le », « la », etc.) ;
- pour la SNOMED, retirer certains suffixes inutiles pour l'alignement comme par exemple la chaîne « SAI » (*i.e.* Sans Autre Indication).

Nous utilisons ensuite la distance de Levenshtein normalisée[9] avec un seuil à $0,97$, ce qui est (empiriquement) suffisant pour rattraper une partie des fautes d'orthographe sans impliquer un trop grand nombre d'erreurs d'alignement. Nous complétons ce premier alignement par un calcul suivant la distance de Stoilos, avec un seuil à $0,9$. Les deux étapes sont complémentaires : la distance de Levenshtein est robuste face aux fautes d'orthographes alors que la distance de Stoilos est plus efficace pour l'analyse des sous-chaînes de caractères.

4.2.2 Validation et complétion manuelles des alignements

Malgré les méthodes précédentes, un alignement à la main est nécessaire pour compléter et valider l'alignement obtenu, par exemple :

- pour les abréviations : « ALAT » $\equiv$ « aspartate aminotransférase », « AVC » $\equiv$ « accident vasculaire cérébrale » ;
- pour les différences de notation : « Dosage de la ... » $\equiv$ « Mesure de la ... », « Syndrome de ... » $\equiv$ « Maladie de ... » ;
- pour les erreurs induites par les mesures : Par exemple, l'alignement entre « Dosage du facteur X » de OntoPneumo et « Dosage du facteur V » de la SNOMED.

Pour simplifier notre travail, nous avons développé un outil permettant facilement d'aligner une ontologie avec la SNOMED (figure 1). Ce logiciel affiche sur la partie gauche de l'écran la hiérarchie de l'ontologie considérée avec un code de couleur en fonction de la saisie de l'utilisateur dans la partie droite :

- le vert lorsqu'un équivalent existe. Il est alors noté dans la colonne de droite ;
- le rouge lorsqu'il n'existe pas d'équivalent du concept dans la SNOMED ;
- le jaune lorsque le concept n'a pas actuellement de statut.

Cette interface simple permet d'éditer les alignements, de récupérer les statistiques de chaque nœud (*i.e.* obtenir pour le sous-arbre de chaque nœud le nombre d'appariements effectués, le nombre de concepts notés sans appariements possibles et calculer les proportions respectives) et offre un outil de visualisation permettant de mieux appréhender la répartition des alignements en fonction des branches de l'ontologie considérée.

[8] Nous remercions le relecteur qui nous a suggéré cette distance pour améliorer cette étape.

[9] Dans l'intervalle $[0, 1]$, 1 pour des chaînes identiques et 0 pour des chaînes totalement différentes.

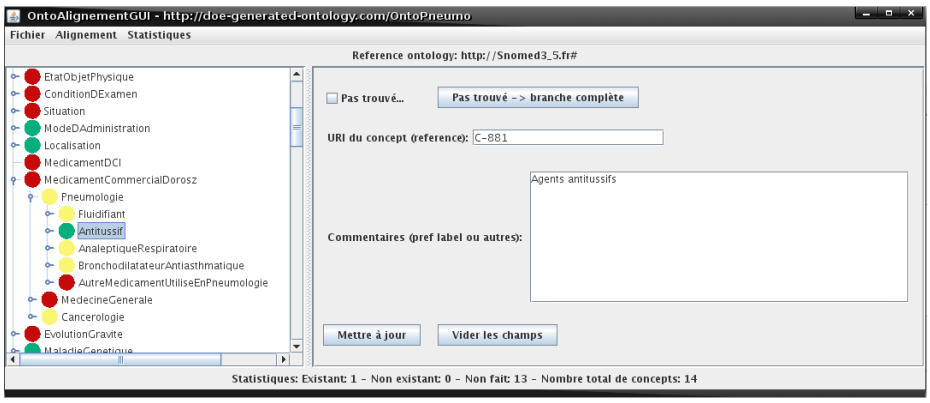


FIG. 1 – Logiciel d'alignement entre la SNOMED et une ontologie, ici OntoPneumo.

Ontologie	Nb. concepts	Nb. appariements automatiques		Nb. final appariements
		Directs	Validés	
OntoPneumo	1114	669	613	787
OntoHTA	506	159	144	228
OntoReaChir	2039	1046	987	1187

TAB. 1 – Résultats des alignements.

Cette partie est importante, l'alignement manuel effectué en complément permet d'augmenter sensiblement le nombre d'appariements (pratiquement le double pour HTA).

5 Résultats et discussion

Cette section présente les alignements obtenus entre nos trois ontologies et la SNOMED (tableau 1 et figure 2). Nous commencerons par discuter des alignements ontologie par ontologie, puis nous conclurons avec une discussion générale de ces résultats.

5.1 Résultats par ontologies de spécialité

5.1.1 OntoPneumo

OntoPneumo est l'ontologie de spécialité parmi les trois que nous étudions qui s'aligne le mieux avec la SNOMED (avec 75% de recouvrement). Ceci peut s'expliquer pour deux raisons :

– OntoPneumo ne contient pas de *top*-ontologie, ses concepts « hauts » s'apparient

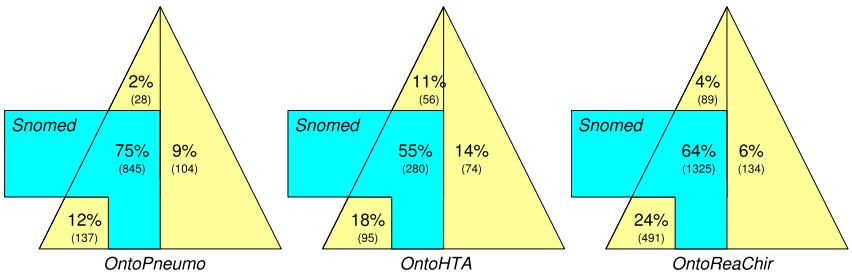


FIG. 2 – Schéma de la répartition moyenne des alignements pour les trois ontologies de spécialités. La partie droite représente la proportion de sous-arbres complets (*i.e.* de la racine jusqu'aux feuilles) sans aucun alignement. En plus des concepts qui sont directement alignés, la partie « SNOMED » inclut aussi tous les concepts qui ont à la fois un ancêtre *ET* un descendant alignés, et sont donc conceptuellement inclus dans la SNOMED (ce qui explique pourquoi ce nombre est plus grand que le nombre d'alignements noté au tableau 1).

donc bien avec la SNOMED, car ils ne sont pas très génériques (seul 2% des concepts hauts de OntoPneumo ne s'alignent pas avec la SNOMED, figure 2).

– OntoPneumo inclut le thésaurus de spécialité de la pneumologie (*cf.* section 3.2). Or, ce thésaurus est construit en utilisant la CIM-10 qui est elle-même incluse dans dans la SNOMED.

Ainsi, les axes de diagnostics sont pratiquement complètement représentés dans la SNOMED. Par contre, dans le cas de diagnostics composés, la SNOMED est assez faible. On entend par « diagnostic composé » les diagnostics liant une pathologie à une partie du corps, tel que « mésenthéliome pleural », « ablation du pouls carotidien » ou encore « granulome hyalin pulmonaire ». En effet, la SNOMED définit de manière générale une pathologie mais ne propose pas d'instances en fonction de l'anatomie sous-jacente. Ainsi, les pathologies pulmonaires de OntoPneumo existent dans la SNOMED, mais sous leur forme générale. Ceci implique que pour ce type de sous-arbres, les feuilles de l'ontologie de spécialité ne s'apparient pas dans la SNOMED (12% pour OntoPneumo), mais que les concepts plus hauts (et plus génériques), oui. A l'inverse, dans le cas de pathologies annexes (cas des maladies cardio-vasculaire ou psychiatrique par exemple), la SNOMED est alors totalement suffisante.

La SNOMED n'ayant pas vocation à coder les concepts non médicaux, les sous-arbres traitant de « rôle hospitalier » comme « médecin » ou « infirmière », ou traitant d' « instrument » comme « bistouri » ou « sonde », ainsi que les sous-arbres couvrant les médicaments ne sont absolument pas alignables avec la SNOMED[10]. Ceci est représenté par les 9% de concepts inclus dans des sous-arbres totalement non alignables.

[10]Le cas des médicaments est néanmoins particulier, car la SNOMED définit les principes actifs (« acide acétylsalicylique » ou « paracétamol »), mais pas les dénominations commerciales (« Aspegic », « Doliprane »). Un alignement est donc envisageable, mais sans équivalence stricte.

5.1.2 OntoHTA

L'ontologie de l'hypertension artérielle est celle qui s'aligne le moins bien avec la SNOMED. La raison principale réside dans la différence de la sémantique de spécialisation de la hiérarchie de subsomption. Autrement dit, les liens hiérarchiques de OntoHTA sont souvent assez peu compatibles avec les liens hiérarchiques de la SNOMED. Par exemple, les procédures de OntoHTA sont triés en « procédure par appareil », « procédure par pathologie », etc. alors que les procédures de la SNOMED sont classées par spécialités médicales, « procédure dentaire », « procédure psychiatrique », etc. ce qui implique que beaucoup de concept sont proches mais pas équivalents.

Comme dans OntoPneumo, les parties liées à la spécialité étudiée sont plus détaillées que dans la SNOMED. Par exemple, le concept « abolition du pouls » (présent dans les deux modèles de connaissances), n'est pas spécialisé dans la SNOMED, alors que OntoHTA la spécialise 18 fois en fonction des pouls possibles (*e.g.* « abolition du pouls carotidien », « abolition du pouls pédieux », etc.). Un autre point intéressant concerne la révision possible de OntoHTA grâce à l'alignement obtenu ou le repérage d'imprécisions dans la SNOMED. Par exemple, les concepts « polykystose rénale » et « maladie kystique congénitale du rein » sont considérés comme synonymes dans la SNOMED, alors qu'ils sont pères et « fils unique »[11] dans OntoHTA. Cette particularité peut dénoter une incomplétude d'OntoHTA ou une imprécision de la SNOMED, assumée ou pas par les constructeurs.

5.1.3 OntoReaChir

OntoReaChir possède une très grande *top*-ontologie, très détaillée. Il faut ainsi en moyenne parcourir 7 à 8 nœuds de profondeur pour arriver sur un nœud proche de la conceptualisation de la SNOMED.[12] La *top*-ontologie s'aligne ainsi assez peu. De plus, 9% des arbres sont totalement non alignables avec la SNOMED (de la racine aux feuilles), décrivant des concepts abstraits, utiles pour la définition de concepts définis, non applicables à la SNOMED.

Malgré tout, cette ontologie a été construite dans le but de couvrir les concepts du thésaurus de la spécialité du domaine. Or, à l'instar de OntoPneumo, ce thésaurus utilise les concepts de la CIM-10, couverts par la SNOMED. Ceci implique qu'un grand nombre de concepts de OntoReaChir s'apparie avec la SNOMED (64% de recouvrement, soit 1325 concepts). Ainsi, au final, la répartition des alignements est sensiblement équivalente à la répartition de OntoPneumo, malgré le nombre extrêmement différent de concepts.

On constate aussi des classements hiérarchiques très différents d'avec la SNOMED, impliquant une non-cohérence de la structure de l'alignement. Par exemple, la SNOMED classe les os par nom, puis les sous-classes représentant les parties de ces os (*e.g.* le concept « humérus » a comme fils « épiphyse de l'humérus »). Dans OntoReaChir, c'est l'inverse, les os sont classés par type de partie, puis instanciés au vrai os (*e.g.* le concept « épiphyse » a comme fils « épiphyse de l'humérus »).

[11] Nous entendons par « fils unique » un concept sans frères et étant ainsi l'unique fils du nœud initial.

[12] Exemple de descente dans la hiérarchie à partir de la racine : « Concept », « ObjetAbstrait », « ObjetDescription », « ObjetFonction », « ObjetFonctionPhysique », « ObjetFonctionPhysiquePathologie », ...

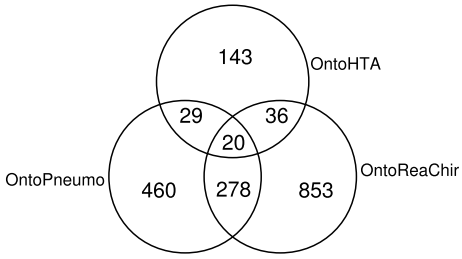


FIG. 3 – Recoupement et répartition des alignements obtenus entre les trois ontologies spécifiques étudiées.

5.1.4 Recoupement entre les alignements des ontologies de spécialité

Comme nous l'évoquions dans la section 4.1, ces alignements à la SNOMED peuvent permettre d'aligner aussi les ontologies spécifiques entre elles. Nous avons donc étudié la répartition des alignements communs à plusieurs ontologies spécifiques (figure 3).

Il y a 20 alignements communs aux trois ontologies. Il est intéressant de constater que ces concepts communs à toutes les ontologies tournent autour de la description de l'état de santé du patient (*e.g.* « tension artérielle »), de la description d'examen générique (*e.g.* « echocardiographie », « échographie Doppler ») et de la description des symptômes (*e.g.* « AVC »).

Il existe 56 alignements communs entre OntoHTA et OntoReaChir (36+20). Les alignements communs supplémentaires tournent autour des diagnostics sur le rein (*e.g.* « néphropathie ») et les maladies cardiaques (*e.g.* « souffle systolique »). Pour les 49 alignements communs entre OntoHTA et OntoPneumo, les alignements communs sont majoritairement issus des examens sanguins standards (*e.g.* « dosage du cholestérol LDL », « dosage du potassium ») ainsi que les maladies emboliques (*e.g.* « thromboembolisme aortique »).

Enfin, il existe un nombre étonnement grand d'alignements communs entre OntoPneumo et OntoReaChir (au total 298 alignements). La majeure partie (131) de ces alignements communs concerne la description morphologique du corps humain (*e.g.* « jambe » ou « prostate »). Le reste est divisé de manière a peu près équitable entre la définition d'organismes néfastes (*e.g.* « virus de l'hépatite C »), d'examen sanguin (*e.g.* « ASAT ») et de pathologique pneumologique (*e.g.* « détresse respiratoire aiguë »).

L'étude de ces alignements communs montre que même dans une ontologie de spécialité donnée il reste des informations issues de la médecine généraliste. Cette conclusion tend à montrer l'importance de l'appariement entre des classifications généralistes comme la SNOMED et les ontologies spécifiques.

5.2 Discussion

La SNOMED est une classification générique. En ce sens, elle manque de précisions pour énoncer les spécialité d'un domaine (*e.g.* les pathologies pulmonaires dans OntoPneumo). D'autre part, elle manque parfois de granularité, et deux concepts père/fils de la

SNOMED peuvent facilement être séparés par plusieurs descendants dans les ontologies de spécialité. Par exemple, dans l'ontologie de la réanimation chirurgicale, le concept « thorax » possède le fils « hémi-thorax » qui lui-même possède deux fils « hémi-thorax droit » et « hémi-thorax gauche ». Dans la SNOMED, « hémi-thorax droit » et « hémi-thorax gauche » sont directement des fils de « thorax ». Autre exemple tiré de OntoPneumo, le concept « lobectomie » est défini et possède plusieurs fils, alors que dans la SNOMED ce concept n'existe pas, mais il existe directement les opérations précises comme « lobectomie thyroïdienne unilatérale » (22 définitions de « lobectomie » au total !). Ceci tend à confirmer l'importance de la définition d'ontologies de domaines, et non d'une utilisation directe d'un modèle dit générique tel que la SNOMED.

Un autre exemple pour étayer cette conclusion réside non plus dans la structure, mais dans le vocabulaire. Celui des ontologies de spécialité, hérité de comptes rendus hospitaliers, est plus proche du véritable vocabulaire médical que celui de la SNOMED. Par exemple, la SNOMED ne définit aucune abréviations, là ou les abréviations tels que « AVC », « ASAT », « ALAT », sont couramment utilisées en clinique et apparaissent dans les ontologies de spécialité.

L'une des autres remarques concerne l'utilisation des thésaurus de spécialité spécifiés à partir de la CIM-10. Cette situation, alliée au fait que la SNOMED contient totalement la CIM-10, simplifie et optimise grandement les alignements. Ainsi, dans un objectif d'utilisation pratique de la SNOMED, son rapport à la CIM-10 permet d'envisager des applications d'aide au codage médical beaucoup plus précises que précédemment.

6 Conclusion

Nous avons présenté dans cet article les résultats d'une étude de cas sur l'alignement de trois ontologies de domaine (OntoPneumo, OntoHTA et OntoReaChir) avec la classification SNOMED v3.5. La répartition de ces alignements tend à montrer l'utilité de ces ontologies de domaine par rapport à l'utilisation directe d'un modèle dit générique. En effet, la granularité interne est plus adaptée dans les ontologies de spécialité, ainsi que le niveau de détail des concepts les plus spécifiques. D'autre part, le formalisme d'ontologie est plus complet et pensé pour la définition de concepts définis issus d'une étape de post-coordination, ce qui implique que des blocs entiers ontologies de spécialités ne soient absolument pas représentés dans la SNOMED.

Une évaluation naturelle de ce travail serait d'étudier les alignements avec la SNOMED-CT. La SNOMED-CT possédant les liens vers la SNOMED v3.5, nos alignements sont a priori directement récupérables comme base d'étude. D'autre part, l'ontologie OntoHTA a été construite sur la base de la SNOMED-CT (à l'inverse des deux autres ontologies qui ont plutôt utilisé la CIM-10). Ainsi, il est fort probable que cette étape fournisse des résultats d'étude très intéressants.

Deuxièmement, les ontologies de domaine de cet article ne possèdent pas de concepts définis issus de la post-coordination ou s'il y en a – c'est le cas pour OntoPneumo–, ils n'ont pas été utilisés. Or, il pourrait être intéressant d'étudier comment ces concepts particuliers peuvent s'aligner avec la SNOMED, et quelles en seraient les conséquences.

Enfin, la difficulté des alignements additionnés de la nécessité de développements d'ontologies de domaines spécialisés pousse à envisager l'alignement avec la SNOMED

durant la construction de la ressource elle-même.

Remerciements

Nous remercions particulièrement Audrey Baneyx et Olivier Steichen de leur participation à la lecture et à l'interpretation des résultats correspondant à leurs ontologies.

Références

ALEKSOVSKI Z., TEN KATE W. & VAN HARMELEN F. (2006). Exploiting the structure of background knowledge used in ontology matching. In *Proc. Workshop on Ontology Matching in ISWC2006* : CEUR Workshop Proceedings.

BANEYX A. (2007). *Construire une ontologie de la pneumologie : aspects théoriques, modèles et expérimentations.* PhD thesis, Université Pierre et Marie Curie (Paris VI). Disponible à `http://tel.archives-ouvertes.fr/tel-00136937/fr/`.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, (25), 131–51. numéro spécial « sémantique et corpus ».

BREITMAN K., FELICÍSSIMO C. & CASANOVA M. (2005). CATO–A Lightweight Ontology Alignment Tool. *Proc. 17th Conf. on Advanced Information Systems Engineering (CAISE'05)*.

CHARLET J., BACHIMONT B., BOUAUD J. & ZWEIGENBAUM P. (1996). Ontologie et réutilisabilité : expérience et discussion. In N. AUSSENAC-GILLES, P. LAUBLET & C. REYNAUD, Eds., *Acquisition et ingénierie des connaissances : tendances actuelles*, chapter 4, p. 69–87. Cepaduès-éditions.

CHARLET J., BANEYX A., STEICHEN O., ALECU I., DANIEL C., BOUSQUET C. & JAULENT M.-C. (2008). Utiliser et construire des ontologies en médecine : Le primat de la terminologie. *Techniques et Sciences Informatiques. À paraître*.

EUZENAT J. & SHVAIKO P. (2007). *Ontology matching*. Heidelberg (DE) : Springer-Verlag.

ICHISE R., TAKEDA H. & HONIDEN S. (2003). Integrating multiple internet directories by instance-based learning. *Proceedings of the eighteenth International Joint Conference on Artificial Intelligence (IJCAI03)*.

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.

ROSENBLOOM S. T., MILLER R. A. & JOHNSON K. B. (2006). Interface terminologies : facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc*, **13**(3), 277–88.

STEICHEN O., DANIEL-LE BOZEC C., JAULENT M.-C. & CHARLET J. (2007). Construction d'une ontologie pour la prise en charge de l'hypertension artérielle. In F. TRICHET, Ed., *Actes des 18 ^{es} Journées Ingénierie des Connaissances*, p. 241–52, Grenoble, France : Cépaduès. ISBN 978.2.85428.790.5.

STOILOS G., STAMOU G. & KOLLIAS S. (2005). A string metric for ontology alignment. *Lecture notes in computer science*, **3729**, 624.

Patrons de gestion de changements OWL

Rim Djedidi[1] et Marie-Aude Aufaure[2]

[1] Département Informatique, Supélec Campus de Gif
Plateau du Moulon – 3, rue Joliot Curie – 91192 Gif sur Yvette Cedex, France
rim.djedidi@supelec.fr
[2] Laboratoire MAS, Chaire SAP Business Object – Centrale Paris
Grande Voie des Vignes, F-92295 Châtenay-Malabry Cedex, France
marie-aude.aufaure@ecp.fr

Résumé : Tout au long de leur cycle de vie, les ontologies évoluent pour répondre à différents besoins de changements. Nous nous intéressons particulièrement aux problèmes inhérents à la gestion des changements d'une ontologie dans un contexte local et nous présentons une approche d'évolution d'ontologies à base de patrons. Les patrons modélisés correspondent aux dimensions *changement*, *incohérence* et *alternative de résolution*. Sur la base de ces patrons et des liens conceptuels entre eux, nous proposons un processus automatisé permettant de guider et contrôler l'application des changements tout en assurant la cohérence de l'ontologie évoluée.
La gestion des changements étant fortement liée au modèle dans lequel est représentée l'ontologie, nous nous focalisons sur le langage OWL et nous tenons compte de l'impact des changements sur la cohérence logique de l'ontologie telle que spécifiée dans la couche OWL DL.

Mots-clés : Evolution d'ontologies, Gestion de changements, Patrons, Cohérence, OWL DL.

1 Introduction

Tout au long de leur cycle de vie, les ontologies évoluent pour répondre à différents besoins de changements : la dynamicité de l'environnement où elles sont appliquées, l'évolution du domaine modélisé, la modification des besoins utilisateurs, les corrections et restructurations apportées à leur conceptualisation, et leur réutilisation pour d'autres applications.

L'évolution d'ontologies est une problématique complexe (Stojanovic, 2004) (Klein, 2004). Outre l'identification même des besoins de changements à partir de différentes sources (domaine, environnement d'usage, conceptualisation interne, etc.), la gestion de l'application d'un changement – de sa formulation jusqu'à son application et sa validation finale – nécessite de spécifier le changement requis, d'analyser ses effets sur la cohérence de l'ontologie et les résoudre, de l'implémenter et de valider son application finale. Dans un contexte collaboratif ou distribué, il est aussi nécessaire de propager le changement appliqué localement à l'ontologie, aux artefacts dépendants (les applications et/ou les ontologies dépendantes) et de valider les changements globalement. De plus, la traçabilité des changements doit être gardée

pour pouvoir les justifier, les expliquer, voire les annuler et gérer les différentes versions d'une ontologie.

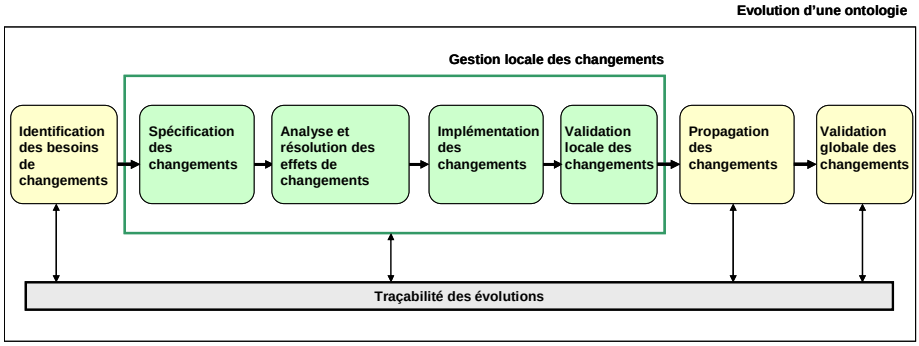


Fig. 1 – Processus global d'évolution d'ontologies

Dans nos travaux[1], nous nous intéressons aux problèmes inhérents à la gestion des changements d'une ontologie dans un contexte local. Conduire l'application des changements tout en maintenant la cohérence de l'ontologie est une tâche cruciale et couteuse en termes de temps et de complexité. Un processus automatisé est donc essentiel. Mais comment conduire l'application d'un changement tout en assurant la cohérence de l'ontologie évoluée ? Comment analyser les effets d'un changement et les résoudre ? Et si plusieurs solutions sont possibles, laquelle choisir et selon quels critères ?

Pour répondre à toutes ces questions, nous avons défini une méthodologie de gestion de changements ***Onto-Evo^{a}l*** (*Ontology Evolution-Evaluation*) qui s'appuie sur une modélisation à l'aide de patrons. Ces patrons spécifient des classes de *changements*, des classes d'*incohérences* et des classes d'*alternatives de résolution*. Sur la base de ces patrons et des liens entre eux, nous proposons un processus automatisé permettant de conduire l'application des changements tout en maintenant la cohérence de l'ontologie évoluée. La méthodologie intègre aussi une activité d'évaluation basée sur un modèle de qualité d'ontologies. Ce modèle est employé pour guider la résolution des incohérences en évaluant l'impact des alternatives de résolution proposées sur la qualité de l'ontologie et ainsi choisir celles qui préservent la qualité de l'ontologie évoluée.

La gestion des changements étant fortement liée au modèle dans lequel est représentée l'ontologie, nous nous focalisons sur le langage OWL et nous tenons compte de l'impact des changements sur la cohérence logique de l'ontologie telle que spécifiée dans la couche OWL DL.

L'article est organisé comme suit : dans la section 2, nous détaillons le processus de gestion de changements. Les patrons de gestion de changements sont présentés et illustrés dans la section 3. Avant de synthétiser les différents points de l'approche définie et présenter nos travaux en cours, une discussion et une comparaison avec les travaux existants sont présentées à la section 4.

[1] Ces travaux sont financés par l'Agence Nationale de Recherche dans le cadre du Projet-RNTL DAFOE.

2 Processus de gestion des changements

L'objectif d'un processus de gestion de changements est de conduire de manière automatisée l'application d'un changement tout en assurant la cohérence de l'ontologie. Ceci nécessite de formuler explicitement le changement requis, de détecter les incohérences dues à son application, de proposer des solutions pour les résoudre, de guider l'ingénieur dans le choix des résolutions et de l'assister dans la validation finale (Stojanovic, 2004).

Pour réaliser cet objectif, nous avons défini des patrons de gestion de changements CMP (*Change Management Patterns*). Ces patrons permettent de ressortir et de classer des types de changements en se basant sur le modèle OWL, des types d'incohérences logiques en se référant aux contraintes de OWL DL et des types d'alternatives de résolution d'incohérences.

Le processus de gestion de changements est conduit à travers quatre phases (figure 2): spécification du changement, analyse du changement, résolution du changement et application du changement.

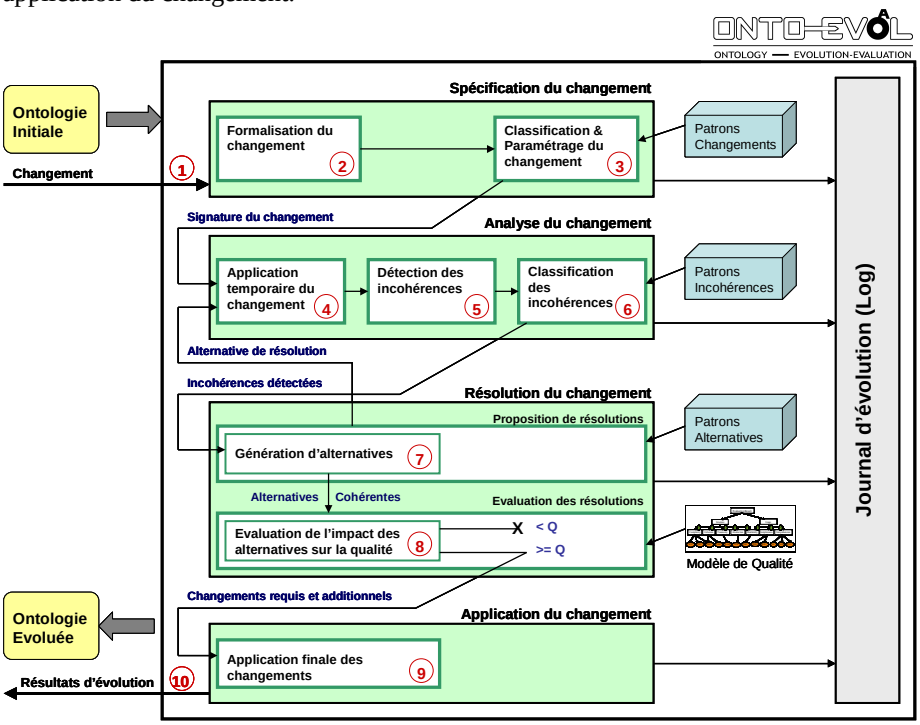


Fig. 2 – Architecture de l'approche de gestion de changements à base de patrons

2.1 Spécification du changement

La phase de spécification est lancée suite à la demande d'un changement à appliquer sur une ontologie initiale supposée cohérente (figure 2 **(1)**). L'utilisateur demande un changement élémentaire ou composé sans avoir à définir et ordonner les opérations de changements intermédiaires nécessaires à son application. Réaliser ces tâches manuellement est très couteux en termes de temps et de risque d'erreurs.

La phase de spécification a un rôle plus large qu'une simple représentation du changement requis dans le langage de l'ontologie. Elle vise à expliciter le changement de manière formelle et compréhensible pour préparer les phases d'analyse et de résolution. Le changement est tout d'abord explicité et formalisé dans le modèle OWL (figure 2 **(2)**). Ensuite, il est classé selon les types prédéfinis par les patrons de changements et enrichi par un ensemble d'arguments permettant de préparer les phases suivantes (figure 2 **(3)**). Le patron correspondant est instancié aux spécificités réelles du changement requis (valeurs, références des entités concernées, etc.). L'ensemble des spécifications dérivées constitue la signature du changement.

2.2 Analyse du changement

Le changement formellement spécifié est appliqué à une version temporaire de l'ontologie pour analyser ses impacts (figure 2 **(4)**). L'étape suivante permet de détecter les incohérences causées (figure 2 **(5)**). Les incohérences détectées sont alors classées selon les types prédéfinis par les patrons d'incohérences (figure 2 **(6)**). L'instanciation des patrons d'incohérences permet de préparer la phase de résolution notamment à travers le lien avec les patrons d'alternatives (figure 3).

Seule la cohérence logique est considérée. La cohérence structurelle – se référant aux contraintes du langage et l'utilisation de ses constructeurs – est vérifiée automatiquement au début du processus. Pour la vérification de la cohérence logique, nous employons le raisonneur Pellet[2] en interfaçage avec le système de gestion de changements que nous avons développé. Pellet supporte aussi bien le niveau terminologique *TBox* (classes et propriétés) que le niveau assertionnel *ABox* (instances) de OWL DL (Sirin et al., 2007). Cependant, certaines incohérences logiques, notamment celles se référant aux propriétés, ne sont pas détectées par Pellet et sont prises en charge par notre système. Par ailleurs, Pellet ne permet pas de préciser les axiomes qui ont causé les incohérences ni comment résoudre les incohérences détectées. L'identification des axiomes causant les incohérences est basée sur les travaux de (Plessers & De Troyer, 2006).

2.3 Résolution du changement

La résolution du changement comprend deux principales activités : proposition de résolutions et évaluation des résolutions.

[2] http://clarkparsia.com/pellet/

2.3.1 Proposition de résolution

A partir des instances d'incohérences détectées, les patrons d'alternatives sont instanciés pour générer les alternatives potentielles de résolution (figure 2 **(7)**). Chaque alternative représente des opérations de changements additionnels à appliquer pour résoudre une incohérence. Cependant, elle ne doit pas causer d'autres incohérences. C'est pourquoi, toutes les alternatives sont vérifiées et résolues selon un mécanisme récursif et seules les alternatives cohérentes sont retenues.

2.3.2 Evaluation des résolutions

Plusieurs solutions peuvent parfois être proposées pour une incohérence. Plutôt que de présenter les différentes alternatives de résolution à l'ingénieur d'ontologies, nous proposons de guider le choix de l'alternative à appliquer en évaluant l'impact de chacune des alternatives sur la qualité de l'ontologie (figure 2 **(8)**). L'évaluation se base sur un modèle de qualité considérant les aspects structure et usage de l'ontologie à travers un ensemble de critères et de métriques. La description détaillée du modèle de qualité ne faisant pas l'objet de ce papier, le lecteur peut se référer à la référence (Djedidi & Aufaure, 2008). L'alternative qui préserve la qualité de l'ontologie, peut être automatiquement choisie. L'évaluation de l'impact sur la qualité participe à l'automatisation du processus en guidant le choix des résolutions à travers des alternatives annotées et évaluées.

2.4 Application du changement

Cette phase correspond à l'application finale du changement. Elle est optimisée par l'emploi de techniques d'évaluation de qualité permettant de guider la résolution des incohérences et de minimiser la dépendance à l'utilisateur. Ainsi, si les résolutions préservent la qualité, le changement requis et ses changements dérivés seront directement validés et appliqués (figure 2 **(9)**) et l'ontologie évoluée. Si par contre, les alternatives ont un impact négatif sur la qualité, les résultats des différentes phases du processus seront présentés à l'ingénieur d'ontologies en complément à son expertise pour qu'il décide du changement (figure 2 **(10)**). Cette phase permet de garder conjointement à l'automatisme du processus, une certaine flexibilité permettant à l'utilisateur de contrôler et de décider du changement et de sa validation finale.

L'ensemble des traitements d'analyse et de maintenance de la cohérence est appliqué à une version temporaire de l'ontologie qui peut être abandonnée si les changements sont finalement annulés, l'ontologie initiale sera alors préservée. Dans le cas contraire, une version modifiée de l'ontologie est définie.

Notons que tout au long du processus, les différents résultats sont sauvegardés dans le journal d'évolution. Le journal d'évolution est une structure permettent de conserver l'historique des évolutions de l'ontologie et les détails de traitement de ces évolutions sous forme de séquences chronologiques d'informations. Il facilite le suivi de l'évolution, le retour arrière, la justification des changements, la gestion des versions et dans une perspective future l'apprentissage de nouveaux patrons de changements, d'incohérences et d'alternatives.

3 Patrons de gestion de changements

Les *Design Patterns* (patrons de conception) représentant des directives (guidelines) partagées qui aident à la résolution de problèmes de conception, ont aussi été adoptés en ingénierie logiciel et ontologique (Gangemi, 2005). Tout comme l'idée de modéliser des *Design Patterns* pour la construction d'ontologies OWL (Gangemi et al., 2007), les patrons de gestion de changements CMP –*Change Management Patterns*– sont proposés comme une solution permettant de ressortir des invariances observées répétitivement lors d'un processus d'évolution d'ontologies. De la modélisation des CMP en ressort trois catégories de patrons : des patrons de *changements*, des patrons d'*incohérences* et des patrons d'*alternatives de résolution*. L'intérêt de cette modélisation est d'offrir différents niveaux d'abstraction, d'établir des liens entre ces trois catégories de patrons déterminant les incohérences qui peuvent être potentiellement causées par un type de changements et les alternatives de résolution possibles pour un type d'incohérences et par là, d'assurer un processus automatisé de gestion de changements.

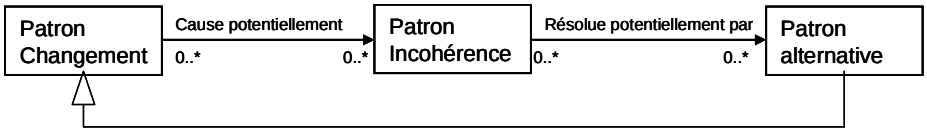


Fig. 3 – Modèle conceptuel des patrons de gestion de changements

3.1 Patrons de changements

Les patrons de changements sont définis sur la base du modèle OWL DL. L'idée est de catégoriser les changements, définir formellement leur signification, leur portée et leurs implications potentielles. Les changements OWL sont classés selon deux grandes catégories : des changements basiques et des changements complexes (Klein, 2004). Les patrons de changements couvrent tous les changements OWL basiques et un premier noyau de changements complexes.

Les composants d'un patron de changement sont:

- Le type des entités concernées correspondant aux primitives conceptuelles de OWL (classe, propriété, instance) ;
- Les arguments regroupant l'ensemble des paramètres nécessaires à l'application du changement. Le contenu varie selon le type du changement et le type des entités concernées et peut comprendre entre autres :
  - La référence des entités réelles sur lesquelles porte le changement (leurs identifiants dans l'ontologie),
  - La référence (les identifiants) des entités intermédiaires impliquées dans le changement (par exemple les superclasses d'une classe à ajouter),

- Les valeurs du changement (par exemple les valeurs d'une restriction de cardinalité).

• Les contraintes à satisfaire pour que le changement puisse être appliqué sans altérer la cohérence logique de l'ontologie. Ce sont des pré-conditions qui peuvent être vérifiées avant l'application du changement. Elles préparent la prévision des incohérences pouvant être causées par un changement. Par exemple, à la demande d'un ajout d'une relation d'équivalence entre deux classes, la pré-condition serait que ces deux classes ne soient pas disjointes dans leurs hiérarchies respectives.

3.1.1 Patrons de changements basiques

Les patrons de changements basiques correspondent à des changements indivisibles qui ne modifient qu'une seule caractéristique du modèle de connaissances de l'ontologie (tel que la suppression d'une relation « *is-a* »).

Exemple1. Soit le patron d'un changement basique correspondant à l'ajout d'une relation de sous-classe. Ce patron est décrit comme suit (table 1) :

Table 1. Exemple de patron de changement basique

Type	**Entités concernées**	**Arguments**	**Contraintes**	**Axiome OWL DL**
P_Chgt_Bas_ Ajouter_ Sous_Classe	Classe, Classe	Sub_classID Super_classID	¬(Sub_classID disjointWith Super_classID)	SubClassOf (Sub_classID, Super_classID)

Pour illustrer une instanciation possible de ce patron, prenons un exemple simple d'une ontologie OWL *O* définie par les axiomes suivants :

{Animal ⊑ Faune-Flore, Plante ⊑ Faune-Flore, Herbivore ⊑ Animal, Carnivore ⊑ Animal, PlanteCarnivore ⊑ Plante, Plante ⊑ ¬Animal}.

Et soit le changement *Ch1* définissant la classe *PlanteCarnivore* comme sous-classe de la classe *Animal*. L'instanciation du patron (table 1) par *Ch1* permet de spécifier la signature suivante (table 2) :

Table 2. Exemple d'instanciation d'un patron de changement basique

Type	**Entités concernées**	**Arguments**	**Contraintes**	**Axiomes OWL DL**
P_Chgt_Bas_ Ajouter_ Sous_Classe	Classe, Classe	Animal, PlanteCarnivore	¬(PlanteCarnivore disjointWith Animal)	SubClassOf (PlanteCarnivore, Animal)

3.1.2 Patrons de changements complexes

Les patrons de changements complexes correspondent à des changements composites et riches renfermant des séquences logiques de changements basiques et incorporant des informations sur leur implication (tel qu'élargir le co-domaine d'une propriété à la superclasse de la classe qui le spécifiait).

Les patrons de changements complexes regroupent plus de détails puisqu'ils décrivent un ensemble de changements intermédiaires. Si un changement requis consiste à ajouter une classe en spécifiant sa place dans la hiérarchie à travers la liste de ses superclasses et/ ou en définissant une collection d'unions ou d'intersections de classes agrémentée d'un ensemble de restrictions, alors c'est un changement complexe correspondant à un patron de changement complexe lui-même décrit par une séquence de patrons de changements basiques (figure 4) et/ou complexes.

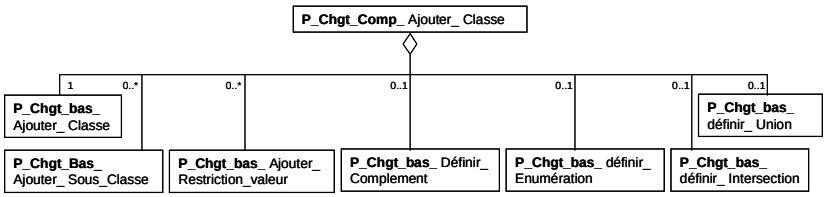


Fig. 4 –Modèle conceptuel d'un patron de changement complexe

3.2 Patrons d'incohérences

Durant la phase d'analyse, le système vérifie les contraintes spécifiées dans la signature du changement, les liens entre le patron de changement instancié et les patrons des incohérences qu'il peut potentiellement causer (figure 3) tout en tenant compte de toutes les incohérences logiques concrètement détectées par le raisonneur Pellet. Toutes les incohérences constatées sont alors classées selon les patrons d'incohérences pour être résolues. Le module de classification se base sur l'interprétation des résultats de Pellet ainsi que les compositions d'axiomes et les dépendances entre axiomes pour identifier les axiomes causant les incohérences.

Les composants des patrons d'incohérences peuvent varier d'un type d'incohérences à un autre mais les principaux composants d'un patron d'incohérence sont :

- Les identifiants de toutes les entités impliquées directement ou indirectement dans l'incohérence logique. Ces informations permettent d'expliquer l'incohérence et facilitent sa localisation ;
- Les identifiants des entités concernées par l'incohérence. Ces informations facilitent la détermination des axiomes causant l'incohérence détectée et préparent la proposition de résolution ;
- Les axiomes concernés par l'incohérence.

Exemple2. Reprenons l'exemple de changement basique *Ch1* (exemple 1), l'instanciation du patron d'incohérence de disjonction correspondant est décrite comme suit (table 3) :

Table 3. Exemple d'instanciation d'un patron d'incohérence de disjonction

Type	**Entités Impliquées**	**Entités Concernées**	**Axiomes OWL DL concernés**
P_Incons _Disj	Animal, Plante, PlanteCarnivore,	Animal, Plante	`Plant ⊑ ¬Animal, PlanteCarnivore ⊑ Plant`

3.3 Patrons d'alternatives

Dans la phase de résolution du changement, la proposition de résolutions se base sur les liens entre le patron d'incohérence instancié et les patrons d'alternatives qui peuvent potentiellement le résoudre (figure 3). Un patron d'alternative représente un changement additionnel à appliquer pour résoudre une incohérence logique. Il est décrit comme un changement (basique ou complexe) et hérite des propriétés d'un patron de changement (figure 3), ce qui implique qu'il peut lui-même causer des incohérences (figure 2). D'autres informations peuvent aussi décrire les patrons d'alternatives telles que les pré-conditions à satisfaire pour choisir le patron comme solution de résolution (table4).

Exemple3. Reprenant l'exemple de changement basique *Ch1* (exemple1), deux alternatives de résolution sont possibles pour résoudre l'incohérence de disjonction causée (exemple2). Leurs patrons et instanciations respectifs sont décrits comme suit :

Table 4. Exemple de patron d'alternative résolvant une disjonction (*al1*)

P_Alt_Disj_ Chgt_Bas_ Ajout_ Sous_Classe **(*al1*)**				
Entités Concernées	**Arguments**	**Pré-conditions**	**Contraintes**	**Axiomes OWL DL**
Classe, Classe	Sub_classID, Super_classID, Id1_cls_disj, Id2_cls_disj	`SuperClass (Id1_cls_disj) ∩ SuperClass (Id2_cls_disj) = Super_classID`	`¬(Sub_classID disjointWith Super_classID)`	`SubClassOf (Sub_classID, Super_classID)`

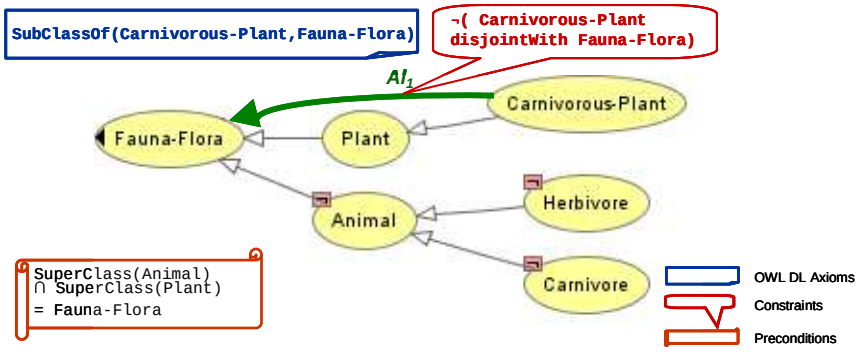


Fig. 5 –Instanciation du patron de l'alternative *al1*

Table 5. Exemple de patron d'alternative résolvant une disjonction (*al2*)

P_Alt_Disj_Chgt_Comp_ Rattacher_ Classe_Hybride **(*al2*) (version synthétisée)**		
Entités Concernées	**Arguments**	**Composants Intermédiaires : Axiomes OWL DL**

Classe,
Classe
	Id_HybridClass,
	Id_sub_class,
	Id1_cls_disj, Id2_cls_disj
		`Class(Id_HybridClass,`
		`UnionOf(Id1_cls_disj,Id2cls_disj))`
		`SubClassOf(Id_HybridClass, Id_sub_class)`

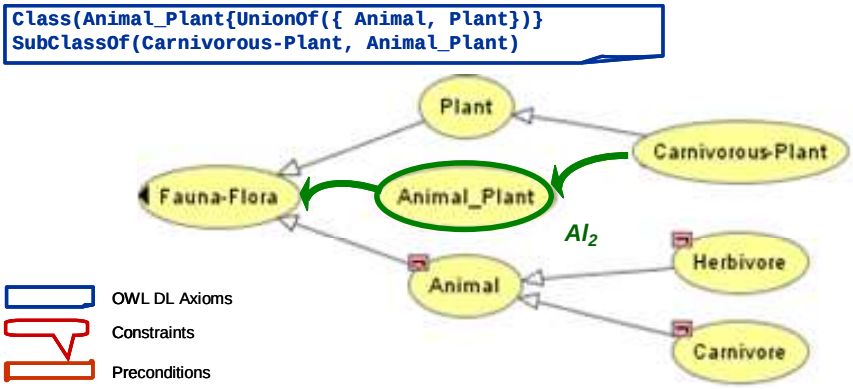


Fig. 6 –Instanciation du patron de l'alternative *al2*

4 Discussion et travaux existants

La modélisation par patrons a été adoptée ces dernières années, dans la conception d'ontologies pour le Web afin de proposer des guides de bonnes pratiques et de fournir des catalogues de composants ontologiques réutilisables[3]. La notion de patrons de conception d'ontologies a été introduite par (Gangemi et al., 2004), (Rector & Roger, 2004), (Svatek, 2004). D'un point de vue théorique, les CMP (*Change Management Patterns*) se rapprochent des ODP (*Ontology Design Patterns*), plus particulièrement des patrons logiques (*Logical Ontology Patterns LOP*). Tous les deux sont appliqués dans le cycle d'ingénierie ontologique. Tout comme les CMP, les LOP sont indépendants du domaine modélisé par l'ontologie et peuvent être appliqués plus d'une fois dans une ontologie pour résoudre un problème d'évolution (CMP) ou de conception (LOP). L'implémentation des CMP et les expressions formelles des LOP sont toutes les deux issues du modèle de OWL DL. Mais d'un point de vue pragmatique, ils sont assez différents : les CMP proposent des solutions réutilisables pour guider la gestion des changements dans un processus d'évolution locale d'une ontologie. Les ODP sont conçus comme des guides de bonnes pratiques réutilisables dans un contexte de conception collaborative d'ontologies en réseau et les LOP représentent des compositions de constructions logiques pour résoudre des problèmes d'expressivité (Presutti et al., 2008).

[3] Exemple : http://odps.sourceforge.net/

Pour les travaux existants en évolution d'ontologies, nous nous positionnons particulièrement par rapport à ceux qui traitent les ontologies OWL, la gestion des changements étant dépendante du modèle de représentation de l'ontologie. Dans (Haase & Stojanovic, 2005), les auteurs ont introduit des stratégies de résolution basées sur les contraintes de OWL Lite. La résolution des incohérences logiques se limite à la détermination des axiomes causant ces incohérences et devant être supprimés et à leur présentation à l'utilisateur pour qu'il puisse décider. Dans notre approche, nous tendons à minimiser les solutions de suppression d'axiomes en proposant des patrons d'alternatives résolvant les incohérences en fusionnant, divisant, généralisant ou spécialisant les concepts pour préserver les connaissances existantes et nous assistons l'ingénieur dans le choix des alternatives à appliquer en faisant intervenir l'évaluation. Dans (Plessers et al., 2006), les auteurs définissent une approche de détection de changements, les changements sont exprimés sous forme de requêtes temporelles appliquées sur un journal de versions. Dans (Plessers et De Troyer, 2006), les auteurs définissent un algorithme localisant les axiomes causant les incohérences dans une ontologie OWL DL que nous adoptons dans notre phase d'analyse de changements. Une autre approche intéressante a aussi été appliquée dans la résolution d'incohérences (Parsia et al. 2005), (Wang et al., 2005) : c'est le débogage d'ontologies. L'objectif est d'offrir à l'ingénieur des explications plus compréhensibles des incohérences que celles fournies par les raisonneurs standards. Deux types de techniques sont distingués : la technique *black-box* considérant le raisonneur comme une boite noire et appliquant des inférences pour localiser les incohérences et la technique *glass-box* qui modifie le mécanisme interne du raisonneur pour expliciter les incohérences et compléter les résultats des raisonneurs.

5 Conclusion et travaux en cours

Dans cet article, nous présentons une approche d'évolution ***Onto-Evo^{a}l*** dont l'objectif est d'optimiser et automatiser la gestion des changements tout en assurant la cohérence et la qualité de l'ontologie évoluée. La principale contribution de nos travaux est d'intégrer une modélisation par patrons dans un processus de gestion des changements d'ontologies permettant de définir et de classer des types de changements, d'incohérences et d'alternatives et de faire ressortir des liens entre ces patrons pour guider et contrôler l'analyse et la résolution des impacts de changements. De plus, l'évaluation de l'impact sur la qualité permet de mieux conduire le processus de gestion des changements et de préserver la qualité de l'ontologie évoluée.

Les patrons définis couvrent les changements OWL DL basiques et un sous-ensemble de changements complexes, un premier noyau d'incohérences logiques et d'alternatives pouvant les résoudre. Un module d'apprentissage est envisagé pour compléter et enrichir ces patrons à travers l'application de l'approche sur des ontologies tests. En effet, il n'est pas évident de fournir un système de gestion complet qui gère tous les types de changements.

Références

DJEDIDI R. & AUFAURE M.A. (2008). Enrichissement d'ontologies : maintenance de la consistance et évaluation de la qualité, 19èmes journées francophones d'Ingénierie des Connaissances (IC'08), Nancy.

GANGEMI A. (2005). Ontology Design Patterns for Semantic Web Content, In. Y. Gil, E. Motta, V..R. Benjamins, M. A. Musen (Eds.) (ISWC'05), Publication Springer-Verlag, LNCS 3729, pp. 262—276. Springer.

GANGEMI A., CATENACCI C. & BATTAGLIA M. (2004). Inflammation ontology design pattern: an exercise in building a core biomedical ontology with descriptions and situations. In D.M. PISANELLI (Ed.) *Ontologies in Medicine. IOS Press*, Amsterdam.

GANGEMI A., GOMEZ-PEREZ A., PRESUTTI V. & SUAREZ-FIGUEROA, M.C. (2007). Towards a Catalog of OWL-based Ontology Design Patterns, CAEPIA 07, Publications du projet Neon (http://www.neon-project.org).

HAASE P. & STOJANOVIC L. (2005). Consistent Evolution of OWL Ontologies, European Conference on Semantic Web Proceedings (ECSW'05), Lecture Notes in Computer Science, (vol. 3532):182-197.

KLEIN M. (2004). Change Management for Distributed Ontologies. Thèse de doctorat, Dutch Graduate School for Information and Knowledge Systems.

PARSIA B., SIRIN E. & KALYANPUR A. (2005). Debugging OWL ontologies. 14ème conférence internationale sur le World Wide Web (WWW2005), Chiba, Japan.

PLESSERS P. & DE TROY O. (2006). Resolving Inconsistencies in Evolving Ontologies. European Conference on Semantic Web Proceedings (ECSW'06), Lecture Notes in Computer Science.

PLESSERS P., DE TROYER O. & CASTELEYN S. (2006). Understanding ontology evolution: A change detection approach. *Journal of Web Semantics*.

PRESUTTI V., GANGEMI A., DAVID S., AGUADO DE CEA G., SUAREZ-FIGUEROA M., MONTIEL-PONSODA E. & POVEDA M. (2008). Library of design patterns for collaborative development of networked ontologies. Deliverable D2.5.1, NeOn project.

RECTOR A. & ROGERS J. (2004). Patterns, properties and minimizing commitment: Reconstruction of the Galen upper ontology in owl. In A. GANGEMI & S. BORGO (Eds.), EKAW'04 Workshop on Core Ontologies in Ontology Engineering. CEUR.

SIRIN E., PARSIA B., CUENCA GRAU B., KALYANPUR A. & KATZ Y. (2007). Pellet: A practical OWL DL reasoner. *Journal of Web Semantics*, 5(2):51-53.

STOJANOVIC L (2004). *Methods and Tools for Ontology Evolution*. Mémoire de thèse, Université de Karlsruhe.

SVATEK V. (2004). Design patterns for semantic web ontologies: Motivation and discussion. 7ème conférence Business Information Systems, Poznan.

WANG H., HORRIDGE M., RECTOR A., DRUMMOND N. & SEIDENBERG J. (2005). Debugging OWL-DL ontologies: A heuristic approach. In Y. GIL, E. MOTTA, V..R. BENJAMINS, M. A. MUSEN (Eds.) (ISWC'05), Publication Springer-Verlag, ISBN 978-3-540-29754-3, Galway, Ireland.

Du texte à la connaissance : annotation sémantique et peuplement d'ontologie appliqués à des artefacts logiciels

Florence Amardeilh[1], Danica Damljanovic[2]

[1] Mondeca. 3, cité Nollez, 75018 Paris, France
florence.amardeilh@mondeca.com
[2] Department of Computer Science. University of Sheffield,
Regent Court, 211 Portobello St, Sheffield S1 4DP, UK
d.damljanovic@dcs.shef.ac.uk

Résumé : Les applications logicielles possèdent généralement une courbe d'apprentissage considérable pour les nouveaux développeurs et pour ceux qui souhaitent en intégrer des parties dans leurs propres applications. L'attrait d'utiliser ici une technologie à base de sémantique repose sur son potentiel à associer un réseau de connaissance aux artefacts logiciels existants, structurés ou non. Ceci se traduit notamment par deux étapes clefs, l'annotation sémantique et le peuplement d'ontologie, qui restent d'importants challenges à résoudre. Nous présentons ici le Content Augmentation Manager, une plateforme servant de médiateur entre les outils d'annotation et les référentiels sémantiques. L'annotation des artefacts logiciels est réalisée par l'outil d'extraction d'information KCIT, capable de produire automatiquement des annotations sur la base d'une ontologie donnée.

Mots-clés : annotation sémantique, acquisition de connaissances, peuplement d'ontologie, artefacts logiciels, référentiels sémantiques, ontologies.

1 Introduction

Réussir la réutilisation de code et éviter les anomalies dans l'ingénierie logicielle requiert de nombreuses qualités, à la fois à la source du code et dans l'équipe de développement. Parmi ces qualités se trouvent la facilité d'identification des composants pertinents et la facilité de compréhension de leurs paramètres et de leurs profils d'usage. L'attrait à utiliser une technologie sémantique pour adresser ce problème repose sur son potentiel à associer un réseau sémantique de connaissance à la documentation logicielle. Cette documentation logicielle provient de multiples artefacts logiciels, comprenant aussi bien des données non structurées comme les articles des forums, les spécifications et les divers manuels que des données structurées issues du code source et des fichiers de configuration. La documentation enrichie peut alors être exploitée afin d'ajouter de nouvelles fonctionnalités, fournir au développeur de nouvelles façons de localiser et d'intégrer les composants, etc.

Deux étapes primordiales de cet enrichissement sont l'annotation sémantique et le peuplement d'ontologie. Dans ce contexte, nous définissons l'**annotation sémantique**

comme une représentation formelle d'un contenu, exprimée à l'aide de concepts, relations et instances décrits dans une ontologie, et liée à la ressource originelle. Les instances sont généralement stockées dans une base de connaissance, indépendamment de la ressource annotée. Elles peuvent donc être réutilisées pour annoter d'autres ressources, offrant par là un point de vue centralisé et consensuel de la connaissance du domaine. L'action d'ajouter des instances à une base de connaissance est appelée **peuplement d'ontologie**. Notons que la notion de peuplement d'ontologie doit être clairement distinguée de celle d'enrichissement d'ontologie. Dans ce cas, il s'agit plutôt d'ajouter de nouveaux concepts et relations au modèle formel de l'ontologie.

Bien que faisant déjà l'objet de nombreuses recherches, ces tâches d'annotation sémantique et de peuplement d'ontologie restent un véritable challenge, pas seulement dans celui de l'ingénierie logicielle mais quelque soit le domaine étudié. En effet, le rôle des humains demeure bien souvent irremplaçable et l'automatisation reste un des plus grands besoins pour de tels outils, particulièrement lorsqu'il s'agit d'annoter de grandes collections de documents (Uren et al, 2006).

Dans ce papier nous présentons notre solution, en mettant l'accent sur deux objectifs : premièrement extraire automatiquement des informations relatives aux artefacts logiciels par rapport à une ontologie de domaine, et deuxièmement résoudre le problème du peuplement automatique de l'ontologie et de l'annotation sémantique des artefacts à partir de ces extractions. Cette solution, appelée Content Augmentation Manager (CA Manager), a pour objectif d'aider à combler le fossé existant entre les outils d'extraction d'information et les référentiels sémantiques qui sont utilisés pour stocker la connaissance qui a été collectée. L'outil d'extraction d'information utilisé ici est le Key Concept Identification Tool (KCIT), capable de produire automatiquement des annotations sémantiques à partir d'artefacts logiciels sans configuration majeure. Son seul paramètre d'entrée est l'ontologie du domaine.

Dans la Section 2 nous discutons des travaux relatifs à notre problématique. Puis nous présentons KCIT dans la Section 3 et le CA Manager dans la Section 4. Les résultats de nos évaluations sont détaillés dans la Section 5. Finalement, nous concluons et envisageons nos futurs travaux dans la Section 6.

2 L'annotation sémantique en pratique

Un outil d'annotation sémantique permet de créer et de gérer un ensemble d'annotations sémantiques à partir d'un document donné. Leur objectif consiste à alléger le fardeau de l'annotation manuelle quelque soit la ressource concernée, et surtout lorsqu'il s'agit de grands volumes de données. La plupart de ces outils ont évolué vers des d'environnements de plus en plus automatisés grâce aux méthodes issues des champs de l'Extraction d'Information et de l'Apprentissage Automatique (Corcho, 2006). Ils peuvent aussi être utilisés pour peupler une ontologie, comme Melita (Ciravegna, 2002), les deux tâches convergeant vers les mêmes types d'outils considérés comme un moyen de capturer la connaissance d'un domaine.

KIM (Kiryakov et al., 2005) s'appuie sur son ontologie générale PROTON[1] pour annoter des pages Web en identifiant automatiquement, à l'aide de dictionnaires et de patrons d'extraction définis dans GATE, les phrases clefs et les entités nommées (personnes, organisations, lieux, etc.). Puis KIM est capable de les relier à l'URI d'une instance particulière dans l'ontologie. Les annotations sont ensuite exploitées pour l'indexation et la recherche sémantique, la co-occurrence et l'analyse des tendances de popularité.

A l'Université de Technologie d'Helsinki en Finlande, des chercheurs ont développé un cadre pour l'annotation automatique (Vehvilinen et al., 2006) et implémentent un outil spécifique, appelé Poka, appliqué au domaine de la Finnish General Upper Ontology YSO[2]. Comme KIM, Poka extrait des entités nommées des textes soumis en entrée mais en considérant leur lemmatisation comme KCIT.

D'autres travaux similaires comprennent les outils S-Cream (Handschuch et al., 2002), MnM (Vargas-Vera et al., 2002), Artequakt (Alani et al., 2003) et OntoSophie (Valarakos et al., 2004). Nous pouvons noter d'importantes différences entre le cadre proposé par le CA Manager et ces approches : certains utilisent les techniques d'apprentissage automatique et d'autres celles du traitement automatique du langage, certains se basent sur des ontologies génériques comme PROTON ou YSO et d'autres sur des ontologies de domaine, ou bien encore soit ils peuplent l'ontologie soit ils annotent les ressources documentaires. La principale différence entre ces plateformes et le CA Manager est que ce dernier préserve l'indépendance entre les outils d'extraction d'information et le référentiel sémantique utilisé, proposant ainsi une capacité d'adaptation importante pour différents besoins applicatifs, comme cela est souvent demandé dans un cadre industriel.

Pour plus de détails sur les plateformes existantes d'annotation sémantique, nous référons le lecteur aux études (Uren et al., 2006) et (Reeve & Han, 2005).

3 L'extraction d'information à partir d'artefacts logiciels

L'extraction d'information est habituellement la première étape pour opérer des tâches de peuplement d'ontologie et d'annotation sémantique. Le processus automatique de production d'extractions sur la base d'une ontologie n'est pas trivial. En effet, le langage utilisé pour décrire les concepts et les relations dans les ontologies peuvent grandement différer du langage des contenus textuels et le langage naturel humain est lui-même bien connu pour son ambiguïté et sa complexité (Church & Patil, 1982). La plupart des approches utilisent des listes statiques de dictionnaires et repèrent seulement le terme exact issu d'un contenu parmi ceux de la liste. Nous avons développé l'outil KCIT (Key Concept Identification Tool) pour retrouver automatiquement des concepts clefs depuis des contenus textuels par rapport à une ontologie de référence. Les extractions sont créées en se basant sur l'hypothèse qu'une partie spécifique d'un contenu se réfère à une instance particulière si les deux lemmes

[1] Site web de PROTON : http://proton.semanticweb.org/

[2] Site web de YSO : http://www.seco.tkk.fi/ontologies/yso/

correspondent. En faisant se correspondre ces lemmes, nous nous assurons que toutes les flexions morphologiques des termes pertinents sont repérées.

Le processus d'extraction de notre outil KCIT comprend plusieurs étapes :

- *Construire une liste des termes pertinents*. Etant donné une ontologie, la lexicalisation de toutes les ressources ontologiques (classes, instances, propriétés, valeurs de propriétés) sont lemmatisées et ajoutées à la liste du vocabulaire contrôlé, dite "gazetteer list" dans GATE.
- *Annoter les contenus*. Le contenu des artefacts logiciels (plus ou moins structuré selon la nature de ces artefacts) est d'abord lemmatisé puis comparé avec le vocabulaire contrôlé construit à l'étape précédente.
- *Résoudre les conflits*. Les extractions sont filtrées afin de résoudre les problèmes d'ambiguïté tels que supprimer les extractions redondantes.

Les prochaines sections de cet article décrivent ces étapes en détail.

3.1 Construire une liste des termes pertinents

Pour initialiser l'outil KCIT, nous prétraitons les ressources ontologiques (classes, instances, propriétés et valeurs de propriétés) afin d'extraire toute lexicalisation compréhensible par un humain : la partie identificatrice de leur URI, leurs libellés (i.e. label) et les valeurs de leurs propriétés instanciées. Un ensemble de règles heuristiques est appliqué à chaque item de cette liste. Bien qu'il ne soit pas nécessaire de configurer spécifiquement KCIT pour l'utiliser avec différentes ontologies, ce paramétrage permet néanmoins d'obtenir de meilleurs résultats. Voici ci-dessous des exemples de règles appliquées par défaut par KCIT :

- Les caractères tiret ("-") ou soulignement ("_") sont remplacés par des espaces blancs. Exemple : *Project_Name* devient *Project Name*.
- Les mots comme *camelCased* sont découpés dans leurs mots constituants. Exemple : *projectName* devient *Project Name*.
- Si le texte contient des "stop words" comme *of, in, the*, etc., KCIT ignore aussi le texte qui suit. Exemple : *pos tagger for french* donne deux syntagmes, le texte original et *pos tagger*.

Chaque item de cette liste est analysé séparément par l'application Onto Root (en haut à droite de la Fig. 1), composée de plusieurs modules génériques de traitement du langage fournit par GATE (Cunningham et al., 2006). Chaque item est tout d'abord découpé en unité textuelle (Tokeniser) à laquelle est assignée une information syntaxique (POS Tagger) et un lemme (Morpho). C'est ce lemme (ou ensemble de lemmes) qui sera ajouté au vocabulaire dynamique (Ontology Resource Root Gazetteer). Par exemple, si une ressource est identifiée *ProjectName* par son URI et libellée *project names* dans l'ontologie, la liste créée avant d'exécuter Onto Root contient les termes suivants : *ProjectName* comme identifiant, *Project Name* comme découpage de l'identifiant et *project names* comme libellé. Chacun de ces items est alors analysé séparément des autres par Onto Root, produisant la même sortie pour *ProjectName* et *Project Name* mais ajoutant le lemme *project name* à partir du libellé *project names*.

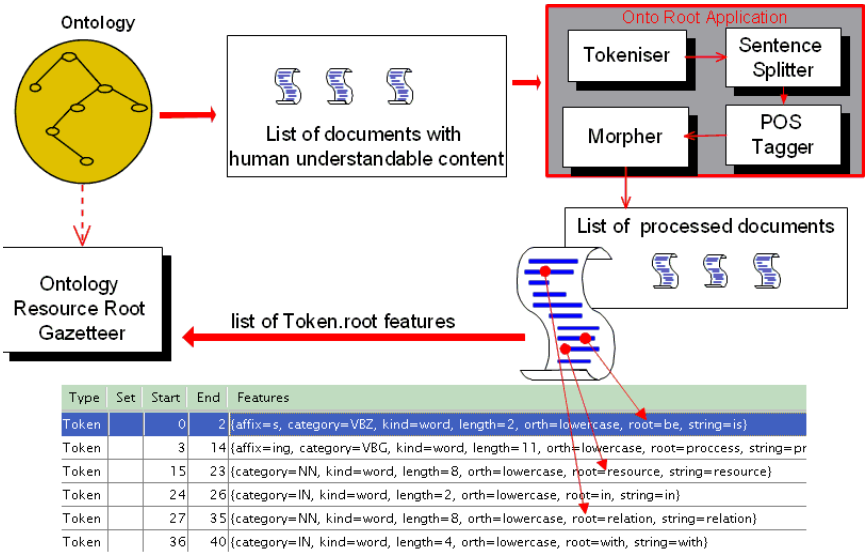


Fig. 1 – Construction automatique d'un vocabulaire à partir de l'ontologie

3.2 Annoter l'artefact logiciel avec les termes extraits

Une fois la liste des termes pertinents créée, la recherche de termes peut être appliquée sur notre corpus d'artefacts logiciels. A cause des variations morphologiques existant en anglais, comme d'en beaucoup d'autres langues, le comportement par défaut n'est souvent pas suffisant pour fournir la flexibilité requise et faire correspondre toutes les flexions morphologiques des termes pertinents. Afin de pouvoir comparer les lemmes créés lors de l'extraction vis-à-vis du vocabulaire dynamique généré à partir des termes ontologiques, la création des lemmes du contenu de l'artefact se fait aussi au moyen de l'outil Onto Root. Ainsi les lemmes extraits des ressources ontologiques et ceux de l'artefact analysé peuvent être comparés en faisant correspondre toutes les flexions morphologiques existantes des termes identifiés comme pertinents.

Les artefacts logiciels présentent toutefois un nouveau challenge pour les outils génériques de traitement du langage servant à délimiter les unités textuelles et les phases. En effet, certains de ces artefacts sont semi-structurés et contiennent des noms de variables, composés de un ou plusieurs mots (e.g., *getDocumentName*). Par conséquent, les ressources ANNIE English Tokeniser et Sentence Splitter (Cunningham et al., 2006) de GATE ont été paramétrées pour pouvoir analyser un code source Java et la JavaDoc associée. Ainsi *getDocumentName* est séparé en *get*, *Document*, et *Name*, avant d'être soumis aux algorithmes d'extraction.

3.3 Résoudre les conflits – Répondre au problème de l'ambiguïté

Nous l'avons dit, le langage humain est ambigu et il est possible d'utiliser la même expression dans différents contextes avec différents sens. Le processus d'analyse de

KCIT peut aussi conduire à la production d'annotations ambiguës à partir de la même unité textuelle ou d'un ensemble d'unités textuelles. Nous appliquons donc plusieurs règles heuristiques basées sur la structure de l'ontologie afin de résoudre ce problème d'ambiguïté. Par exemple, dans l'ontologie de domaine de GATE[3], il existe une instance, "*ANNIE POS Tagger*", partageant la même valeur pour son libellé et sa propriété *resourceHasName*. Cette expression comprend aussi le libellé de la classe *POS Tagger*. Lorsque le texte *ANNIE POS Tagger* apparaît dans un document, trois annotations sont créées qui nécessitent ensuite d'être désambiguïsées. Dans l'interface graphique, elles apparaissent comme des étiquettes se chevauchant (cf. *Start* et *End* de la Fig. 2). Dans l'ontologie GATE, la classe *POS Tagger* possède quatre instances dont *ANNIE POS Tagger*. Il devient donc possible de désambiguïser facilement les annotations à partir de l'instance correcte.

En sortie du processus, nous obtenons un ensemble d'extractions et leurs propriétés : l'URI de la ressource ontologique auquel le terme se réfère, son type (instance, classe, ou propriété) et d'autres caractéristiques pouvant être utilisées ultérieurement.

Type	Set	Start	End	Features
OntoRes		0	16	{instanceURI=http://gate.ac.uk/ns/gate-ontology#ANNIEANNIEPOSTagger, propertyName=resourceHasName
OntoRes		0	16	{URI=http://gate.ac.uk/ns/gate-ontology#ANNIEANNIEPOSTagger, type=instance}
OntoRes		6	16	{URI=http://gate.ac.uk/ns/gate-ontology#POSTagger, type=class}

Fig. 2 – Annotations ambiguës pour la chaîne d'entrée *ANNIE POS Tagger*

4 La plateforme Content Augmentation Manager

L'outil KCIT a été encapsulé dans une plateforme plus générale de production d'annotations sémantiques et de peuplement d'ontologie, nommée CA Manager. En fait, la philosophie au cœur du CA Manager est de pouvoir combler le fossé existant entre des outils d'extraction tel que KCIT et des référentiels sémantiques tel qu'ITM[4] ou Sesame[5]. Il a donc été conçu comme un médiateur, proposant une infrastructure modulaire et flexible qui lui permet de s'adapter à divers flux d'applications clientes et ce, quelque soit l'ontologie du domaine. Il est aussi capable de contrôler la qualité et la validité des résultats de l'extraction d'information par rapport à une ontologie donnée, de les confronter avec d'autres ressources existantes (internes ou externes) et de les enrichir.

Pour ce faire, le CA Manager repose à la fois sur les recommandations formulées par la communauté Web Sémantique (formats RDF/OWL, services web) ainsi que sur une infrastructure UIMA[6] (Unstructured Information Management Architecture) qui a été enrichie et configurée pour les tâches d'annotation et de peuplement.

[3]Ontologie décrivant l'application GATE : http://gate.ac.uk/ns/gate-kb

[4] ITM : http://mondeca.com/index.php/en/intelligent_topic_manager

[5] Sesame : http:// www.openrdf.org/

[6] UIMA : UIMA: http://www.alphaworks.ibm.com/tech/uima

4.1 Une plateforme web sémantique basée sur UIMA

L'infrastructure UIMA a pour objectif de fournir une plateforme de développement pour les outils de traitement automatique du langage naturel. Grâce à sa facilité de composition et d'intégration de modules internes ou externes, il a été rapidement accepté et recommandé par la communauté de l'Extraction d'Information (EI) et c'est pourquoi nous l'avons choisi comme fondation du CA Manager.

Néanmoins, bien qu'UIMA fournisse les bases pour développer un processus d'acquisition de connaissance, il ne donne pas pour autant des conseils sur ses étapes et leur enchaînement, notamment pour contrôler la validité des nouvelles annotations et instances. D'autre part, le Common Analysis Structure" (CAS) définit un schéma d'annotation de haut niveau qui doit être redéfini plus finement pour chaque nouveau besoin. Enfin, UIMA utilise une manière propriétaire d'exposer les web services (the Vinci IBM protocol), qui n'est pas réutilisable avec les standards prônés par la communauté Web Sémantique.

Dans la mise en œuvre de la plateforme du CA Manager (Martin et al., 2008), nous avons donc à la fois repris les préceptes formulés par UIMA comme le développement d'une architecture basée sur plusieurs moteurs d'analyse (Analysis Engines) permettant une configuration flexible et aisée des différents processus des applications clientes. Mais nous avons aussi tenté de faire évoluer l'infrastructure UIMA vers une exploitation des standards du Web Sémantique. Nous avons ainsi formalisé un schéma d'annotation en RDF, inspiré du CAS afin qu'il soit échangeable et enrichi au fur et à mesure du flux d'acquisition de connaissance, mais surtout générique aux tâches d'annotation et de peuplement quelque soit le domaine traité par l'ontologie de l'application. Enfin, nous avons développé la possibilité de déployer un processus de manière distribuée pour en améliorer la performance comme UIMA mais accessible via des web services reposant sur les langages et protocoles du Web Sémantique.

4.2 Extraire, Consolider et Stocker la connaissance

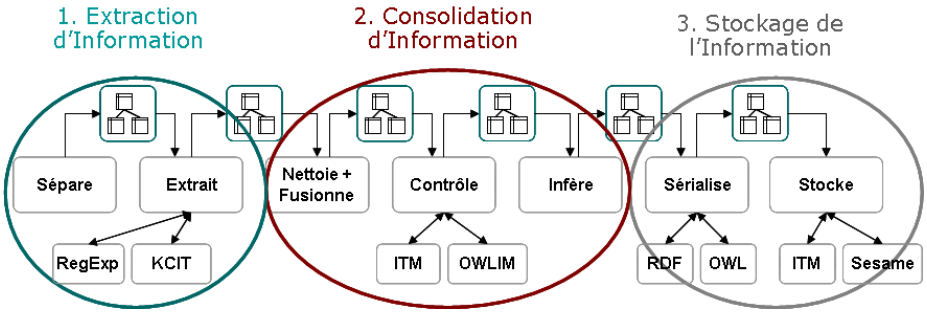


Fig. 3 – CA Manager : un processus modulaire et flexible

Le CA Manager propose donc un flux d'étapes logiques, cf. Fig. 3, dont certaines sont optionnelles, enrichissant progressivement le schéma d'annotation prédéfini. Ces

étapes peuvent être groupées en 3 modules principaux : 1) *Extraire* la connaissance pertinente et annoter le contenu; 2) *Consolider* les résultats vis à vis du modèle de l'ontologie et du référentiel sémantique; 3) Sérialiser la connaissance et le schéma d'annotation dans divers formats et les *stocker* dans le référentiel sémantique.

4.2.1 Extraire la connaissance depuis les textes

Le module d'Extraction d'Information est composé de deux étapes, *sépare* et *extrait*. La première étape, *sépare*, est optionnelle. Le contenu soumis en entrée au CA Manager est divisé en de multiples parties, e.g. un corpus en un ensemble de documents, ou un document en sous-sections bien identifiées.

La seconde étape, *extrait*, appelle l'outil d'extraction paramétré dans le flux, ici KCIT pour le traitement des artefacts logiciels, et récupère les occurrences des entités trouvées dans les artefacts. Un arbre conceptuel est généré en sortie, souvent dans un format propriétaire XML, qui a besoin d'être traduit dans le schéma d'annotation du CA Manager. Un mapping est préalablement créé manuellement entre toutes les entités pouvant être extraites et l'ontologie du domaine afin de générer ensuite automatiquement le schéma d'annotation de l'artefact analysé. Ce mapping est implémenté sous la forme d'un ensemble de règles d'acquisition de connaissance (RAC) tel que décrit dans (Amardeilh, 2007).

4.2.2 Consolider la connaissance grâce au référentiel

Comme indiqué dans (Alani et al., 2003), rares sont les outils pour le peuplement d'ontologie et l'annotation sémantique qui décrivent, ou même mentionnent, cette phase de consolidation dans leurs processus. Pourtant, cette phase est extrêmement importante afin d'assurer l'intégrité et la qualité du référentiel de l'application au moment du peuplement de l'ontologie. En fait, la plupart d'entre eux reposent seulement sur une validation manuelle pour contrôler les instances nouvellement ajoutées à la base de connaissance. A contrario, le CA Manager se distingue de ces outils en mettant l'accent sur l'automatisation de la consolidation des informations extraites en fonction de l'ontologie et de la base de connaissance. Afin de préserver l'intégrité de cette base de connaissance, la phase de consolidation est effectuée avant la création des instances dans le référentiel. Chaque algorithme de consolidation du CA Manager prend en compte deux axes : 1) la ressource ontologique concernée (instance de classe ou valeur de propriété); 2) les contraintes à contrôler (non redondance de l'information, restrictions de domaine et de couverture, cardinalités). Pour plus de détails sur ces algorithmes de consolidation, voir (Amardeilh, 2007). Dans le CA Manager, le module de consolidation se découpe en trois étapes, principalement *fusionne* et *contrôle*, et optionnellement *infère*.

La première étape, *fusionne*, vérifie que l'entité extraite n'existe pas en double dans le schéma d'annotation mais aussi dans la base de connaissance. Le référentiel sémantique est donc requêté afin de retrouver l'URI d'une instance présente dans la base de connaissance et correspondant à l'entité extraite. Les requêtes peuvent être simples (ex : nom de la classe et la chaîne de caractère identifiée dans le texte) ou multicritères (ex : nom de la classe et un ensemble de valeurs de propriétés identifiant

de manière non ambiguë une instance dans le référentiel). Par exemple, une personne peut être recherchée par son nom, comme l'instance *Niraj Aswani* de la classe *GATEDeveloper*. Mais dans des cas d'homonymie par exemple, rechercher le nom d'une personne n'est pas suffisant et il est plus pertinent d'interroger une combinaison de propriétés particulières et discriminantes (comme la date de naissance) pour filtrer parmi plusieurs instances ayant le même libellé.

L'étape *contrôle* vérifie que l'entité extraite est valide par rapport au modèle de l'ontologie. Ceci signifie contrôler le respect de l'appartenance à la classe la plus spécifique dans la hiérarchie, des domaines et couvertures pour les propriétés, des cardinalités, des formats des dates et des numériques, etc. Les triplets invalides sont étiquetés comme tels par une métadonnée de statut et stockés à part dans le référentiel sémantique sur le serveur. Ils peuvent être retrouvés ultérieurement afin notamment d'être présentés à un utilisateur final pour être manuellement validés.

La dernière étape de ce composant, *infère*, est optionnelle. Le composant peut se connecter à un moteur de raisonnement et d'inférence afin de découvrir de nouvelles entités pour relations entre elles mais aussi pour contrôler la cohérence globale et la qualité du référentiel sémantique après ces nouveaux ajouts.

4.2.3 Enregistrer les annotations et la connaissance dans les référentiels

Le composant Information Storage possède deux étapes, *sérialise* et *stocke*. L'étape *sérialise* parcourt le schéma d'annotation enrichi et consolidé par les étapes précédentes afin de générer une sortie dans le format requis par l'application cible (XML, RDF, OWL, etc.). La seconde étape, *stocke*, est optionnelle suivant si l'application utilise directement le format sérialisé ou si elle stocke les résultats dans un référentiel comme ITM et/ou dans un serveur d'annotation comme Sesame.

Le processus décrit ci-dessus est exposé comme un web service. Le flux de traitement de chaque application (ontologie du domaine, enchaînement des étapes, appels vers outils externes et format de sérialisation) doit être préalablement décrit dans un fichier de configuration situé sur le serveur. Une interface de test est fournie à l'adresse http://62.210.155.132/ca-test/. Nous avons aussi développé une interface pour la validation simultanée des annotations et des instances créées à partir du même artefact. Celles ayant été jugées *invalides* par les algorithmes de consolidation sont également récupérées et présentées à l'utilisateur comme "à valider". Il peut alors les corriger, sans risque de rendre le référentiel incohérent puisque l'interface repose entièrement sur l'ontologie qui contraint la saisie de l'utilisateur.

5 Evaluation

Dans le cadre du projet TAO[7] nous avons implémenté plusieurs processus pour tester la flexibilité du CA Manager : 1) ontologie PROTON + corpus d'articles de presse + expressions régulières simples + référentiel Sesame; 2) ontologie GATE + corpus d'artefacts logiciels + KCIT + référentiel Sesame; 3) ontologie GATE +

[7] TAO website : http://www.tao-project.eu

corpus d'artefacts logiciels + KCIT + référentiel ITM. Nous avons réalisé une évaluation de la chaîne complète de traitement sur la base du dernier processus à partir des mesures classiques de précision et de rappel sur la base d'un corpus constitué de 20 documents de la plateforme d'ingénierie textuelle GATE. Nous avons choisi différents types d'artefacts pour constituer un corpus représentatif, décomposé comme suit: 4 articles de forum de la mailing liste de GATE ; 3 classes java du code source de GATE ; 7 chapitres du manuel utilisateur de GATE ; 3 publications au sujet de GATE ; 2 pages Web accessibles à partir du site web gate.ac.uk ; 1 guide du développeur de GATE.. Mais comme les résultats dépendent grandement de la complexité et de la nature de l'ontologie utilisée, nous réétudions actuellement cette évaluation avec les mesures Learning Accuracy (Hahn & Schnattinger, 1998) et Balanced Distance Measure (BDM) (Maynard et al., 2008), qui permettent de prendre en compte ces contraintes.

Table 1. Precision and recall measures for the selection of the GATE software artefacts

	Precision	**Recall**
4 Forum posts	98.6111111	100.0
7 chapters of the GATE User Manual	96.0007672	96.9611548
2 Web pages	94.379845	95.0787402
3 publications	89.796798	95.8392268
3 java classes	97.2592593	98.8636364
1 GATE application developers guide	96.484375	98.4063745
Total	94.2830463	96.8763326

Nous avons aussi évalué plus spécifiquement la performance de l'outil d'extraction KCIT (cf. section 3) sur le même corpus à l'aide des mesures de précision et de rappel. Nous avons tout d'abord annoté manuellement ces documents afin de créer un corpus étalon de référence. Puis, nous l'avons traité automatiquement avec KCIT et comparé les résultats avec le corpus étalon. Les résultats sont présentés dans la Table 1.

Pour les 20 documents sélectionnés, 4523 annotations créées sont correctes, 41 annotations sont partiellement correctes, 126 annotations manquent, et 255 sont fausses. En regardant de plus près les documents annotés, la majorité des annotations manquantes ou fausses sont dues à la sortie du Morphological Analyser. Par exemple, l'analyseur n'a pas réussi à extraire correctement la racine des mots lorsque le pluriel des acronymes (ou des termes *camelCased*) était utilisé. Par exemple, la racine extraite pour *LanguageResources* restait *LanguageResources*. D'autre part, beaucoup d'annotations ont été créées dues au terme *learn*, bien que seulement une minorité soit correcte. La raison est que KCIT ne repose pas sur le contexte pour décider si un terme est pertinent ou non. Ainsi, chaque occurrence du mot *learn*, même dans la phrase « *learn GATE using movie tutorials* » était annoté comme référant au plugin appelé *learning*. De plus, certaines annotations se chevauchant n'étaient pas filtrées par KCIT, ceci affectant la performance globale. Par exemple, le terme *Stemmer PR* est annoté en référence à *Stemmer PR*, instance de la classe *Processing Resource*. Mais la partie libellée *PR* est aussi annotée, en référence à la classe *Processing*

Resource. Pourtant, bien que correcte, cette seconde annotation doit être supprimée durant la phase de filtrage de KCIT, car elle est redondante.

En général, nous pouvons conclure que la performance de KCIT est satisfaisante dans les cas où les documents sont courts et spécifiques au domaine. Par exemple, les annotations des articles de forums ou des classes java étaient produites avec une précision et surtout un rappel très bons. Plus les documents sont importants, moins le vocabulaire utilisé est spécifique au domaine (comme dans le cas des publications), et ainsi la performance se dégrade, mais toujours en restant à un niveau raisonnable. La raison de ceci, comme déjà mentionné, est que KCIT ne fait pas usage du contexte et échoue à filtrer les annotations qui sont étiquetées avec des concepts qui sont clairement en dehors du périmètre du domaine, mais partagent le même nom. Par exemple, *features* dans le contexte des artefacts logiciels de GATE est souvent utilisé pour décrire les caractéristiques des annotations générées par GATE. Cependant, lorsque le contexte traite des caractéristiques globales d'un composant logiciel, cette annotation du terme *features* a besoin d'être supprimée.

6 Conclusion et travaux futurs

Nous avons présenté la plateforme du CA Manager et son outil d'extraction d'information KCIT. KCIT est capable d'enrichir automatiquement ses dictionnaires à partir des libellés extraits et lemmatisés de l'ontologie de domaine pour repérer des termes liés à des artefacts logiciels. Le CA Manager sert de médiateur et consolide ces extractions par rapport au référentiel sémantique, peuple l'ontologie, crée les annotations sémantiques décrivant ces artefacts et stocke la nouvelle connaissance dans le référentiel. Cette plateforme innove par la combinaison d'une infrastructure UIMA et des technologies du Web Sémantique. Nous avons présenté les résultats d'évaluation de l'outil KCIT qui ont montré que sa performance dans des domaines restreints, comme celui de l'ingénierie logicielle, est raisonnable. Comme la performance de l'outil dépend de la couverture de l'ontologie du domaine, nous pensons qu'il est assez fiable pour couvrir n'importe quel domaine du moment que l'ontologie contient assez de données. Cependant, comme montré par les résultats d'évaluation, la phase de filtrage nécessite d'être améliorée, ce qui devrait réduire le nombre d'annotations ambiguës basées sur le contexte. Enfin, nos travaux futurs impliquent de finaliser l'évaluation du CA Manager en mettant à l'épreuve de nouvelles métriques comme la Learning Accuracy (LA) et la BDM.

Remerciements. Cette recherche est partiellement financée par le projet Transitioning Applications to Ontologies (TAO) du EU Sixth Framework Program (FP6-026460).

Références

ALANI H., KIM S., MILLARD D. E. (2003). Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation. In Knowledge Capture Conference (K-CAP'03), Workshop on Knowledge Markup and Semantic Annotation, Florida.

AMARDEILH F. (2007). Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle. Thèse de doctorat, Univ. Paris X.

CHURCH K. & PATIL R. (1982). Coping with syntactic ambiguity or how to put the block in the box. In American Journal of Computational Linguistics, 8(3-4).

CIRAVEGNA F., DINGLI A., PETRELLI D., WILKS Y. (2002). User-system cooperation in document annotation based on information extraction. In 13th Int. Conf. on Knowledge Engineering and Management (EKAW'02), LNCS 2473, Springer-Verlag. p. 122-138. Madrid.

CORCHO O. (2006). Ontology based document annotation: trends and open research problems. In Int. J. Metadata, Semantics and Ontologies, 1(1), Inderscience. p. 47-57.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V., URSU C., DIMITROV M., DOWMAN M., ASWANI N., ROBERTS I. (2006). Developing Language Processing Components with GATE Version 3.1 (a User Guide).

FELLBAUM C. (1998). WordNet - An Electronic Lexical Database. MIT Press.

HAHN U. & SCHNATTINGER K. (1998) Towards text knowledge engineering. In 15th National Conference on Artificial Intelligence (AAAI). p. 524-531, Menlo Park, CA, MIT Press.

HANDSCHUCH S., STAAB S., CIRAVEGNA F. (2002) S-CREAM - Semi-automatic creation of metadata. In 13th Int. Conf. on Knowledge Engineering and Management (EKAW'02), LNCS 2473, Springer-Verlag. p. 379-391. Madrid.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D., KIRILOV A., GORANOV M. (2005). Semantic annotation, indexing, and retrieval. In J. Web Semantics, Science, Services and Agents on the WWW, 2(1), Elsevier. p. 49-79.

MARTIN J., HERRERO G., CAPELLINI A., FRANCART T., AMARDEILH F., MARINOVA Z. (2008). TAO Suite: Architecture and integration requirements and specifications. Deliverable D5.2, TAO project IST-2004-026460.

MAYNARD D., PETERS W., YAOYONG LI. (2008). Evaluating evaluation metrics for ontology-based applications: infinite reflections. In Int. Conf. on Language Resources and Evaluation, Marrakech, Morocco.

REEVE L. & HAN H. (2005). Survey of semantic annotation platforms. In Symposium on Applied Computing (SAC'2005), Santa Fe, New Mexico, USA, p. 1634-1638.

UREN V., CIMIANO P., HANDSCHUH S., VARGAS-VERA M., MOTTA E., CIRAVEGNA F. (2006). Semantic annotation for knowledge management: requirements and a survey of the state of the art. In J. Web Semantics, Science, Services and Agents on the WWW, 4(1). p. 14-26.

VALARAKOS A., PALIOURAS G., KARKALETSIS V., VOUROS G. (2004). Enhancing the Ontological Knowledge through Ontology Population and Enrichment. In 14th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW'04), Lecture Notes in Artificial Intelligence, Vol. 3257, Springer-Verlag. p. 144-156.

VARGAS-VERA M., MOTTA E., DOMINGUE J. (2002). MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In 13th Int. Conf. on Knowledge Engineering and Management (EKAW'02), LNCS 2473, Springer-Verlag. p. 379-391. Madrid.

VEHVILINEN A., HYVNEN E., ALM O. (2006). A semi-automatic semantic annotation and authoring tool for a library help desk service. In First Semantic Authoring and Annotation Workshop.

WITTE R., ZHANG Y., RILLING J. (2007). Empowering software maintainers with semantic web technologies. In 4th European Semantic Web Conference.

Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du Web sémantique

Ines Jilani[1], Florence Amardeilh[2]

[1]INSERM, UMR S 872, Éq. 20, Les Cordeliers, Paris, F-75006 France ; Université Pierre et Marie Curie-Paris6, UMR S 872, Paris, F-75006 France ; Université Paris Descartes, UMR S 872, Paris, F-75006 France,
ines.jilani@spim.jussieu.fr
[2]Modyco, UMR 7114, Université Paris 10, F-92001 Nanterre Cedex, France
florence.amardeilh@mondeca.com

Résumé : Collecter, lire, interpréter et annoter une grande masse de données textuelles n'est pas chose facile depuis le développement des nouvelles technologies dont Internet qui propose pléthore d'informations. Ces tâches sont d'autant plus fastidieuses à mener dans le domaine de la biologie où les intervenants doivent constamment être informés des nouveautés mais aussi réaliser des expériences sur la paillasse pour publier à leur tour leurs travaux et rester concurrentiels. Cet article propose de construire une ontologie, de la peupler automatiquement grâce à une méthode de traitement automatique des langues : les patrons lexico-syntaxiques. Une évaluation de l'extraction de connaissances est réalisée et présente une précision de 72% ainsi qu'un rappel de 50%.

Mots-clés : Ingénierie des connaissances, Apprentissage machine, Jeunes chercheurs en Intelligence Artificielle, Peuplement ontologique.

1 Introduction

Les nouvelles techniques apportées par les progrès dans le domaine de la biologie moléculaire comme le séquençage des génomes, ou les puces à ADN produisent une masse de données si grande que les biologistes ont du mal à s'y retrouver. L'accès à ces données mais aussi leur interprétation ainsi que leurs annotations sont des tâches trop fastidieuses à réaliser avec les outils classiques du biologiste.

Une autre difficulté vient s'ajouter à l'augmentation du volume des données, il s'agit des nombreuses terminologies employées par les biologistes : chacune est établie pour un sous domaine précis de la biologie. Or, un même terme peut parfois avoir des significations différentes selon que l'on s'intéresse à un sous domaine ou à un autre. Par ailleurs, un grand nombre de bases de données spécialisées voient le jour et sont accessibles à tous sur la toile. Cependant elles utilisent généralement des entrées différentes, mais aussi des terminologies propres à chacune. Ainsi, les

biologistes ne peuvent avoir une vision globale d'une connaissance disponible à partir d'une seule et même source car il n'existe pas de base de données qui mutualise toutes celles disponibles en biologie.

Nous exposons tout d'abord un bref état de l'art concernant le domaine de travail et les méthodes utilisées, puis nous introduisons le contexte de l'étude. Ensuite, nous présentons la modélisation de l'ontologie, l'acquisition des connaissances sur les micro Acides RiboNucléiques (miARN) puis son peuplement effectif grâce aux connaissances extraites automatiquement avant de finir par conclure.

2 L'apport du Web Sémantique

Le Web Sémantique consiste à décrire le contenu de ses ressources en les annotant avec des informations non ambiguës afin de favoriser l'exploitation de ces ressources par des agents logiciels (Prié & Garlatti, 2004). Or, les données actuelles du Web sont souvent écrites en langage naturel, car destinées aux humains. Le langage naturel étant par essence trop ambigu, des alternatives formelles et sémantiquement explicites doivent être mises en place pour lever les ambiguïtés du langage naturel, aussi bien dans le contenu des ressources que dans leurs annotations.

Les ontologies, originaires des techniques de modélisation de la connaissance notamment développées en intelligence artificielle, sont utilisées dans le domaine de la biologie afin de proposer un ensemble structuré de tous les termes représentant le sens d'un champ d'information, une sorte de description de la structure des informations disponibles sur le sujet. L'ontologie constitue en soi un modèle représentatif de l'ensemble des concepts dans le sous domaine de la biologie concerné, ainsi que les relations entre ces concepts. Autrement dit, elle fournit les moyens d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés dans un langage de représentation des connaissances formel favorisant le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage (Bourigault, Aussenac-gilles, & Charlet, 2004).

L'exploitation des outils du Web Sémantique, et notamment l'exploitation des ontologies du domaine pour la tâche d'enrichissement automatique de bases de connaissances est encore innovante. Néanmoins il existe des systèmes (KEOPS (Brisson & Collard, 2007), (Hignette, 2007)) qui s'intéressent à l'extraction d'informations, à l'annotation sémantique, basées sur des ontologies de domaine ou non.

La mise en œuvre du peuplement de telles ontologies grâce aux solutions proposées par le Web sémantique passe par le traitement du langage naturel, complété par une ontologie de référence. Dans ce cas, la tâche consiste à repérer dans le texte les instances, existantes ou nouvelles, de cette ontologie. Lorsque de nouvelles instances sont identifiées dans le texte, extraites puis reliées à l'ontologie, il s'agit alors de peupler cette ontologie, c.-à-d. d'enrichir la base de connaissances y étant associée avec ces nouvelles instances.

Le peuplement de notre ontologie est réalisé grâce aux connaissances extraites automatiquement avec la méthode des patrons lexico-syntaxiques (PLS) (Jilani, Grabar, & Jaulent, 2006), issue du domaine du traitement automatique des langues et de la théorie des automates. La méthode des PLS a été utilisée pour la structuration automatique de terminologies : avec la détection de relations hiérarchiques et transversales. Cette méthodologie a été proposée la première fois par Hearst (Hearst, 1992) pour l'acquisition d'hyponymes à partir de textes. Elle a ensuite été développée par Morin dans le cadre de l'acquisition des patrons pour le repérage de relations hiérarchiques entre termes (Morin, 1999).

3 Contexte

Notre travail a pour point de départ un besoin spécifique des biologistes de l'Institut Pasteur (IP) de Montevideo (Uruguay). En effet, travaillant sur les miARN de l'espèce humaine, les biologistes ont besoin de récolter un maximum de connaissances liées à ce sujet. Les miARN sont des ARN simple-brin longs d'environ 21 à 24 nucléotides et sont des répresseurs post-transcriptionnels : en s'appariant à des ARN messagers, ils guident leur dégradation, ou la répression de leur traduction en protéine, entraînant l'apparition ou au contraire l'inhibition de maladies. Or, ce champ d'étude est très récent dans le domaine de la biologie, faisant même l'objet d'un prix Nobel en 2006. Par conséquent, beaucoup de chercheurs biologistes se sont penchés sur cette nouvelle problématique et ont mené un certain nombre de travaux ces dernières années afin de découvrir les interactions possibles entre des combinaisons de couples miARN et ARN messager. Du fait de son caractère novateur et du nombre de travaux récemment publiés, les biologistes possèdent une vision très restreinte sur le sujet. Nous pensons qu'un moyen de les aider à recenser les connaissances existantes, les différentes expérimentations menées, leurs résultats ainsi que les terminologies utilisées dans ce domaine, est de développer une application Web sémantique reposant sur une ontologie de ce domaine. La modélisation d'une telle ontologie permettra de fournir un moyen d'accès central aux différentes terminologies, aux instances de la base de connaissances voire même aux ressources des bases de données externes.

Mais cela ne suffit pas à la tâche des biologistes, il faut aussi leur fournir le moyen de mettre à jour cette base de connaissances sur les miARN. Pour cela, nous avons construit une plateforme basée sur un outil d'extraction d'informations, miR Discovery. L'alimentation de la base de connaissances par les annotations générées à partir des articles biologiques disponibles via le portail PubMed de Medline[1] sera exploitée afin d'appliquer dans un second temps des règles d'inférence qui permettront de raisonner sur ces miARN. Nous allons à présent décrire l'ontologie du domaine des miARN que nous avons construite manuellement avant d'aborder le problème de son alimentation automatique.

[1] www.ncbi.nlm.nih.gov/entrez/

4 Modélisation d'une ontologie des miARN

S'il n'existe pas à l'heure actuelle d'ontologie représentant la connaissance au sujet des miARN, plusieurs ressources terminologiques et ontologiques (Gene ontology (Ashburner et al., 2000), Sequence Ontology (Eilbeck et al., 2005), etc.) offrent un point de vue général sur les gènes et leur séquençage. De plus, des bases de données sur les ARNm et les miARN, telles que Tarbase (Sethupathy, Corda, & Hatzigeorgiou, 2006) et miRBase (Griffiths-Jones, 2004), ont vu le jour récemment, témoignant ainsi de l'engouement des biologistes pour cette nouvelle problématique. Pour ce projet, nous avions besoin d'une ontologie de haut niveau qui constituerait un bon point de départ pour élaborer une nouvelle ontologie dédiée à la représentation des miARN et de leurs impacts sur la régulation et la mutation des gènes. Nous avons donc choisi de travailler avec la Sequence Ontology (SO) car elle a pour objectif de décrire les séquences biologiques en général. Les concepts de gène, d'ARNm et de miARN entre autres étaient déjà représentés dans la SO ainsi que certaines relations pertinentes pour notre domaine comme « *is_part_of* » pour décrire la décomposition d'un miARN en « *loop* », « *stem* », « *3'UTR* », « *5'UTR* », etc. Nous avons également réutilisé une autre relation intéressante « *regulated_by* » pour modéliser le phénomène de régulation entre un miARN et un segment d'ARNm bien qu'elle possède à présent le statut « *deprecated* » dans la SO. Enfin, nous avons enrichi et étendu la modélisation existante de la SO avec la connaissance actuelle des miARN détenue par les biologistes de l'IP de Montevideo.

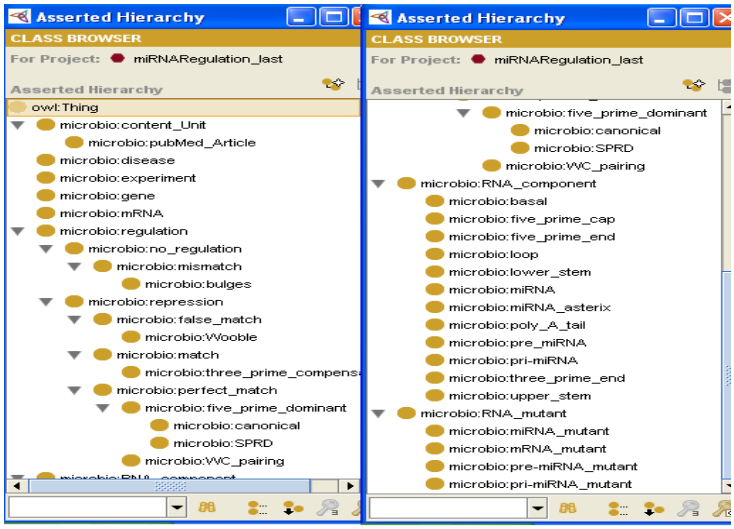


Fig. 1 – Taxonomie de l'ontologie des miARN dans Protégé

Plusieurs versions de la Sequence Ontology existent dans différents formats. Nous avons utilisé le format OBO et l'éditeur OBO Edit[2] pour visualiser cette ontologie et rechercher les concepts pertinents pour notre ontologie. Mais pour la

[2] http://oboedit.org/

modélisation de notre ontologie, nous utilisons le format OWL[3] DL via l'éditeur d'ontologie Protégé[4] largement utilisé par la communauté du Web Sémantique. Ce format OWL nous permet notamment de conserver un lien vers les concepts réutilisés de la SO et de faciliter l'intégration à venir de notre ontologie dans la SO. Lorsqu'un concept provient de la SO ou est équivalent à un concept existant de la SO, nous créons un lien sémantique entre le concept de notre ontologie et celui de la SO via la construction "*owl:equivalentClass*". Par exemple, le concept "miRNA" de la SO est une classe équivalente au concept "miARN" de notre ontologie. Nous appliquons le même principe pour les équivalences entre propriétés via la construction "*owl:equivalentProperty*". Le fait d'opter dans un premier temps pour une modélisation séparée de la SO nous permet de garder une indépendance de conception de notre ontologie tant qu'elle n'aura pas été définitivement validée par les biologistes, tout en conservant un lien fort avec la SO pour le jour où nous leur transmettrons nos résultats pour une demande d'intégration avec leurs ressources.

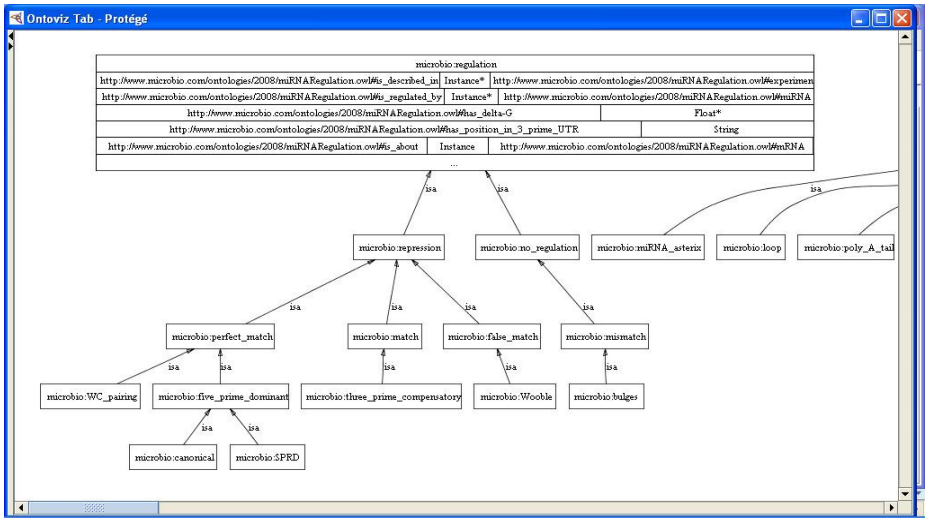


Fig 2 – Concepts, relations et attributs qui ont été ajoutés à la SO afin de représenter les différentes régulations induites par les miARN. La première version de l'ensemble de l'ontologie miARN est actuellement en cours de validation par les biologistes de l'IP.

La figure 1 présente la taxonomie de l'ontologie des miARN. L'ontologie représente aujourd'hui :

- 38 concepts primitifs : ils représentent les objets, abstraits ou concrets, réels ou fictifs, élémentaires ou composites, du monde réel. Ces concepts sont organisés en taxonomie, par l'utilisation de la relation de subsumption, dans laquelle ils peuvent appartenir à plusieurs sur-concepts différents. Par exemple, la classe "3_prime_UTR" est une sous-classe de "RNA_component".

[3] http://www.w3.org/2004/OWL/

[4] http://protege.stanford.edu/

- 13 relations (object properties) : elles représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine. Dans le domaine modélisé, les concepts « Disease » et « RNA_mutant » sont reliés entre eux par la relation sémantique « is_produced_by(Disease, RNA_mutant) » dans laquelle « Disease» est le domaine et « RNA_mutant » la portée (ou « range » en anglais).
- 11 attributs (datatype properties) : les **attributs** correspondent à des caractéristiques, des spécificités particulières attachées à un concept et qui permettent de le définir de manière unique dans le domaine. Leurs valeurs sont littérales, c.-à-d. de type primitif, comme une chaine de caractère ou un nombre entier.

Les **instances** de concepts ne font pas partie à proprement parler de l'ontologie, mais plutôt de la base de connaissances (Handschuh, 2005). En effet, ces dernières permettent de stocker les instances des concepts, mais aussi les instances de relations et les valeurs des propriétés en fonction des contraintes imposées par l'ontologie. C'est l'automatisation de cette tâche à partir d'un corpus d'articles de la littérature biologique que nous développons dans le reste de cet article.

5 Acquisition de connaissances sur les miARN

Afin de peupler cette ontologie du domaine et d'aider à la constitution des ressources pour les patrons d'extraction de notre outil d'annotation, miR Discovery, nous avons exploité un ensemble de connaissances biologiques concernant les miARN.

5.1 Constitution de ressources sur les miARN

Le point de départ de notre travail a d'abord été de mettre en exergue les règles concernant les miARN et les processus biologiques engagés lors de l'appariement des miARN et des ARNm. C'est un champ d'activité nouveau et prometteur sur lequel les biologistes veulent travailler et obtenir des résultats très rapidement pour être concurrentiels.

La table 1 présente un aperçu des règles surlignées par l'expert biologiste, et qui d'après lui contiennent l'information pertinente à extraire. Un type de règle en particulier représente un intérêt fondamental pour les biologistes, il s'agit des passages de textes qui traitent d'un lien éventuel entre une mutation sur un gène ou un miARN et une maladie. Nous retrouvons cette connaissance dans les règles 1 à 5 de la table 1. Notons que la connaissance exprimant un lien entre une mutation et/ou un gène et/ou un miARN sans allusion à aucune maladie est aussi une information potentiellement intéressante pour les biologistes. En effet, le biologiste peut être alerté sur ce lien existant et en vérifier si nécessaire la nature et les conséquences induites. Afin de pouvoir détecter dans les textes les termes relatifs aux mutations, aux gènes, aux miARN ainsi qu'aux maladies, il a été nécessaire d'identifier les terminologies respectives et de les charger dans l'outil d'extraction d'information.

Table 1. Liste des règles surlignées par l'expert biologiste, qui identifient clairement une connaissance d'intérêt pour les biologistes sur les miARN. On y retrouve également le PMID (identifiant unique des articles dans Pubmed) ainsi que les termes importants de la règle qui ont permis d'attirer l'attention du biologiste sur l'intérêt de la règle en question.

N°	Règle	PMID	Termes importants
1	Two novel mutations (C+7T/miR-16-1, G+19A/let-7e) have been shown to differentially modulate miRNA expression in vivo;	18778868	Mutations (C+7T/miR-16-1, G+19A/let-7e) / miRNA expression
2	however, only one (C+7T/miR-16-1) has been reported to be associated (P = 0.038) with a human disease: chronic lymphocytic leukemia (CLL)	18778868	(C+7T/miR-16-1) / chronic lymphocytic leukemia (CLL)
3	Among these, they identified one SNP in the 3' UTR of the cluster of differentiation 86 (CD86) gene, rs17281995 [C/G], that was significantly associated with colorectal cancer	18778868	SNP / CD86 / rs17281995 / colorectal cancer
4	Calin et al. also described at least two more novel, potentially CLL-associated miRNA mutations (G49T/miR-206 and +107A/miR-29b-2) that merit further experimental analysis.	18778868	CLL / associated miRNA mutations / (G49T/miR-206 and +107A/miR-29b-2)
5	The authors observed that the strongest association (P = 0.0001) was with rs12720208 [C/T], a SNP that the authors demonstrate mediates allele-specific in vitro targeting of miR-433 to the FGF20 3' UTR.	18778868	rs12720208 / targeting of miR-433 to the FGF20 3' UTR
6	As the target sites were designed to allow optimal 3' pairing, we conclude that G:U base-pairs in the seed region are always detrimental	15723116	G:U base-pairs / seed region are always detrimental
7	Surprisingly, a single 8mer seed (miRNA positions 1–8) was sufficient to confer strong regulation by the miRNA	15723116	single 8mer seed / miRNA positions 1–8 / strong regulation

Table 2. Terminologies ou bases de données utilisées pour construire les dictionnaires relatifs aux gènes, aux miARN, aux maladies et aux mutations.

Termes	**Source**	**Exemple**	**Nombre de termes récupérés**
Gènes	HUGO[7] SwissProt Tarbase	CD86	119 408
miARN	miRBase	let-7 mir-318	904
Maladies	Sous ensemble de la SNOMED	chronic lymphocytic leukemia (CLL)	4 443
Mutations	Construction manuelle par des graphes (Cf. figure 2)	rs17281995 G49T/miR-206	Infini

Le corpus utilisé a été collecté manuellement en envoyant la requête suivante à PubMed : *SNPs [MH] AND miRNAs [MH] AND human [MH].*

[MH] indique que le terme à gauche est un terme MeSH[8]: le Medical Subject Headings est un thésaurus biomédical proposant 25 186 termes en 2009 pouvant être utilisés pour décrire très précisément le contenu d'un document médical.

[7] http://www.hugo-international.org/

[8] http://www.nlm.nih.gov/mesh/

Le résultat de cette requête correspondait à 35 articles[9], parmi lesquels uniquement 21 étaient gratuitement disponibles en entier.

Le corpus exploité comporte donc 21 articles provenant de journaux différents, équivalents à 533 853 tokens, et d'une taille de 2,2 Mo.

5.2 Extraction de la connaissance relative aux miARN

Nous avons réuni toutes les phrases dans lesquelles nous retrouvons une proposition décrivant un lien entre une mutation, un gène, un miARN et/ou une maladie parmi les règles concernant les miARN surlignées par l'expert biologiste. Il est alors apparu qu'aucune phrase n'a la même structure syntaxique que les autres. En effet, il existe de très nombreuses manières d'exprimer ce lien biologique. Nous avons également noté que les phrases dans lesquelles nous retrouvons une mutation, un gène, un miARN et/ou une maladie sont forcément des phrases qui relatent un lien biologique entre ces différents éléments. Partant de cette hypothèse, la construction de patrons lexico-syntaxiques avec différents chemins syntaxiques n'est pas nécessaire, c.-à-d. que des patrons uniquement lexicaux de tri-occurrences (mutation et (gène ou miARN) et maladie) ou de quadri-occurrences (mutation et gène et miARN et maladie) sont suffisants pour extraire les connaissances qui intéressent les biologistes.

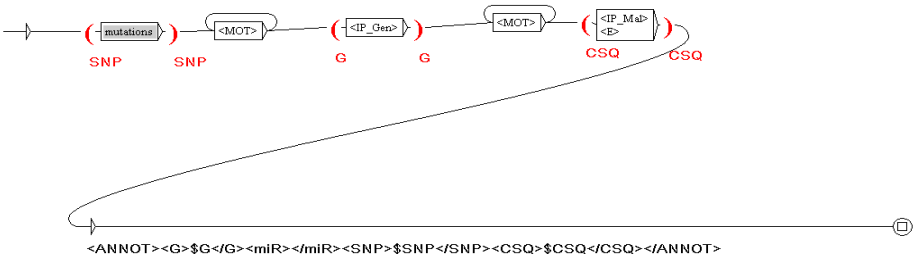


Fig. 3 – Patron de tri-occurrence reconnaissant une mutation suivi d'une succession de motS (<MOT>*), puis un nom de gène (<IP_Gen>) suivi d'une succession de mots (<MOT>*) et enfin une maladie (<IP_Mal>) ou rien (<E>). Si le patron reconnaît une phrase dans le texte, les termes déterminants pour l'extraction sont récupérés grâce aux variables définies : SNP pour les mutations, G pour le gène et CSQ pour la maladie. Le résultat de l'extraction sera donc :<ANNOT><G>$G</G><miR></miR><SNP>$SNP</SNP><CSQ>$CSQ</CSQ></ANNOT>

Les patrons sont donc construits pour détecter une tri-occurrence ou quadri-occurrence dans le texte.

La figure 3 montre un graphe représentant un patron de tri-occurrence pour détecter les phrases exprimant un lien entre une mutation, un gène et/ou une maladie. Il fait appel à un sous-graphe nommé mutations (boîte grisée) présenté à la figure 4. Il a été

[9] Accès à Pubmed le 25 novembre 2008

nécessaire de construire un patron qui reconnaisse une infinité de mutations différentes (Figure 4), car il n'existe pas de base de données répertoriant de manière exhaustive toutes les mutations existantes chez l'humain. Ce patron passe néanmoins à côté de la détection de plusieurs mutations car il existe dans la littérature de très nombreuses façons de les exprimer, variant presque d'un auteur à un autre, et ce, malgré une nomenclature bien définie[10]. La figure 4 illustre quelques façons d'exprimer les mutations.

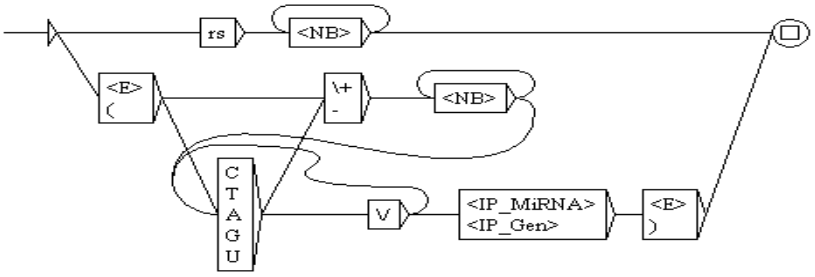


Fig. 4 – Patron de reconnaissance des mutations: il reconnaît des mutations de type : *rs* suivi d'un nombre (<NB>*), mais aussi de type *(G+19A/let-7e)* où *let-7e* est un miARN (<IP_MiRNA>), ou encore *+107C/miR-29b-6.*

5.3 Peuplement automatisé de l'ontologie des miARN

L'Unstructured Information Management Architecture (UIMA)[11] est une infrastructure logicielle initiée par le centre de recherche Alphaworks d'IBM et à présent repris par l'incubateur d'Apache. Elle fournit les bases pour développer un processus d'annotation sémantique, bien qu'elle ne donne pas pour autant des conseils sur la programmation et l'ordre des étapes de ce processus . Le Content Augmentation (CA) Manager développé dans le cadre du projet européen TAO[12] propose sur la base d'UIMA une liste d'étapes logiques, dont certaines sont optionnelles, qui vont être chaînées ensemble, enrichissant au fur et à mesure un schéma d'annotation pré-défini (Figure 5). Ces étapes peuvent être groupées en trois thématiques ou composants principaux : 1) *Extraire* la connaissance pertinente et annoter le contenu; 2) *Consolider* les résultats vis à vis du modèle de l'ontologie et du référentiel sémantique; 3) Sérialiser le schéma d'annotation dans divers formats et le *stocker* dans le référentiel sémantique.

Avant même d'initier la démarche d'annotation sémantique et de peuplement d'ontologie, il nous faut charger l'ontologie des miARN dans un référentiel sémantique comme Sesame[13] ou ITM[14]. Puis, la première étape d'extraction consiste à appeler l'outil

[10] http://www.genomic.unimelb.edu.au/mdi/mutnomen/recs.html
[11] http://www.research.ibm.com/UIMA/
[12] http://www.tao-project.eu
[13] http://www.openrdf.org/
[14] http://mondeca.com/index.php/en/intelligent_topic_manager

d'analyse linguistique décrit ci-dessus, miR Discovery, afin d'annoter automatiquement les articles biologiques de PubMed. Chaque nouvelle annotation ou instance générée (notamment de relations entre les miARN, mutations, gènes et maladies) est contrôlée dans le référentiel sémantique afin de vérifier la non-redondance de la connaissance et la préservation de la cohérence et de la qualité de la base de connaissances. Pour ce faire, un ensemble de règles de consolidations (Amardeilh, 2008) sont définies et appliquées sur chaque nouvelle annotation, instance de classe ou instance de propriété. Ces règles peuvent être simples comme vérifier qu'une instance de même classe et même libellé n'existe pas déjà dans la base de connaissances ou bien beaucoup plus complexes. Dans ce cas, les règles de consolidation peuvent être complétées par des règles de raisonnement et d'inférence. Les informations non valides vis à vis du modèle ontologique ou de la base de connaissances seront mises de côté avec le statut « à valider » afin que si l'application finale repose sur une interface de validation manuelle de ces annotations et instances, le CA Manager soit en mesure de remonter ces informations non valides pour permettre à l'utilisateur de les corriger et les désambiguïser si besoin.

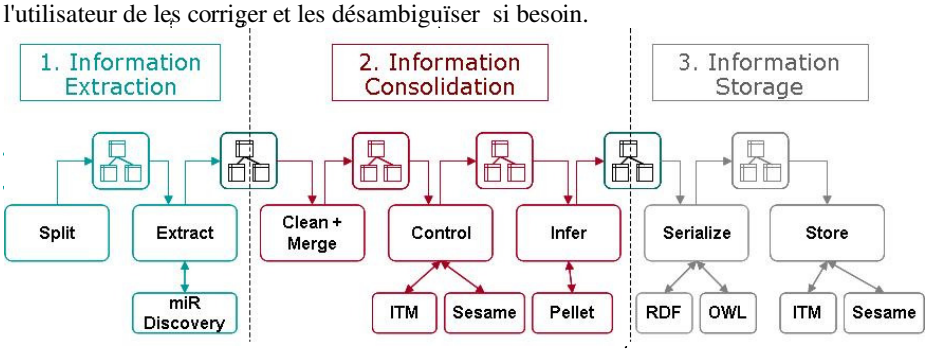


Fig. 5 - Un workflow modulaire basé sur UIMA

5.4 Evaluation et Résultats

Nous avons pu extraire des connaissances concernant les mutations ainsi que des indications de leur emplacement (miARN ou gène) mais aussi parfois leur lien éventuel avec une maladie. Ces connaissances sont extraites et proposées au format XML, comme l'indique la figure 6. Nous avons extrait 35 annotations différentes, c.-à-d. distinctes les unes par rapport aux autres (sur le modèle de la figure 6). Sur la totalité du corpus, 30 annotations provenant de 49 phrases différentes ont été réalisées manuellement par l'expert biologiste. Elles sont utilisées comme référence pour notre évaluation. Sur ces 49 phrases, certaines proposent donc des annotations redondantes.

Notre outil a permis de faire 35 annotations. Sur ces 35 annotations, 25 sont correctes mais certaines sont incomplètes (on ne retrouve pas la maladie par exemple) ou redondantes et, parmi elles, 15 sont strictement identiques aux annotations de référence faites par l'expert biologiste. Nos mesures de rappel et de précision sont données ci-après :

- Précision = 25 / 35 = 0,72
- Rappel = 15 / 30 = 0,50

Le chiffre du rappel est relativement bas car nous n'avons pas pris en compte les variantes morphologiques des maladies par exemple. En effet, notre outil ne détecte pas une phrase qui contient *« lung cancers »* car notre dictionnaire n'inclut que les formes au singulier *« lung cancer »*. Le chiffre de la précision souffre de la synonymie des noms de gènes avec des noms, des prénoms, ou des acronymes utilisés pour référencer des techniques en biologie. Ce problème n'est pas nouveau, et de nombreuses équipes y travaillent.

```
<ANNOT>
  <G>FGF20</G>
  <miR>miR-433</miR>
  <SNP>rs12720208</SNP>
  <CSQ />
  <PMID>18252210</PMID>
  <SENT>miR-433 Inhibits FGF20 Translation at SNP rs12720208 .</SENT>
</ANNOT>
```

Fig. 6 – Exemple de connaissances extraites au format XML, les balises <G> sont pour les gènes, <miR> pour les miARN, <SNP> pour les mutations, <CSQ> pour les maladies et <SENT> pour les phrases dont proviennent les annotations.

6 Conclusion

Nous venons de présenter dans cet article la notion de filtrage sémantique à travers les activités d'annotations sémantiques et de peuplement d'ontologie traitant des miARN. Nous avons vu que ce filtrage est intrinsèquement lié à la modélisation d'une ontologie de domaine dans le cadre du Web Sémantique. En effet, cette ontologie va représenter les concepts, attributs et relations d'un domaine à l'aide d'un langage de représentation des connaissances orienté Web comme OWL. Elle sera instanciée à partir des extractions linguistiques, par exemple les interactions entre des miARN et ARNm identifiés dans les articles de PubMed. L'ontologie que nous avons modélisée sert de médiateur et d'accès aux différentes terminologies de l'application. Celle-ci repose sur une architecture SOA (Service Oriented Architecture) s'appuyant sur les web services proposés par le CA Manager. Plus ce modèle de représentation de la connaissance dans lequel seront exprimées les instances est formel, plus les services proposés seront « intelligents ». Les biologistes pourront interroger les connaissances en fonction de différents points de vue : recherche par mots-clefs (annotations sémantiques), recherche multicritère (en fonction des concepts, relations et attribut du modèle), ou encore recherche par extension sémantique (terminologies). Les agents logiciels pourront également raisonner et inférer de la nouvelle connaissance et ainsi dégager un sens implicite contenu dans le document d'origine (Laublet, 2007). Le peuplement d'ontologie à l'aide des techniques du Web sémantique ouvre donc des perspectives intéressantes à de nombreuses applications comme la recherche d'informations sémantiques, la catégorisation, la composition de documents, etc. L'extraction de connaissances donne un rappel de 50%, et la précision des résultats est de 72%. Ces deux mesures peuvent être améliorées et nous y travaillons.

Remerciements

Cette étude a été financée par le projet Microbio (programme Stic-Amsud).

Références

Amardeilh, F. (2008). Semantic Annotation & Ontology Population. *Semantic Web Engineering in the Knowledge Society, ISI Global*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29.

Bourigault, T. D., Aussenac-gilles, N., & 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,* 87-110.

Brisson, L., & Collard, M. (2007). *Intérêt des systèmes d'information dirigés par des ontologies pour la fouille de données*. Rapport de recherche Projet Execo, Nice, Sophia Antipolis: CNRS.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., et al. (2005). The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology, 6*(5).

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research, 32*(Database Issue), D109-D111.

Hearst, M. A. (1992). *Automatic acquisition of hyponyms from large text corpora.* Paper presented at the International Conference on Computational Linguistics (COLING'92), Nantes.

Hignette, G. (2007). *Annotation sémantique floue de tableaux guidée par une ontologie, Thèse de doctorat,* AgroParistech.

Jilani, I., Grabar, N., & Jaulent, M.-C. (2006). *Fitting the finite-state automata platform for mining gene functions from biological scientific literature.* Paper presented at the Semantic Mining in Biomedicine, Jena (Germany).

Laublet, P. (2007). Web sémantique et ontologies. In Hermès (Ed.), *Humanités numériques Nouvelles technologies cognitives et concepts des sciences sociales* (Vol. 1). Paris.

Morin, E. (1999). Acquisition de patrons lexicosyntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL), 40(1)*, 143-166.

Prié, Y., & Garlatti, S. (2004). Méta-données et annotations dans le Web sémantique. *Revue I3 Information - Interaction - Intelligence, 4*, 45-68.

Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA, 12*, 192-197.

Explorer des actualités multimédia dans le web de données

Raphaël Troncy[1]

CWI Amsterdam, Science Park 123, 1098 XG Amsterdam, The Netherlands
raphael.troncy@cwi.nl

Résumé : Pour faciliter l'échange des actualités, l'IPTC (*International Press Telecommunication Council*) a développé l'Architecture NewsML (NAR) composée d'un modèle XML pour représenter les métadonnées et de vocabulaires contrôlés (IPTC News Codes) pour catégoriser les dépêches de presse. D'autres formats de métadonnées spécifiques au multimédia peuvent être utilisés conjointement mais cela pose des problèmes d'interopérabilité puisque les modèles XML fermés sous-jacents empêchent en particulier de lier ces métadonnées à d'autres connaissances disponibles sur le web. Dans cet article, nous proposons un environnement unique pour chercher et naviguer dans des contenus multimédia d'actualités contextualisés. Nous présentons une ontologie OWL pour l'Architecture NewsML, liée à d'autres ontologies multimédia. Nous montrons comment les métadonnées fournies par les journalistes peuvent être automatiquement enrichies par des méthodes de traitement automatique de la langue et du signal multimédia, pour ensuite être liées à des connaissances formalisées dans le web de données. Nous fournissons des recommandations quant à développer une ontologie à partir d'un schéma et d'en formaliser les connaissances implicites.

Mots-clés : Actualités multimédia, Web de données, Interface d'exploration.

1 Introduction

Dans le cycle de vie d'une actualité, l'information est généralement : *i)* produite par une agence de presse, un journaliste indépendant ou citoyen, *ii)* consommée et enrichie par un quotidien, un magazine ou un diffuseur radio-télévisé et finalement *iii)* livrée à des utilisateurs finaux. Les dépêches de presse sont typiquement accompagnées de métadonnées et d'une description brève du contenu pour faciliter leur indexation et leur recherche dans des archives. Cependant, la plupart de ces métadonnées se perd à cause de problèmes d'interopérabilité entre les différents acteurs des processus métier de production des actualités. De plus, les interfaces utilisateur de consultation utilisent rarement ces métadonnées. Par conséquent, les utilisateurs sont souvent obligés d'utiliser des environnements qui, pour une recherche donnée, contiennent une grande quantité d'information non pertinente, souvent redondante ou peu fiable, avec un accès insuffisant à de la connaissance de contexte pour comprendre ces actualités.

Notre objectif à long terme est de créer un environnement qui permettrait aux utilisateurs de voir les relations causales, logiques et sémantiques entre des actualités mul-

timédia prises individuellement, en utilisant leurs descriptions formalisées et la connaissance disponible sur le web. Notre approche consiste à créer des modèles de connaissance pour améliorer les problèmes d'interopérabilité dans toute la chaîne de production des actualités. Le problème que l'on cherche à résoudre dans cet article couvre les deux extrêmes de cette chaîne de production : comment représenter le sens des métadonnées tout au long du flux d'information et quelle conséquence cette modélisation a sur l'interface utilisateur finale.

Notre contribution est double. Nous présentons tout d'abord la modélisation d'ontologies OWL pour les langages de descriptions standardisés par l'IPTC, nous convertissons les vocabulaires contrôlés dans un thésaurus SKOS et nous montrons comment les métadonnées peuvent être automatiquement enrichies et intégrées à de la connaissance formalisée disponible sur le web. Nous généralisons cette approche et nous fournissons quelques recommandations quant à modéliser une ontologie formelle à partir d'un schéma. Nous discutons ensuite les décisions de modélisation en essayant d'évaluer leurs conséquences sur les interfaces utilisateur. Nous présentons finalement un prototype qui permet de chercher et de naviguer dans des actualités à partir de leurs descriptions formelles (Troncy, 2008).

L'article est structuré de la manière suivante. Nous introduisons brièvement dans la section suivante les standards principaux utilisés par l'industrie des médias. Nous discutons dans la section 3 des méthodes existantes pour construire des ontologies à partir de schémas et nous présentons les différentes tentatives pour intégrer ontologie pour représenter les actualités et ontologie multimédia. Nous détaillons dans la section 4 les étapes que nous préconisons pour construire une infrastructure sémantique pour les actualités. Pour démontrer l'adéquation de cette infrastructure, nous présentons un système de recherche sémantique pour des actualités multimédia (section 5) avant de conclure et d'ouvrir quelques perspectives à ces travaux (section 6).

2 Les standards pour l'actualité et le multimédia

2.1 Les standards pour l'actualité

Historiquement, l'IPTC a développé les formats NITF[1] et NewsML pour décrire la transmission, la structure et le contenu des informations d'actualités. Ces langages XML ont, cependant, démontré leurs limites pour décrire des actualités de plus en plus multimédia, et ont souvent été jugés trop verbeux. L'IPTC a donc récemment produit l'architecture NAR[2] qui fournit un cadre général pour une seconde génération de spécifications (G2). NAR est un modèle générique qui définit quatre objets principaux (`newsItem`, `packageItem`, `conceptItem` and `knowledgeItem`) ainsi que les opérations et traitements associés à leurs structures. Des langages spécifiques pour décrire des actualités (NewsML G2) ou des événements (EventsML G2) étendent ensuite cette architecture. Ainsi, l'élément générique `newsItem` est spécialisé pour prendre en considération les différents médias (dépêche textuelle, image, clip vidéo).

[1] News Industry Text Format : `http://www.nitf.org/`

[2] `http://www.iptc.org/NAR/`

IPTC maintient finalement un ensemble de vocabulaires contrôlés appelés *IPTC NewsCodes* qui sont utilisés pour catégoriser les dépêches d'actualités. Le thésaurus *Subject Code* contient par exemple 1300 termes organisés sur trois niveaux hiérarchiques pour décrire le sujet principal de chaque dépêche.

2.2 Les standards pour le multimédia

Bien que NAR défini des concepts pour représenter différents médias (textuel, photo, audio, vidéo, graphique, animation), une multitude d'autres standards sont utilisés par l'industrie des médias (Hausenblas *et al.*, 2007). Ainsi, les photos prises par les journalistes contiennent des métadonnées EXIF fournies par l'appareil spécifiant les caractéristiques de la photo (e.g. taille, orientation) ou des informations liées à sa prise (e.g. focale, temps d'exposition, flash). Kanzaki[3] et Norm Walsh[4] ont tous deux proposés une ontologie RDFS de EXIF et fournissent un service pour extraire et convertir ces métadonnées contenues dans l'en-tête des images.

Ces métadonnées techniques sont généralement complétées avec d'autres standards dont le but est de décrire le contenu. DIG35 est par exemple une spécification de l'I3A (*International Imaging Association*) qui définit un schéma XML pour représenter les paramètres de l'image, les informations de création, ce que l'image représente (qui, quoi, quand et où), ou encore les droits associés à l'image. En collaboration avec l'université de Ghent, nous avons récemment proposé une ontologie pour ce format[5] dont la modélisation suit les mêmes principes que nous exposerons dans la section 4. XMP fournit un modèle RDF natif pour décrire la gestion, les droits et le contenu d'images en ré-utilisant le format du Dublin Core. IPTC a lui-même intégré XMP dans son propre format de métadonnées pour les images.

Une vidéo peut être décomposée et décrite en utilisant le standard MPEG-7 (MPEG-7, 2001). Ce langage fournit un ensemble important de descripteurs pour décomposer un média, gérer les métadonnées de catalogage, représenter les caractéristiques de bas niveau du signal ou encore définir des concepts plus abstraits. L'ambiguïté et le manque de sémantique formelle de MPEG-7 ont déjà été mises en avant, et plusieurs ontologies OWL ont été proposées et récemment comparées (Troncy *et al.*, 2007). L'ontologie COMM (*Core Ontology for Multimedia Annotation*) propose par exemple une nouvelle conceptualisation du standard en utilisant DOLCE comme ontologie de haut niveau et en créant de nouveaux patrons de conception (*design patterns*) pour le multimédia (Arndt *et al.*, 2007). Chez les diffuseurs, l'EBU[6] a récemment adopté l'architecture NAR pour décrire les vidéos en fournissant quelques extensions pour décrire plus finement le contenu et gérer les droits associés.

En conclusion, la chaîne de production d'actualités utilise de nombreux standards et formats, souvent basés sur XML, mais intrinsèquement fermés, ce qui conduit à des problèmes d'interopérabilité. De plus, ces formats empêchent d'utiliser de nouveaux vocabulaires contrôlés qui n'auraient pas été prévus au préalable, ou plus généralement

[3] `http://www.kanzaki.com/ns/exif`

[4] `http://sourceforge.net/projects/jpegrdf`

[5] `http://multimedialab.elis.ugent.be/users/chpoppe/Ontologies/`

[6] European Broadcaster Union : `http://www.ebu.ch`

de la connaissance disponible sur le web. Nous proposons d'utiliser les langages du web sémantique pour faciliter l'intégration de ces standards. En se basant sur l'état de l'art de la construction d'ontologies décrit dans la prochaine section, nous présentons notre infrastructure sémantique et nous formulons des recommandations pour modéliser une ontologie formelle à partir de schémas existants.

3 Etat de l'art

3.1 Convertir des schémas et des thésauri pour le web sémantique

Les productions en matière de construction d'ontologies sont nombreuses. Uschold et Grüninger s'intéressent par exemple à l'ensemble du processus de conception de l'ontologie et de son cycle de vie, tandis que METHONTOLOGY propose de modéliser l'ontologie au niveau des connaissances en utilisant des représentations intermédiaires. D'autres travaux insistent sur la conceptualisation des taxonomies de concepts et de relations (Troncy & Isaac, 2002), tandis que la méthodologie ARCHONTE se veut constructiviste et préconise un retour au corpus textuel pour sélectionner des concepts et des relations dont leur sens est normalisé avant d'être formalisé (Bachimont, 2004). Toutes ces méthodologies, cependant, ne considèrent pas le cas supposé plus simple où un schéma semi-formel (diagramme UML, schéma XML, thésaurus) modélisant les connaissances du domaine, existe mais doit être formalisé pour être utilisable sur le web sémantique.

Une méthode générale pour convertir des thésauri en SKOS a également été proposée (Assem *et al.*, 2004). Cette méthode préconise quatre étapes : préparation, conversation syntaxique, conversion sémantique et standardisation. Notre méthode s'inspire de ces recommandations mais en ajoute de nouvelles pour conceptualiser l'ontologie à partir de schémas semi-formels. L'alignement entre des thésauri convertis en SKOS et des ressources du web sémantique a été discuté (Tordai *et al.*, 2007). Nous utilisons l'outil AnnoCultor[7] issu de ces travaux pour convertir les vocabulaires contrôlés de l'IPTC en thésauri SKOS.

3.2 NewsML et les ontologies multimédia

Plusieurs travaux ont tenté de construire une ontologie des actualités. Le projet européen NEWS[8] propose une ontologie RDFS multilingue à partir des formats NITF et NewsML et du thésaurus *SubjectCodes* (Fernández *et al.*, 2006). Le projet Neptuno[9] suit la même approche et propose une ontologie RDFS construite également à partir de NewsML et du thésaurus *SubjectCodes* aligné avec d'autres vocabulaires contrôlés internes à l'agence de presse espagnole (Castells *et al.*, 2004) dans le but de décrire son archive. MESH[10] est finalement un projet européen en cours qui met l'accent sur l'analyse multimédia pour enrichir automatiquement les métadonnées et construire des

[7] `http://sourceforge.net/projects/annocultor`

[8] `http://www.news-project.com/`

[9] `http://seweb.ii.uam.es/neptuno/`

[10] `http://www.mesh-ip.eu/`

résumés personnalisés. Une ontologie des actualités semble avoir été développée mais celle-ci n'est pas encore disponible.

A la différence de ces projets, notre approche consiste à dissocier les thésauri utilisés pour valuer les propriétés des métadonnées, de l'ontologie qui décrit la gestion et le contenu des dépêches selon le point de vue du journaliste. Cette séparation fournit une architecture plus flexible où, par exemple, le thésaurus *SubjectCodes* peut être indépendamment aligné avec d'autres vocabulaires contrôlés. Nous publions également ces thésauri sur le web sémantique, en fournissant pour chaque terme un URI déréférençable[11]. Finalement, notre ontologie est conforme aux standards les plus récents de l'IPTC (NAR) et nous la lions à d'autres ontologies suivant les bons principes du web (sémantique).

La méthodologie *XML Semantics Reuse* consiste à transformer automatiquement un schéma XML en une ontologie OWL[12]. Elle a été utilisée dans le domaine du journalisme pour convertir les formats NITF, NewsML et MPEG-7 ainsi que le thésaurus *SubjectCodes* de l'IPTC en OWL/RDF (Garcia *et al.*, 2008). L'ontologie résultante, cependant, ne formalise pas la sémantique informelle de ces standards puisque celle-ci n'est pas présente dans les schémas XML originaux (Troncy *et al.*, 2007). De plus, la transformation automatique re-créée des structures complexes imbriquées (e.g. des éléments intermédiaires correspondants à des types ou des ensembles d'éléments) qui n'ont pas lieu d'être dans une ontologie. Nous préconisons au contraire de re-modéliser l'ontologie en suivant les bonnes pratiques détaillées ci-dessous.

4 Une infrastructure sémantique pour les actualités

NAR est un modèle générique pour décrire le contenu des actualités, et la manière de les gérer et de les échanger. Ce modèle partage assez naturellement les mêmes principes que le web sémantique :

– les actualités sont des ressources distribuées qui ont besoin d'être identifiées de manière unique et pérenne ;
– les actualités sont décrites avec des vocabulaires contrôlés et partagés.

NAR est cependant défini à l'aide d'un schéma XML et par conséquent, sa sémantique implicite n'est pas formellement représentée (par exemple, un `NewsItem` peut être un `TextNewsItem`, un `PhotoNewsItem` ou un `VideoNewsItem`). L'extension du modèle à d'autres standards est également compliquée puisque il est difficile d'établir l'équivalence de deux éléments XML. Nous décrivons dans la suite les étapes nécessaires pour modéliser notre infrastructure ontologique[13].

4.1 Étape 1 : modéliser l'ontologie NAR

La première étape consiste à formaliser la sémantique implicite des standards de l'IPTC. Bien que ces modèles existent sous forme de diagrammes UML, leur "ontologi-

[11] Le déréférencement consiste à accéder à la représentation d'une ressource identifiée par une URI. L'expression *deferencing an URI* a fait l'objet d'une longue discussion au sein du W3C pour finalement aboutir à la résolution communément appelée httpRange-14.

[12] Voir le projet ReDeFer : `http://rhizomik.net/redefer`

[13] Les différentes ontologies sont disponibles à `http://newsml.cwi.nl/ontology/`

sation" n'est pas triviale. Nous discutons ci-dessous quelques décisions de modélisation.

Aplatir la structure XML. XML Schema permet de définir des structures relativement riches, mais il est limité quant à définir leur sémantique puisque le langage fournit principalement un système de typage pour des données structurées. Ainsi, le modèle NAR contient un certain de nombre de structures intermédiaires, sorte de container dont le but est juste de regrouper des éléments sans avoir un sens particulier. Ces structures ne doivent pas être représentées dans l'ontologie puisqu'elles ne seront pas instanciées et correspondront à des noeuds anonymes dans le graphe RDF. Pour conceptualiser l'ontologie, nous préconisons donc d'aplatir le schéma XML en ne gardant que les propriétés qui seront instanciées.

Indiquer la provenance. Les assertions contenues dans les dépêches ont souvent besoin d'être réifiées. Par exemple, un éditeur enregistré comme `team:md` peut indiquer qu'une dépêche a été catégorisée `diplomatie` le `2005-11-11T08:00:00Z`, ce qui se traduit en RDF par :

```
{<> nar:subject cat:11002000} dc:creator   team:md ;
                  dc:modified  ''2005-11-11T08:00:00Z'' .
```

Le mécanisme de réification de RDF n'a cependant pas de sémantique formelle dans la théorie des modèles. Pour représenter la provenance des informations, nous préconisons donc d'utiliser la technique des graphes liés et nommés où les relations entre graphes sont décrites à l'aide de requêtes SPARQL et de vues (Schenk & Staab, 2008).

4.2 Étape 2 : la lier avec d'autres ontologies

Comme nous l'avons vu dans la section 2.2, d'autres standards tels que EXIF, Dublin Core, XMP, DIG35 ou MPEG-7 sont utilisés par l'industrie des médias. Ces standards ont généralement été traduits, ou existent nativement en OWL/RDF et peuvent donc s'intégrer naturellement à notre architecture. Par conséquent, nous préconisons d'ajouter des axiomes dans l'ontologie pour expliciter les relations entre concepts provenant d'ontologies différentes mais qui se recouvrent partiellement. Ainsi, l'ontologie NAR contient les axiomes suivants :

```
nar:subject  owl:equivalentProperty  dc:subject
nar:Person   owl:equivalentClass     foaf:Person
```

Les moteurs de recherche du web sémantique tels que Sindice[14], Watson[15] ou Falcon[16] peuvent être utilisés pour découvrir de nouveaux concepts ou propriétés partageant le même sens que ceux définis dans notre ontologie et auxquels ils pourraient être liés.

4.3 Étape 3 : convertir les IPTC News Codes en thésauri SKOS

Les IPTC *NewsCodes* sont définis dans 36 thésauri, de tailles variables, dont les termes sont utilisés dans les métadonnées décrivant les actualités. Bien que ces termes

[14] `http://sindice.com/`

[15] `http://watson.kmi.open.ac.uk/WatsonWUI/`

[16] `http://www.falcons.com.cn/`

soient parfois organisés en taxonomie, la relation de subsomption n'est jamais explicite mais encodée dans les noms de termes. Ainsi, "cancer" (`cat:07001004`) est plus spécifique que "maladie" (`cat:07001000`) qui est lui même plus spécifique que "santé" (`cat:07000000`) simplement parce qu'ils partagent les mêmes quatre premiers chiffres. Nous avons converti ces thésauri en SKOS en explicitant cette relation de subsomption à l'aide des constructeurs `skos:narrower` et `skos:broader`.

Cette compatibilité RDF nous permet de définir plus en avant certains concepts de l'ontologie NAR en terme de `owl:Restriction` : la valeur d'une propriété peut être un `skos:Concept` ou doit provenir d'un `skos:ConceptScheme` particulier. Par exemple, la propriété `nar:subject` ne peut avoir comme valeur qu'un terme provenant du `skos:ConceptScheme` *SubjectCodes*.

Finalement, nous avons exposé[17] tous ces thésauri sur le web sémantique en suivant le guide des bonnes pratiques promues par le W3C[18]. Chaque terme est donc identifié par un URI déréférençable. Ainsi, une requête http dont le type est `Accept:text/html` retournera la description XML originale de l'IPTC tandis que le type `Accept:application/rdf+xml` retournera la version SKOS/RDF utilisable par les machines du thésaurus.

4.4 Étape 4 : enrichir automatiquement les métadonnées

Une fois l'ontologie NAR modélisée et liée à d'autres ontologies populaires sur le web sémantique, la conversion des métadonnées de chaque dépêche en RDF selon notre architecture est triviale. Cependant, nous préconisons une ultime étape dont le but est d'enrichir automatiquement les métadonnées en suivant les principes des données liées[19]. Ainsi, nous utilisons des techniques de traitement automatique de la langue et du signal multimédia pour extraire d'avantages de métadonnées (Figure 1).

Le traitement automatique de la langue consiste à extraire les entités nommées telles que les personnes, les organisations, les lieux, les marques, etc. à partir de la légende d'une photo ou d'une dépêche textuelle. Nous utilisons désormais le service OpenCalais[20] après avoir expérimenté avec les plate-formes GATE[21] et SPROUT[22]. Une fois les entités nommées extraites, nous les alignons avec des ressources disponibles dans le web des données, à savoir, avec Geonames pour les lieux, ou avec DBPedia pour les noms de personnes et d'organisations. Le traitement du signal fournit également un autre type de métadonnées utilisé ultérieurement pour classifier les résultats d'une requête. Ainsi, il est possible d'effectuer une classification non supervisée de toutes les images montrant le footballeur Zinedine Zidane en utilisant les descripteurs de texture et d'histogramme de couleur, et ainsi de différencier les photos où il apparaît en costume pour recevoir un prix et où il est sur le terrain.

[17] `http://newsml.cwi.nl/NewsCodes/`

[18] `http://www.w3.org/TR/swbp-vocab-pub/`

[19] Linked Data : `http://linkeddata.org/`

[20] `http://www.opencalais.com/`

[21] `http://gate.ac.uk/`

[22] `http://sprout.dfki.de/`

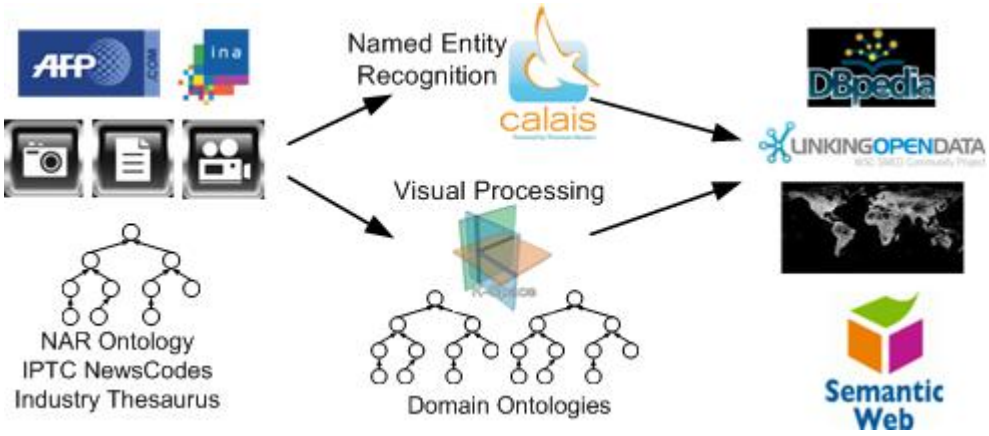


FIG. 1 – Enrichissement automatique des métadonnées des actualités

5 Explorer des actualités multimédia

Dans le but de démontrer l'utilité de notre architecture sémantique, nous présentons dans cette section un prototype pour chercher et explorer des actualités multimédia. Nous utilisons Cliopatria[23], une plate-forme basée sur SWI-Prolog comprenant un entrepôt pour des données RDF, l'implémentation des langages de requêtes SeRQL/SPARQL, les bibliothèques d'interface utilisateur de Yahoo ! (YUI) et des routines pour la recherche sémantique (Hildebrand *et al.*, 2006; Wielemaker *et al.*, 2008). A la différence d'*Exhibit* de Simile, ClioPatria offre une architecture client-serveur qui retourne des objets auxquels il est possible d'appliquer des styles de présentation et des stratégies d'interaction personnalisés. Nous présentons dans la suite les données utilisées pour notre expérimentation et nous montrons comment la formalisation des métadonnées permet d'obtenir des dimensions pour présenter les résultats d'une recherche ou pour guider l'utilisateur dans une navigation par facettes.

5.1 Données utilisées

Le jeu de données utilisé dans notre expérimentation comprend : 100000 dépêches de presse en anglais et en français, 2557 photos et 8 heures de journaux télévisés couvrant la période juin et juillet 2006 (Tableau 1). Suivant les quatre étapes détaillées dans la section 4, nous avons analysé ces dépêches pour enrichir automatiquement les métadonnées. Ainsi, l'analyse de la légende des 2557 photos fournit 217 personnes connues dans DBPedia et 426 lieux connus dans Geonames. L'évaluation manuelle des résultats montre que le service Geonames a tendance à toujours trouver d'abord une ville américaine pour chaque requête. Les dépêches contiennent toutefois la plupart du temps un nom de ville et de pays, ce qui permet d'avoir une reconnaissance précise du lieu mentionné dans la dépêche. Les quelques erreurs que nous avons observées pro-

[23] `http://e-culture.multimedian.nl/software/ClioPatria.shtml`

viennent d'un mauvais typage de l'entité nommée (e.g. `Australia` a parfois été typé comme une `Person`). Nous sommes actuellement en train d'évaluer des algorithmes plus sophistiqués tels que IdentityRank (Fernández *et al.*, 2007) pour minimiser ces problèmes d'homonymie.

Description	Nb de triplets RDF
Ontologies générales : NAR, NewsML-G2, DC, VRA, FOAF	7,336
Ontologies de domaine et BC (football)	104,358
Thésauri : IPTC NewsCodes, Thesaurus INA	34,903
Ressources externes : Geonames, DBPedia	53,468
Fil des dépêches de l'AFP en anglais (Juin et Juillet 2006)	804,446
Photos de l'AFP de la coupe du monde 2006	61,311
Journaux télévisés archivés à l'INA (Juin et Juillet 2006)	1,932
Total	1,067.754

TAB. 1 – Nombre de triplets RDF chargés dans ClioPatria

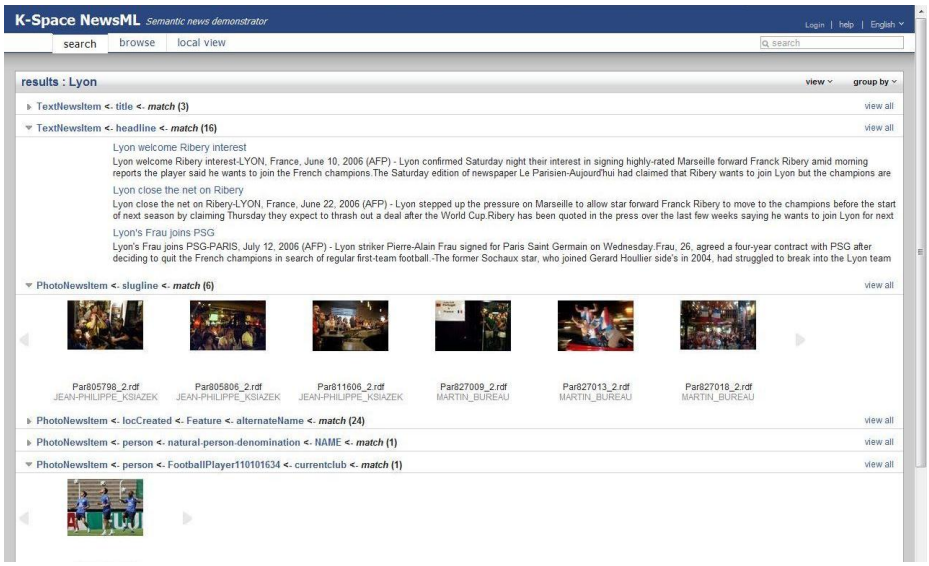


FIG. 2 – Recherche avec le terme "Lyon" dans notre moteur de recherche sémantique

5.2 Recherche sémantique d'actualités

La figure 2 montre le résultat de notre système pour la requête "Lyon". Les actualités sont groupées selon le chemin dans le graphe RDF qui mène à la propriété pour laquelle la valeur est pertinente pour la requête. Dans notre cas, le système retourne les dépêches où "Lyon" apparaît dans les propriétés `title`, `headline`, `slugline`,

etc. De plus, le type des dépêches permet de personnaliser leur présentation. Ainsi, les trois premières lignes d'une dépêche textuelle seront affichées alors que les images sont présentées sous la forme d'un carrousel.

Le dernier groupe affiché dans la figure 2 contient une photo unique avec trois footballeurs. De manière surprenante, les métadonnées de cette photo ne contiennent pas le terme "Lyon". En revanche, la légende mentionne le joueur Juninho Pernambucano, personne connue dans DBPedia qui fournit des informations supplémentaires sur cette personne telles que sa date de naissance ou les clubs dans lesquels il a joué, et en particulier son club actuel : "Lyon". Cet exemple illustre la puissance (et les limites) de notre système : il est désormais possible de trouver des documents ayant une relation, même lointaine, avec la requête et d'expliciter cette relation (le graphe RDF). Ce résultat figurera toutefois en queue d'un classement de pertinence dû à la longueur du graphe.

De manière similaire, la requête "Saksamaa" retourne un groupe de 679 photos pour lesquelles ce terme n'apparaît jamais dans les métadonnées. L'explication provient du fait que toutes ces photos ont été prises pendant la coupe du monde en Allemagne, une entité nommée reconnue comme un lieu et donc lié à la base Geonames, qui fournit également le libellé du pays dans toutes les langues, *Saksamaa* signifiant Allemagne en éthiopien.

5.3 Exploration sémantique des actualités

En plus d'une interface de recherche, notre prototype contient un navigateur à facettes pour mieux explorer le contenu d'une archive. Les facettes correspondent à certaines propriétés jugées d'intérêt particulier dans les métadonnées et sont configurables par l'utilisateur. Ainsi, nous avons défini une vue dédiée au football qui contient les propriétés `subject`, `slugline`, `locCreated`, `location` et `person`. Le lien avec la base de connaissance Geonames permet de proposer des vues plus riches pour présenter les actualités. La figure 3 montre par exemple toutes les photos du joueur Zinedine Zidane sur une carte à l'emplacement où les photos ont été prises (drapeau bleu) ou à l'emplacement mentionné dans les dépêches (drapeau rouge).

Cette vue présente parfois un réel intérêt. Deux séries de photos dont les mots clés sont `FBL-WC2006-MATCH64ITA-FRA` et `FBL-WC2006-MATCH64-FRA-ITA` semblent en effet concerner le même événement. Leur affichage sur une carte permet immédiatement de comprendre ce qui les différencie : les dépêches ont été produites soit en Italie, soit en France adoptant un point de vue inévitablement biaisé de la finale.

6 Conclusion et perspectives

Nous avons décrit dans cet article une méthode composée de quatre étapes pour construire une infrastructure basée sur les ontologies pour les actualités. Ces recommandations sont complémentaires des patrons de conception ontologique et des guides de bonnes pratiques pour publier des ontologies dans le web de données. Nous avons discuté nos décisions de modélisation : au niveau ontologique, nous préconisons d'aplatir les structures XML quand l'ontologie est construite à partir d'un schéma et de ré-utiliser le plus possible les vocabulaires existants ; au niveau des instances, nous recommandons

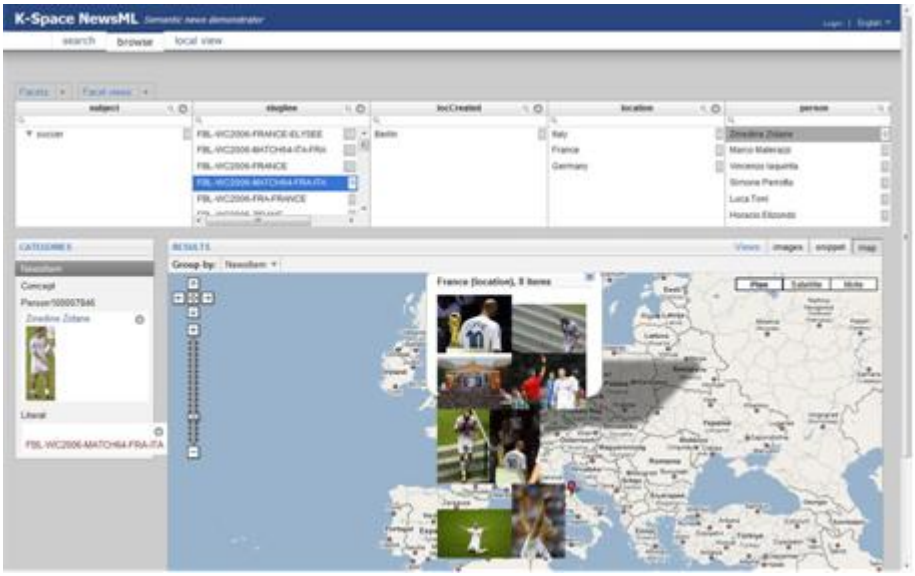


FIG. 3 – Explorer les photos prises pendant la finale de la coupe du monde

d'enrichir et de lier les métadonnées avec des thésauri SKOS ou de la connaissance formalisée disponible dans le web de données telle que les bases Geonames ou DBPedia. L'ontologie NAR est actuellement en cours de revue par l'IPTC et pourrait être approuvée prochainement. Nous avons présenté un prototype pour chercher et explorer des actualités. Les interfaces utilisent la richesse des métadonnées sémantiques pour grouper, classer et présenter le résultat des requêtes. Cet environnement est publiquement disponible à `http://newsml.cwi.nl/explore/search`.

Le temps est une dimension fondamentale dans le domaine des actualités et notre système possède également des vues temporelles. Raisonner sur des données temporelles est néanmoins un problème complexe. Nous planifions d'inclure prochainement une ontologie du temps[24] et les relations temporelles de l'ontologie DOLCE dans le but de proposer une vue des dépêches agrégées par sujet, par jour, mois ou année. Notre système fonctionne actuellement avec des données statiques qui ont été préalablement analysées. Une évolution naturelle consiste à proposer un environnement dynamique alimenté par des fils de dépêches en continu. Finalement, une évaluation de notre prototype par des journalistes de l'AFP est planifiée.

Références

ARNDT R., TRONCY R., STAAB S., HARDMAN L. & VACURA M. (2007). COMM : Designing a Well-Founded Multimedia Ontology for the Web. In *6^{th} International Semantic Web Conference (ISWC'07)*, p. 30–43, Busan, Korea.

[24] `http://www.w3.org/TR/owl-time/`

ASSEM M. V., MENKEN M. R., SCHREIBER G., WIELEMAKER J. & WIELINGA B. (2004). A Method for Converting Thesauri to RDF/OWL. In *3rd International Semantic Web Conference (ISWC'04)*, p. 17–31, Hiroshima, Japan.

BACHIMONT B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Habilitation à diriger des recherches, Université de Compiègne.

CASTELLS P., PERDRIX F., PULIDO E., RICO M., BENJAMINS R., CONTRERAS J. & LORÉS J. (2004). Neptuno : Semantic Web Technologies for a Digital Newspaper Archive. In *1st European Semantic Web Symposium (ESWS'04)*, p. 445–458, Heraklion, Crete.

FERNÁNDEZ N., BLÁZQUEZ J. M., ARIAS J., SÁNCHEZ L., SINTEK M., BERNARDI A., FUENTES M., MARRARA A. & BEN-ASHER Z. (2006). NEWS : Bringing Semantic Web Technologies into News Agencies. In *5th International Semantic Web Conference (ISWC'06)*, p. 778–791, Athens, Georgia, USA.

FERNÁNDEZ N., BLÁZQUEZ J. M., SÁNCHEZ L. & BERNARDI A. (2007). IdentityRank : Named Entity Disambiguation in the Context of the NEWS Project. In *4th European Semantic Web Conference (ESWC'07)*, p. 640–657, Innsbruck, Austria.

GARCIA R., PERDRIX F., GIL R. & OLIVA M. (2008). The semantic web as a newspaper media convergence facilitator. *Journal of Web Semantics*, **6**(2), 151–161.

HAUSENBLAS M., BOLL S., BÜRGER T., CELMA O., HALASCHEK-WIENER C., MANNENS E. & TRONCY R. (2007). Multimedia Vocabularies on the Semantic Web. W3C Multimedia Semantics Incubator Group Report. `http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/`.

HILDEBRAND M., OSSENBRUGGEN J. V. & HARDMAN L. (2006). /facet : A Browser for Heterogeneous Semantic Web Repositories. In *5th International Semantic Web Conference (ISWC'06)*, p. 272–285, Athens, Georgia, USA.

MPEG-7 (2001). Multimedia Content Description Interface. ISO/IEC 15938.

SCHENK S. & STAAB S. (2008). Networked Graphs : A Declarative Mechanism for SPARQL Rules, SPARQL Views and RDF Data Integration on the Web. In *17th International World Wide Web Conference (WWW'08)*, Beijing, China.

TORDAI A., OMELAYENKO B. & SCHREIBER G. (2007). Semantic Excavation of the City of Books. In *Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM'07)*, p. 39–46.

TRONCY R. (2008). Bringing the IPTC News Architecture into the Semantic Web. In *7th International Semantic Web Conference (ISWC'08)*, p. 483–498, Karlsruhe, Germany.

TRONCY R., CELMA O., LITTLE S., GARCÍA R. & TSINARAKI C. (2007). MPEG-7 based Multimedia Ontologies : Interoperability Support or Interoperability Issue ? In *1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MAReSO)*, Genova, Italy.

TRONCY R. & ISAAC A. (2002). DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. In *13th Journées d'Ingénierie des Connaissances (IC'02)*, p. 63–74, Rouen, France.

WIELEMAKER J., HILDEBRAND M., OSSENBRUGGEN J. V. & SCHREIBER G. (2008). Thesaurus-based search in large heterogeneous collections. In *7th International Semantic Web Conference (ISWC'08)*, p. 695–708, Karlsruhe, Germany.

Méta-modèle général de description de ressources terminologiques et ontologiques

Pierre-Yves Vandenbussche[1,2], Jean Charlet[1,3]

[1] INSERM UMRS 872, éq.20, 15, rue de l'école de médecine, 75006 Paris, France

[2] MONDECA, 3, cité Nollez, 75018 Paris, France
pierre-yves.vandenbussche@etu.jussieu.fr

[3] DSI AP-HP, Paris, FRANCE
jean.charlet@spim.jussieu.fr

Résumé : L'intégration des ressources terminologiques et ontologiques d'un domaine est un enjeu majeur en vue de leur pleine exploitation par des organisations. Cette intégration est rendue difficile par l'hétérogénéité des ressources et de leur formalisme de représentation (SKOS, BS 8723, etc.). Ces formalismes se différencient principalement par leur richesse d'expressivité. Dans cet article, nous proposons un nouveau méta-modèle de représentation de terminologies et d'ontologies. Celui-ci a une double particularité. Il propose un formalisme de représentation plus général car il fait l'union de chacune des spécificités des formalismes existants tout en définissant de nouveaux constructeurs qui apportent un pouvoir d'expressivité supplémentaire aux ressources terminologiques. Il se base sur les technologies d'Ingénierie Dirigée par les Modèles, en vue de permettre une intégration automatique de ressources terminologiques provenant d'un formalisme. **Mots-clés** : méta-modèle, terminologie, ontologie, interopérabilité, opérationnalisation.

1 Introduction

Depuis l'apparition de l'informatique et du Web, les ressources terminologiques et ontologiques, des plus linguistiques aux plus formelles, ont pris une position centrale qui rend possible le partage d'informations normalisées. Ces ressources complémentaires doivent souvent être gérées de façon cohérente, les unes par rapport aux autres et dans le temps afin de répondre aux demandes croissantes des utilisateurs. Mais la diversité et le nombre des ressources rendent leur mise en œuvre difficile.

Nous proposons un méta-modèle général de description des ressources terminologiques et ontologiques. Notre approche se positionne sur un même plan que les standards et les normes de représentation spécifiques à un type de ressources mais apporte le moyen de gérer de manière générale une ou plusieurs ressources en conservant l'expressivité de chacune.

Dans la suite de notre article nous rappelons dans la section 2 les motivations de

notre recherche. Après avoir retenu certains critères d'analyse, nous étudions dans la section 3 quelques ressources existantes. De cette étude nous identifions les éléments de modélisation nécessaires à notre méta-modèle et les limites d'utilisations que nous résolvons en partie dans notre méta-modèle. Dans la section 4, nous mettons en exergue à partir des normes et des standards de représentation adaptés à des types de ressources particuliers, les modélisation utiles pour notre méta-modèle que nous présentons dans la section 5. Les discussions et conclusions apportent pour finir des réflexions pour la suite de ce travail.

2 Le contexte de recherche

Depuis plus d'une dizaine d'années, la discipline de l'ingénierie de la connaissance travaille à l'élaboration de ressources terminologiques et ontologiques et à la mise à disposition de connaissances. Notre approche de manière générale et les travaux exposés dans cet article en particulier se situent au cœur de ces enjeux en proposant d'une part, le moyen de représenter et de faire coexister l'ensemble de ces ressources et d'autre part, des services satisfaisant les besoins des utilisateurs grâce à leur modélisation. Dans cette section nous revenons sur certaines définitions des ressources auxquelles nous sommes confrontés puis nous étudions quelles attentes nous avons de ces ressources.

2.1 La diversité des ressources

Il existe tout un panel de systèmes de structuration pour représenter les connaissances : liste contrôlée, classification, thésaurus, terminologie, ontologie... Commençons par replacer l'éventail des ressources auxquelles nous sommes confrontés et dont nous empruntons certaines définitions aux écrits existants sur le domaine.

Une *classification* est « la répartition systématique en classes, en catégories d'êtres, de choses ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude ; c'est aussi le résultat de cette opération » (Bourigault *et al.*, 2004). Notons que les classes peuvent être organisées entre elles hiérarchiquement selon un principe générique-spécifique. Un exemple de classification est la CIM-10 [1].

Un *thésaurus* est un ensemble structuré de termes nécessaires à son utilisation au sein d'une hiérarchie de concepts liés par des relations sémantiques. Eurovoc (cf. 3.2.1) et le MeSH (cf. 3.2.2) sont des thésaurus.

Une *terminologie* « est une liste de termes d'un domaine ou sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques, cette liste étant ou non structurée » (Lefèvre, 2000). Contrairement à un thésaurus, dans une terminologie, l'accent est mis sur l'exhaustivité des termes (synonymes, abréviations...). Un exemple de terminologie est la SNOMED 3.5.

Les attendus des *ontologies* ont beaucoup changé depuis le début de leur utilisation en informatique dans les années 1990. En effet leurs objectifs sont devenus plus modestes et plus réalistes. On peut les définir comme un ensemble de concepts et de relations

1. 10e édition de la classification internationale des maladies. Voir : `http://taurus.unine.ch/icd10/`

pour une utilisation particulière d'un domaine déterminé ; cette structure repose sur une formalisation avouée afin d'effectuer des inférences dans un système informatique. Un exemple d'ontologie est la SNOMED-CT (cf. 3.2.3).

Dans la suite de cet article, nous désignons par *Ressource Terminologique et Ontologique* (RTO) l'ensemble de ces artefacts (Bourigault *et al.*, 2004).

Face à ce pluralisme, un constat peut être fait : la complexité au sein de chaque structure de ressources rend difficile leur catégorisation. Il existe ainsi des terminologies plus ou moins formelles, des thésaurus avec une structure plus ou moins complexe, des ontologies avec ou sans contraintes sur leurs relations... Le classement d'une ressource dans un type particulier de structuration est une tâche ardue en témoignent les critiques de certaines ressources qui prétendent être ce qu'elles ne sont pas.

2.2 Les attentes d'utilisation

Comme nous venons de le voir, il existe une diversité dans la représentation et l'organisation des connaissances qui s'explique par des utilisations différentes. Néanmoins ces structurations ont toutes pour vocation d'appréhender de l'information, de la partager et de permettre un traitement humain et computationnel. Identifions les souhaits d'utilisation de ces ressources :

2.2.1 La recherche d'information

Ce premier point est le plus critique : le but même d'une RTO est de faciliter l'accès à des informations normalisées et plus ou moins formalisées. La recherche d'une information par une personne peut se faire par trois approches différentes : par recherche textuelle ; par recherche arborescente ; par recherche sur le réseau.

- La recherche textuelle est fortement dépendante de la **richesse linguistique** de la ressource. La prise en compte de l'aspect terminologique et la finesse de l'expressivité linguistique vont être des facteurs importants. Par exemple, la représentation des relations de synonymie et de méronymie entre termes accroît considérablement les recherches relatives à un terme ou concept donné.
- La recherche arborescente est liée à la **structuration** même de la RTO. L'organisation d'une terminologie sous forme arborescente (par des relations génériques spécifiques ou partitives) devient nécessaire quand le **volume** est trop important et qu'une simple liste ne suffit plus. Si nous prenons l'exemple de la SNOMED-CT (cf. 3.2.3), on peut entrer dans la terminologie par 19 hiérarchies de haut niveau (p. ex. « Clinical finding/disorder » ou encore « Body structure »). Ces arborescences posent toutefois des problèmes, tels que l'adhésion aux principes de structuration sur lesquels nous reviendrons (cf. 3.3).
- La recherche sur le réseau de connaissances repose sur les **relations sémantiques définies** dans ce réseau. Il s'agit d'obtenir l'ensemble des éléments ou parties du réseau sémantique qui vérifient une requête donnée sur le système. On pourrait ainsi rechercher dans une ontologie de médecine « Toutes les maladies ayant pour localisation l'abdomen ». Cette recherche implique pour être possible, que cette ontologie possède par exemple : les entités *Maladie*, *Abdomen* ; une relation *localisation* d'une maladie dans une région du corps.

2.2.2 L'interopérabilité

Avec les développements du Web et la mondialisation, le partage d'information est au centre des problématiques actuelles. Dans le milieu médical par exemple, les attentes tendent vers des systèmes d'information partagés entre les services et entre les établissements de santé. Parallèlement, les pratiques et les utilisations des ressources ont évolué et il devient caduc de penser qu'il existe *une* terminologie qui capture l'ensemble des connaissances d'un domaine (Aussenac-Gilles, 2005). Ce besoin d'échange d'informations et la prise en compte de plusieurs ressources traduisent la notion d'interopérabilité que P. Miller définit par : « process of ensuring that the systems, procedures and culture of an organisation are managed in such a way as to maximise opportunities for exchange and re-use of information, whether internally or externally. » (Miller, 2000).

Considérons l'interopérabilité sous 3 niveaux (Ferreira da Silva *et al.*, 2006) :

- L'interopérabilité syntaxique constitue le premier niveau et concerne le format de représentation des connaissances. Le fait qu'une ressource soit exprimée ou puisse être traduite sous un format **standardisé ou normé** conduit vers l'interopérabilité syntaxique.
- L'interopérabilité structurelle, à un second niveau, fait référence à l'organisation de l'information au sein d'une ressource qui est régie par son **modèle** sous-jacent.
- L'interopérabilité sémantique est l'interprétation que l'on a d'une représentation d'un domaine fondée sur un consensus. La sémantique est donnée par les **symboles** que l'on a définis sur les primitives du modèle.

La sémantique se définit sur les primitives du modèle exprimées dans une syntaxe donnée. Ces trois niveaux sont donc interdépendants.

3 Expressivité et utilisation des RTOs : trois études de cas

Des attentes d'utilisation que nous venons d'énoncer, nous pouvons maintenant extraire des indicateurs pour juger de l'expressivité et de l'utilisation des RTOs qui seront repris dans une synthèse :

3.1 L'identification de critères d'analyse

- **Le périmètre du domaine décrit** : la ressource doit avoir une définition claire de ses prétentions. Tout d'abord, la finalité de la ressource doit être connue, c'est-à-dire l'application pour laquelle elle a été construite. Si une ressource a un usage donné, alors elle décrit un domaine particulier avec une granularité de l'information représentée.
- **La volumétrie** : dépendant de son périmètre et de la granularité voulue, le volume d'une RTO peut fortement varier.
- **La richesse linguistique** : la linguistique est un point d'entrée pour la recherche par une personne. L'expressivité de la linguistique va dépendre des primitives dé-

finies dans le modèle telles que la gestion de la langue sur un terme, les relations de synonymie et de traduction...
– **L'expressivité formelle** : le caractère formel sous forme d'une logique mathématique permet d'opérer des traitements automatiques sur une ressource, par exemple grâce à la définition formelle d'une relation de subsomption.
– **La conformité aux normes, aux standards et aux recommandations** : l'utilisation de standards pour la construction d'une ressource ou pour l'échange de cette ressource favorise l'interopérabilité.

3.2 Étude de trois RTOs

Pour comprendre la diversité d'expressivité des RTOs actuelles, nous avons étudié les ressources suivantes en nous basant sur les critères que nous venons d'énoncer : Eurovoc, SNOMED-CT, MeSH.

3.2.1 Eurovoc

Eurovoc « est un thésaurus multilingue (21 langues) couvrant tous les domaines de l'activité de l'Union Européenne, Il permet d'indexer les documents dans les systèmes documentaires des institutions Européennes et de leurs utilisateurs »[2]. La construction de ce thésaurus est conforme aux normes ISO 2788-1986 et ISO 5964-1985 et se compose de *descripteurs* ou termes préférentiels, de *non-descripteurs* ou termes non-préférentiels organisés au sein d'une classification hiérarchique à deux niveaux (domaines et microthésaurus). Les relations sémantiques utilisées sont au nombre de quatre : relation d'appartenance au microthésaurus (MT) ; relation d'équivalence de synonymie entre un terme préférentiel et un terme non-préférentiel (UF,USE pour Used For et Use) ; relation hiérarchique (BT,NT pour Broader Term et Narrower Term) ; relation associative (RT pour Related Term). En terme de volumétrie, il existe 6645 concepts reflétés par autant de termes préférés et 10 000 relations dans chaque langue pour plus de 260 000 termes (préférés ou non) toutes langues confondues en 2007.

Le périmètre d'Eurovoc a été établi : le but est de répondre aux besoins de systèmes documentaires généraux sur les activités de l'Union Européenne ; il ne convient toutefois pas à l'indexation et à la recherche de documents spécialisés. Le thésaurus indique qu'il ne peut pas prétendre couvrir les différentes réalités nationales à un niveau suffisamment spécifique (p.ex. existence en Belgique du *Conseil supérieur de la Justice*). Il est toutefois possible au travers d'une fiche de maintenance, de spécifier un nouveau besoin.

3.2.2 MeSH

Le MeSH (Medical Subject Heading) est le thésaurus de référence dans le domaine biomédical produit par le NLM (U.S. National Library of Medicine)[3]. La structure du MeSH est à trois niveaux : un ensemble de termes dont un préférentiel, désigne

2. Voir `http://europa.eu/eurovoc/`
3. Voir `http://ist.inserm.fr:3201/inserm08/index.html`

un *concept* qui fait partie d'une classe de concepts appelée *descripteur*. La navigation dans le thésaurus se fait par recherche ou en entrant par les *Main Headings*. Le MeSH contient près de 25 000 descripteurs et plus de 455 000 termes en janvier 2009. Trois types de relations, ne reposant pas sur la logique formelle, sont utilisées : hiérarchique, synonymique et de proximité sémantique.

Produit depuis 1960, le MeSH est utilisé pour l'indexation par de nombreuses bibliothèques et institutions à travers le monde. le thésaurus a été traduit dans de nombreuses langues mais il est à déplorer qu'il n'existe aucune possibilité de navigation à travers le multilinguisme [4]. Le MeSH n'est conforme à aucune norme, il permet seulement des échanges via le format XML.

3.2.3 SNOMED-CT

La SNOMED-CT (Systematized Nomenclature of MEDicine-Clinical Terms) est une ontologie multilingue (pour l'instant en deux langues et un dialecte) de la santé clinique [5]. Il s'agit d'une structure hiérarchique de concepts désignés par des *descriptions* (termes) sur plus de 31 niveaux de subsomption. la SNOMED-CT est conforme au modèle HL7 version 3 et repose sur une sémantique formelle (Logique de description). Elle contient plus de 311 000 concepts, près de 800 000 termes et 1 360 000 relations en janvier 2008.

Cette terminologie a pour vocation d'être utilisée pour tous documents cliniques allant du dossier patient électronique et des systèmes informatiques des hôpitaux jusqu'à la télé-médecine.

3.3 Synthèse des RTOs étudiées

RTO / Caractéristiques	Eurovoc	MeSH	SNOMED-CT
Périmètre	L'activité de l'UE. Couverture assez large. Faible granularité (7 niveaux de prof.)	Le domaine biomédicale. Couverture assez large. Granularité normale (11 niveaux de prof.)	La santé clinique. Couverture très large. Granularité très fine (31 niveaux de prof.)
Volumétrie	6645 concepts 260 000 termes 210 000 relations	25 000 concepts 455 000 termes	311 000 concepts 800 000 termes 1 360 000 relations
Richesse linguistique	synonymie, associative.	synonymie, associative	synonymie
expressivité formelle	relation hiérarchique	relation hiérarchique	Les concepts et la relation IS_A définis par la logique de description
conformité aux normes	ISO 2788-1986 ISO 5964-1985		HL7 version 3

4. Seul l'INSERM ayant la responsabilité de sa traduction en Français met à disposition une version bilingue Anglais-Français.

5. Voir `http://www.ihtsdo.org/snomed-ct/`

De l'étude de ces RTOs (dont certains résultats ont été présentés dans la partie précédente), on peut dégager certains enseignements :

- On peut considérer qu'un volume trop important peut nuire à l'utilisation effective d'une ressource par une personne. Si nous prenons la SNOMED-CT, la quantité d'information à disposition est très importante en raison d'un périmètre très large et une granularité très fine. La simple recherche arborescente d'un concept dans cette terminologie est d'autant plus ardue. En pratique, l'utilisation qui pourrait en être faite par exemple au sein d'un service hospitalier pour coder un acte, se restreindrait à un sous-ensemble défini des concepts de la RTO de référence. *Comment peut-on permettre l'utilisation d'un sous-ensemble d'une RTO de référence ?*
- Les besoins en recherche textuelle imposent une représentation minimale de primitives linguistiques, mettant à minima en avant un terme préférentiel lié à plusieurs termes synonymes comme c'est le cas dans les RTOs étudiées. La représentation de synonymes permet d'élargir le nombre de réponses à une recherche textuelle, les termes non préférentiels amenant le résultat sur le terme préférentiel. Néanmoins d'autres relations existent (p.ex. *RelatedTerm* dans Eurovoc) et permettent de présenter plus d'information à l'utilisateur. *Quelle expressivité linguistique est utile pour l'exploitation ?*
- L'un des premiers rôles d'une terminologie est de réduire l'ambiguïté. Le sens donné à un concept vient aussi bien de sa position dans le réseau de connaissances que de la définition logique qu'on lui a donnée. Seule la SNOMED-CT décrit de manière formelle ses concepts.
- L'interopérabilité syntaxique et structurelle, quant à elle, a besoin de standardisation. Nous le voyons dans notre étude : mis à part le MeSH qui ne repose sur aucune norme, chaque ressource est conforme à des normes différentes même si ces ressources sont complémentaires. Les enjeux autour de cette interopérabilité inter-référentiels sont multiples : Différencier une RTO d'interface et une RTO de référence (Rosenbloom *et al.*, 2006) ; permettre la communication entre deux utilisateurs faisant référence à deux RTOs différentes... *Quels sont les pré-requis à une interconnexion entre des référentiels différents et comment dès lors garantir l'interopérabilité ?*
- En plus des points soulevés par cette analyse, une autre question nous semble importante : celle de l'adhésion aux principes de modélisation. Toute élaboration d'une ressource terminologique ou ontologique repose d'une part sur une méthode de conceptualisation propre au modélisateur de cette ressource (ou à un consensus) et d'autre part sur la finalité dans laquelle la ressource s'inscrit. *Comment rendre la recherche d'un concept plus efficace en ne connaissant pas à priori le mode de catégorisation utilisé dans l'élaboration de cette RTO ?*

4 Les standards et normes de terminologies et d'ontologies

Dans le domaine des RTOs certaines normes existent et facilitent ainsi l'interopérabilité (cf. 2.2.2). Dans cette section nous présentons les normes et les standards principaux plus ou moins spécialisés pour un type de ressources particulier. Notre méta-modèle s'inspire de la modélisation de ces normes avec lesquelles il doit être conforme (plus général) pour prétendre ne pas réduire l'interopérabilité.

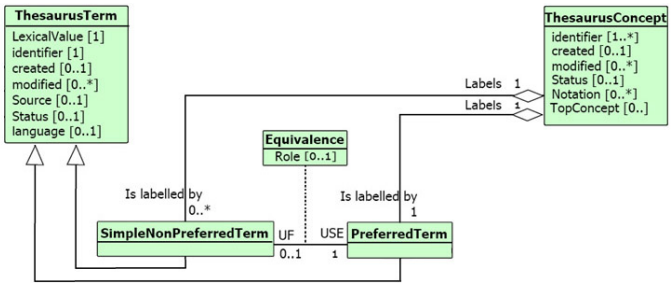


FIGURE 1 – BS8723 : partie du modèle centrée sur la terminologie.

4.1 SKOS [6]

SKOS se définit comme un langage qui permet la représentation de systèmes d'organisation de connaissances tels que thésaurus, taxonomies, ou tout autre type de vocabulaire contrôlé ou structuré. Ce standard met à disposition certaines primitives dédiées à la linguistique : on a d'une part le *Concept* qui représente une notion et d'autre part la terminologie avec pour chaque langue, un terme préféré *prefLabel*, des synonymes *altLabel* et des chaînes de caractères servant à la recherche (p.ex. un code d'un concept) *hiddenLabel*. SKOS est une famille de langages extensibles. L'extension XL (pour eXtended Labels) considère les *Label*s comme des ressources, ce qui pourrait nous permettre de redéfinir des relations (p.ex. une relation de traduction) entre ces *Labels* pour augmenter la richesse linguistique. SKOS définit certaines relations comme transitives, ce qui permet d'effectuer des inférences logiques.

4.2 BS 8723 [7]

Les normes ISO concernant les terminologies sont en train d'évoluer [8] grâce aux travaux de la British Standard et de son projet BS 8723. La gestion de la linguistique est ici plus fine que dans le standard SKOS. Comme nous le montre la figure 1, un terme préféré (terme qui est utilisé dans une langue pour désigner un concept) est ici exprimé sous la forme d'une primitive, de même que le terme non préféré. Ceci leur permet d'exprimer une relation de synonymie directe entre ces élément du modèle.

4.3 OWL

Le standard OWL permet d'exprimer une connaissance en utilisant une sémantique formelle basée sur la logique des prédicats. A cet égard, OWL convient parfaitement à l'expressivité formelle nécessaire à l'interopérabilité et sert souvent de format de définition de méta-modèle dans la représentation de connaissances. Son formalisme et son expressivité permettent de représenter des ontologies. Cependant, son expressivité et son caractère généraliste sont mal adaptés à

6. Simple Knowledge Organisation System (SKOS) développé dans le cadre du W3C depuis 2003. Voir : `http://www.w3.org/2004/02/skos/`

7. Voir : `http://schemas.bs8723.org`

8. le projet ISO 25964 va remplacer les normes ISO 2788 concernant l'élaboration et le développement de thésaurus monolingue et ISO 5964 concernant l'élaboration et le développement de thésaurus multilingue en se basant sur les travaux de la BS8723.

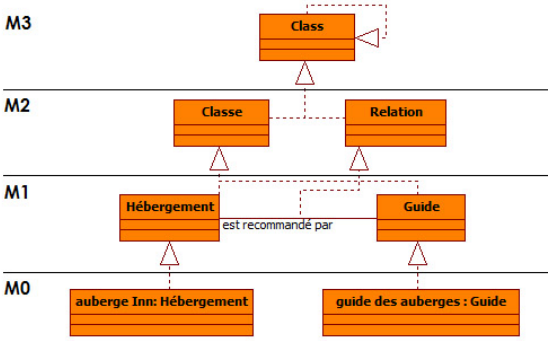


FIGURE 2 – Exemple de méta-modélisation en 4 couches

la description de terminologies ou de thésaurii contrairement au standard SKOS ou à la norme BS 8723. OWL a toutefois les éléments pour décrire ces normes mais ne permet pas nativement de représenter des terminologies.

5 Présentation d'un méta-modèle général de RTO

Certaines normes et standards existent pour décrire un type particulier de RTO. Cependant l'expressivité de leurs primitives ne permet pas de représenter l'ensemble des problèmes conceptuels posés par l'ensemble des ressources terminologiques et ontologiques (El Hachani, 2005). Les motivations qui nous poussent à définir un nouveau méta-modèle général sont doubles : supporter les modèles décrivant les RTOs existantes (p.ex. SKOS, BS 8723 etc.) et proposer grâce à la modélisation une opérationnalisation effective des ressources. L'objectif pour nous est d'abord de définir un méta-modèle en prenant en considération les meilleures pratiques dans les modèles et normes existantes, ensuite d'améliorer cette modélisation par des apports qui rendent possible une meilleure utilisation des ressources. D'autres projets se basent sur une démarche similaire, nous citerons le projet DAFOE. Dans ce projet, un méta-modèle permet également de représenter les RTOs mais à des étapes différentes partant des corpus et menant à l'élaboration d'une ontologie (Charlet *et al.*, 2008). Le méta-modèle du projet DAFOE a pour but d'appréhender ces étapes successives et de faciliter ce processus menant à la formalisation de la connaissance. Notre approche se différencie de par son objectif de mise en relation de ces ressources sur un même plan.

5.1 Méta-modélisation

Face au besoin croissant d'intégration logicielle et d'interopérabilité des systèmes d'information, l'ingénierie logicielle a mis en place depuis 2000, l'Ingénierie Dirigée par les Modèles (IDM) (Bernstein, 2003). Toute base de connaissance afin d'être opérationnalisée est représentée à l'aide d'un modèle exprimé dans un langage particulier. Ceci définit l'activité de méta-modélisation, c'est-à-dire suggérer un formalisme de modélisation, chaque modèle se conformant à un méta-modèle prédéfini. De même, à un niveau d'abstraction supérieur, le méta-modèle a besoin d'être clairement défini par un méta-méta-modèle. Afin d'éviter une décomposition infinie de niveaux d'abstraction, un patron d'architecture en 4 couches, illustré à la figure 2, sert

maintenant de référence. Le modèle M0 étant le monde réel que nous conceptualisons, il serait dans notre exemple sur le domaine de l'hôtellerie : *auberge Inn*, le modèle M1 contiendrait la classe *Hébergement*. Le modèle M2 définirait l'ensemble des éléments pour exprimer le niveau M1 c'est à dire la classe *Classe*, la classe *Relation*. un dernier niveau M3 réflexif (qui s'auto-définit) contiendrait une classe *Class* qui permet de définir les deux classes du niveau 2. Pour bien comprendre cette architecture nous vous renvoyons sur la spécification du MOF [9].

5.2 L'expressivité linguistique

Comme nous avons pu le voir, l'expressivité linguistique est très importante pour l'utilisation finale et pour la maintenance d'une ressource. La séparation entre un niveau terminologique et un niveau conceptuel permet notamment de maintenir la partie terminologique indépendamment des problèmes conceptuels comme l'expose Reymonet *et al.* (2007). C'est dans ce sens que la norme BS 8723 modélise un terme par une classe *Terme* séparée de la modélisation des concepts. Mais de quelles relations et de quelles propriétés avons-nous besoin ? D'après notre expérience, une représentation exhaustive de tous les cas possibles n'est pas rationnelle. La plupart des cas ne nécessitent qu'une représentation simple, ainsi notre méta-modèle doit permettre nativement de représenter les relations et attributs les plus communément utilisés tout en permettant de redéfinir plus précisément certains éléments, ce qui permet aux modèles d'évoluer (cf. figure3).

Ces apports linguistiques permettent de résoudre certaines problématiques précédemment identifiées dans notre synthèse (cf. 3.3) : le méta-modèle distinguant la gestion linguistique et conceptuelle, la maintenance de la RTO s'en voit simplifiée ; le lien d'un terme vers son texte source est capturé par le méta-modèle assurant ainsi une traçabilité très utile particulièrement lors de la construction ou l'enrichissement de connaissances à partir de corpus : la souplesse d'un méta-modèle extensible autorisant l'ajout par exemple de la fréquence d'un terme dans le corpus d'origine.

5.3 Sous-ensemble d'une ressource de référence

La structuration au sein d'un système de connaissances a toujours eu pour but d'organiser les connaissances et de faciliter la recherche d'information. Dès lors que la taille d'une telle ressource est trop importante il devient difficile de retrouver une information. Face à cette problématique soulevée dans notre synthèse, certaines ressources ont adopté des mécanismes de représentation spécifiques. Par exemple : le thésaurus GEMET [10] a mis en place une navigation thématique orthogonale à la hiérarchie verticale habituelle ; la SNOMED-CT a introduit la notion de *RefSet*. C'est un groupement de références de concepts spécialisés pour six utilisations différentes allant d'une simple liste (index) à plat en passant par un groupement par langue jusqu'à une hiérarchie de navigation (taxinomie).

Certains langages permettent la définition de collections [11] n'offrant toutefois pas autant de possibilités d'utilisation que nos deux exemples présentés ci-dessus.

En se basant sur ces fonctionnements, nous avons défini une entité dans notre méta-modèle, nommée ***Concept Group***, qui représente un groupement de concepts. La finalité de cette primitive est de pouvoir définir un sous-ensemble de concepts d'une RTO de référence. Ce sous-ensemble

9. MOF (Meta Object Facility) est un modèle de niveau M3 réflexif, il définit la grammaire d'UML (Unified Modeling Language) au niveau M2. `http://www.omg.org/mof/`

10. GEneral Multilingual Environmental Thesaurus. Voir : http ://www.eionet.europa.eu/gemet

11. Il existe deux primitives RDF permettant de manipuler une collection de concepts : *RDFList* et *RDFSContainer*. SKOS possède la notion de *ConceptScheme* (collection de concepts avec ou sans relation) dont la définition est volontairement permissive.

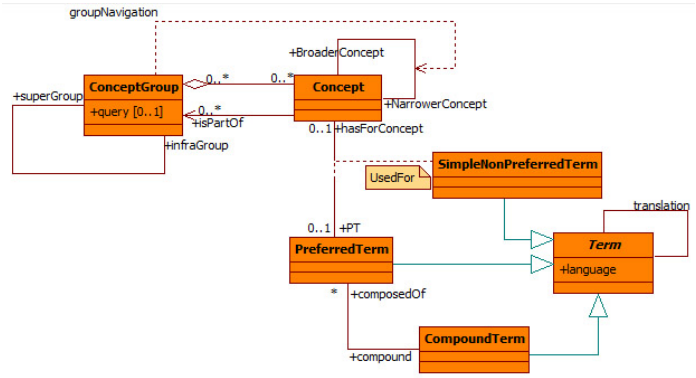


FIGURE 3 – Présentation UML simplifiée d'une partie de notre méta-modèle concernant la linguistique et les groupements de concepts. Cette figure montre l'utilisation de sous-classes de *Term* pour capturer les notions de termes préférés et non préférés reliées à un concept. La notion de *ConceptGroup* faisant référence à un ensemble de concepts définis en intension ou en extension, peut être organisée hiérarchiquement.

correspondant à une utilisation ou à une vue sur la ressource, aurait la possibilité d'être réutilisé ou partagé (pour cela, notre groupement doit avoir un identifiant unique). Nous distinguons deux définitions des concepts dans un *Concept Group*. Premièrement, nous avons ceux définis par **intension** : ensemble des concepts vérifiant la requête d'appartenance au groupement. Deuxièmement, ceux définis par **extension** : ensemble des concepts pointant sur le groupement. Les *Concept Group*s sont également hiérarchisés par une relation.

Ce travail que nous sommes en train de mettre en œuvre a été soumis au groupe de recherche pour l'élaboration de la nouvelle norme ISO 25964. Ils ont ajouté cette primitive mais sans la définition d'appartenance par intension. La définition de la primitive *Concept Group* dans notre méta-modèle répond à la problématique d'utilisation d'un sous-ensemble d'une RTO. Nous avons ainsi introduit par cet élément, de nouveaux points d'entrée dans la RTO : à la recherche arborescente vient s'ajouter une recherche orthogonale par groupement.

6 Discussions et conclusions

L'opérationnalisation de ressources terminologiques et ontologiques gérées de façon cohérente est un enjeu majeur dans l'ingénierie de la connaissance. Les apports d'un méta-modèle général permettent de faciliter l'interopérabilité au sein d'une même application ou entre plusieurs systèmes d'information et d'améliorer l'utilisation de ces ressources. Sur ce dernier point, le choix des primitives définies dans le méta-modèle va déterminer l'exploitation directe des ressources qu'il sera possible de faire. Que ce soit en linguistique ou pour les groupements de concepts, il est possible grâce à l'expressivité d'un méta-modèle, d'améliorer l'utilisation, le partage et la collaboration autour de ces ressources.

Les apports théoriques présentés dans cet article ont déjà été la source d'enrichissements de normes (cf. 5.3). Leurs mises en œuvre au sein de notre outil ITM nous donne l'assurance des résultats d'une telle approche. La description de notre méta-modèle avec une logique mathématique enrichira les traitements automatiques faits par des ordinateurs. Notre modèle ne couvre

pas à ce stade tous les besoins des ressources terminologiques et ontologiques. La gestion dans le temps des versions n'est pas abordée de même qu'une utilisation plus fine des mappings ou projections entre plusieurs ressources. Ces réflexions guideront nos recherches futures.

Références

AUSSENAC-GILLES N. (2005). *Méthodes ascendantes pour l'ingénierie des connaissances* . Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France.

BERNSTEIN P. A. (2003). Applying model management to classical meta data problems. In *CIDR*.

BOURIGAULT D., AUSSENAC-GILLES N. & 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.

CHARLET J., SZULMAN S., PIERRA G., NADAH N., TEGUIAK H. V., AUSSENAC-GILLES N. & NAZARENKO A. (2008). Dafoe : A multimodel and multimethod platform for building domain ontologies. In D. BENSLIMANE, Ed., *2^e Journées Francophones sur les Ontologies*, Lyon, France : ACM.

EL HACHANI M. (2005). *Indexation des documents multilingues d'actualités incluant l'arabe : équivalence interlangues et gestion des connaissances chez les indexeurs*. Thèse de doctorat nr en sciences de l'information et de la communication, Université Lumière Lyon.

FERREIRA DA SILVA C., MÉDINI L., GHAFOUR S. A., HOFFMANN P. & GHODOUS P. (2006). Semantic interoperability of heterogeneous semantic resources. *Electronic Notes in Theoretical Computer Science*, **150**(2), 71–85.

LEFÈVRE P. (2000). *La recherche d'informations (du texte intégral au thèsaurus)*. Hermès Science Publications.

MILLER P. (2000). Interoperability : What is it and why should i want it ? *Ariadne*, **24**.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de ressources termino-ontologiques en owl. In F. TRICHET, Ed., *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 169–180, http ://www.cepadues.com/ : Cépaduès Editions.

ROSENBLOOM S., MILLER R., JOHNSON K., ELKIN P. & BROWN S. (2006). Interface terminologies : Facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association*, **13**(3), 277–288.

Ontologies étendues pour l'annotation sémantique*

Yue Ma, Laurent Audibert, Adeline Nazarenko

Laboratoire d'Informatique de l'université Paris-Nord (LIPN) - UMR 7030
Université Paris 13 - CNRS
99, avenue Jean-Baptiste Clément - F-93430 Villetaneuse, France
[prénom].[nom]@lipn.univ-paris13.fr

Résumé : Cet article tente de formaliser le processus consistant à annoter sémantiquement un texte au regard d'une ontologie. L'annotation sémantique met des fragments de texte en correspondance avec les éléments d'une ontologie, mais toute la difficulté consiste à identifier les fragments à annoter et les étiquettes à leur associer. Nous proposons d'étendre les ontologies par des règles d'annotation sémantique plutôt que par l'ajout de (méta-)propriétés lexicales. Cette solution permet de tirer le meilleur parti des outils de TAL qui produisent chacun à leur niveau des annotations linguistiques. Elle a également le mérite de distinguer clairement le processus d'analyse linguistique et l'interprétation ontologique.
Mots-clés : Ontologie, Annotation sémantique, Lexique, Terminologie, Plateforme d'annotation, Patron linguistiques

1 Introduction

L'annotation sémantique se définit comme le processus qui fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite. Si cette interprétation s'exprime en termes ontologiques, nous parlons d'une interprétation ontologique. C'est le cas que nous considérons ici.

De nombreux systèmes annotent des textes au regard d'une ontologie qui peut être disponible localement ou via une URL (voir la revue proposée par (Dill *et al.*, 2003)). Il s'agit souvent d'annoter manuellement ou semi-automatiquement les texte (Handschuh, 2002; Kogut, 2001) mais les techniques d'apprentissage automatique permettent aussi d'automatiser l'annotation (Ciravegna, 2000).

Dans tous les cas, l'annotation repose sur des ontologies enrichies de connaissances linguistiques qui permettent de mettre en correspondance des éléments de l'ontologie avec des fragments de textes mais la nature de ces connaissances, leur encodage et leur mode d'utilisation varient beaucoup d'un système à l'autre. Nous proposons de

* Ce travail a été réalisé dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche.

les représenter sous la forme de règles d'annotation plutôt que d'adjoindre des propriétés lexicales aux concepts de l'ontologie. Alors que la plupart des systèmes d'annotation mettent l'accent sur l'étiquetage des entités nommées (Kiryakov *et al.*, 2004; Dill *et al.*, 2003), nous considérons le processus d'annotation sémantique dans toute sa complexité.

Une ontologie est une spécification formelle, explicite et consensuelle de la conceptualisation d'un domaine (Gruber, 1993). Une ontologie est constituée d'un ensemble de concepts organisés hiérarchiquement et structurés par des rôles liant ces concepts. Elle peut également comporter des axiomes et être peuplée. Dans ce dernier cas, elle comporte en outre des instances de concepts et des instances de rôles (nous parlons alors de relations entre instances)[1].

2 Annotation sémantique

L'annotation sémantique a pour objectif de formaliser l'interprétation qui peut être faite des textes sous la forme de méta-données attachées aux textes ou à certains de leurs segments. Cette interprétation s'exprime couramment en termes ontologiques quand il s'agit d'associer un type sémantique aux noms des entités mentionnées dans le texte (personnes, gènes, organisations, etc.) ou de les associer à un concept (Kiryakov *et al.*, 2004; Amardeilh *et al.*, 2005). Les entités nommées ne représentent cependant qu'une partie des éléments sémantiquement pertinents et de l'interprétation ontologique qui peut être faite des textes.

Pour illustrer notre propos, voici deux fragments de texte, chacun accompagné d'un exemple d'annotation sémantique respectivement représentés dans les figures 1 et 2 :

1. *Marie lit une pièce de théâtre de Molière.*
2. *The GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK (La protéine GerE inhibe la transcription in vitro du gène sigK qui encode sigmaK).*

2.1 Différents types d'annotation

Nous distinguons les types d'annotations selon la nature de l'élément ontologique auquel elles se rattachent.

Certains mots ou expressions renvoient à des *instances de concepts*. On les désigne traditionnellement sous le terme d'*entités nommées* car ils renvoient à des entités référentielles de manière autonome et conventionnelle (Ehrmann, 2008), comme les mots *Marie* et *SigmaK* ou l'expression *the GerE protein* dans les exemples ci-dessus. Si l'on considère des ontologies peuplées d'instances, le processus d'annotation consiste à créer des instances de concepts et à leur rattacher ces entités nommées. Si on interprète le texte au regard d'une ontologie classique (non peuplée), on néglige l'annotation des entités nommées ou on se contente de les associer à des concepts. Dans certains cas,

[1] Nous ne considérons pas ici de langage particulier pour la représentation des ontologies, mais dans la suite de l'article, les noms de concepts sont en majuscules (CONCEPT), les instances avec seulement une capitale à l'initiale (Instance), les rôles en minuscules (*a-pour-rôle*).

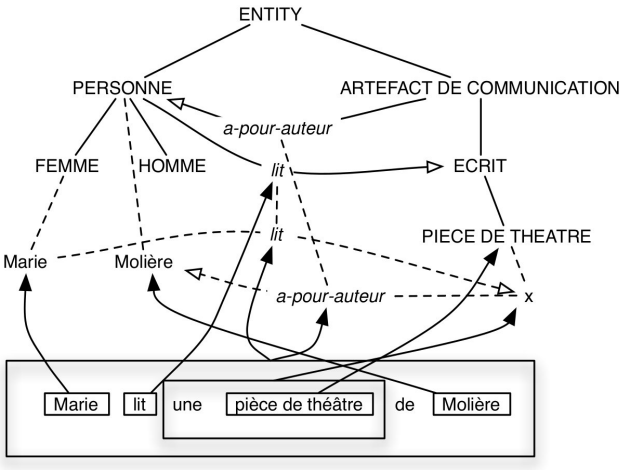


FIG. 1 – Example d'annotation de Marie lit une pièce de théâtre de Molière"

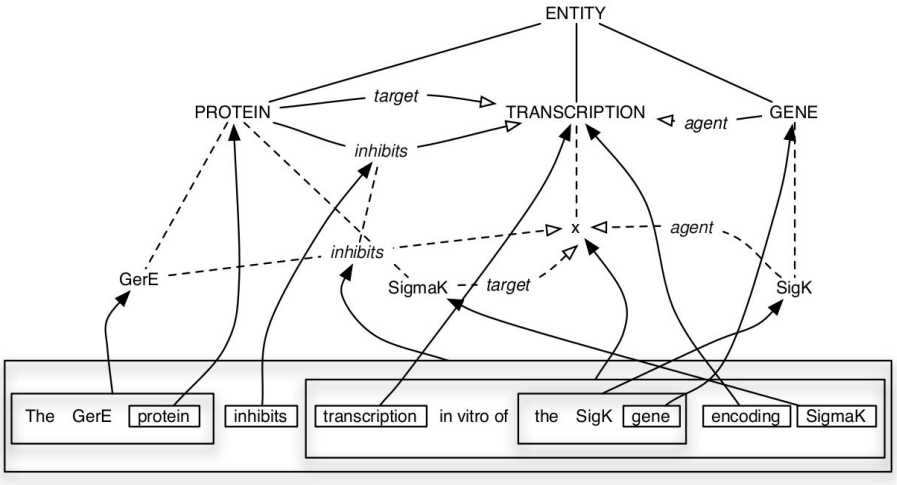


FIG. 2 – Example d'annotation de "The GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK
(La protéine GerE inhibe la transcription in vitro du gène sigK qui encode sigmaK)"

une référence est faite à une instance de concept sans que celle-ci soit nommée : dans l'exemple 1, la "pièce de théâtre de Molière" que "Marie lit" n'est pas mentionnée explicitement. Le processus d'annotation devrait être le même que lorsque l'entité est nommée mais en pratique on néglige ces "entités non nommées" parce que leur repérage est hors de portée des outils de traitement automatique des langues (TAL) courants.

Certains mots ou expressions dénotent des *concepts*. Ils constituent généralement le vocabulaire spécialisé du domaine considéré, *i.e.* la terminologie du domaine. Ces termes (par ex. *protein, transcription, pièce de théâtre*) sont souvent composés de plusieurs mots et ils sont importants à repérer parce qu'ils sont sémantiquement plus pertinents que les mots qui les composent. On a tendance à privilégier les termes nominaux mais des verbes ou groupes verbaux comme *encoding* peuvent aussi être associés à des concepts.

Certains mots ou expressions dénotent des *rôles conceptuels*. De même que les termes peuvent être rattachés à des concepts dans le processus d'annotation, ils peuvent être rattachés à des rôles si les notions sous-jacentes ont été modélisées sous la forme de rôle plutôt que comme concepts. C'est le cas de *lit* ou *inhibits* dans les exemples ci-dessus mais le fragment de texte annoté est souvent plus large, les rôles s'exprimant souvent par tournures de phrases qui ne se réduisent pas à un mot clef.

Certains fragments de texte renvoient à des *relations entre instances* : "une pièce de théâtre de Molière" ou "SigK gene encoding sigmaK" dans les exemples ci-dessus. C'est souvent un large fragment qui est annoté comme une relation entre instance. On s'y intéresse pour peupler les ontologies et on tend à les négliger lorsque l'interprétation se fait au regard d'une ontologie classique.

Certains fragments textuels, enfin, expriment des *axiomes ontologiques*. Par exemple, la phrase "Genes are biological entities" peut s'interpréter comme une relation de subsomption entre les concepts GENE et BIOLOGICAL ENTITY. Si on était capable d'analyser une phrase comme "les pièces de théâtre sont toujours écrites par quelqu'un", on pourrait de la même manière l'associer à un axiome exprimant une restriction de cardinalité du rôle *a-pour-auteur*.

Ces cinq types d'annotations ne sont généralement pas considérés tous ensemble. Selon les cas, on met l'accent sur la population des ontologies et donc sur les instances et leurs relations, sur l'information conceptuelle ou encore sur la découverte d'axiomes. Ils sont néanmoins intéressants à considérer ensemble pour appréhender l'annotation sémantique dans sa globalité.

2.2 Difficultés

Etablir les annotations précédentes pose une double difficulté.

La première difficulté concerne la segmentation du texte parce qu'il est souvent difficile d'identifier précisément les éléments textuels à annoter. C'est un problème connu en reconnaissance d'entités nommées : est-ce que le déterminant et le nom classifieur font partie de l'entité ou est-ce que seul le nom doit être annoté (*the GerE protein* vs. *GerE*) ? Il est souvent difficile de trancher. De manière plus générale, certaines connaissances ontologiques ne se traduisent pas par un simple mot ou expression. C'est souvent un fragment large ou une phrase complète qui véhicule l'information comme dans

le fragment "transcription of the SigK gene encoding sigmaK". Le fait de prendre ou non en compte ces annotations dépend largement de l'objectif visé.

L'ambiguïté inhérente à la langue soulève une seconde difficulté. Lorsque le fragment textuel est ambigu[2], il faut choisir l'élément ontologique avec lequel il doit être mis en correspondance. Le processus d'annotation suppose alors une étape de désambiguïsation qui repose généralement sur des indices figurant dans le contexte du fragment à annoter. Par exemple, la préposition *de* qui s'interprète comme dénotant le rôle *a-pour-auteur* dans "une pièce de théâtre de Molière", pourrait s'interpréter comme une localisation dans d'autres contextes. C'est le type des mots qui l'entourent et notamment le nom qu'elle introduit qui permet de désambiguïser la préposition.

Pour effectuer cette annotation sémantique en résolvant ces problèmes de segmentation et de désambiguïsation, il faut étendre l'ontologie avec des connaissances permettant de mettre en correspondance le texte à annoter et l'ontologie au regard de laquelle il est interprété.

2.3 Mise en oeuvre

Aujourd'hui, l'annotation linguistique des documents est généralement réalisée par des outils de TAL. On a souvent recours à des plates-formes qui permettent d'intégrer et d'exécuter, souvent séquentiellement, un certain nombre d'outils d'annotation existants qui ont été conçus dans des contextes et avec des objectifs différents. Parmi les plates-formes d'annotation linguistique, nous pouvons citer GATE (Cunningham, 2002), Atlas (Bird *et al.*, 2000) ou encore Ogmios (Hamon *et al.*, 2007). Une nouvelle tendance s'appuie sur l'environnement UIMA (Ferrucci & Lally, 2004).

Ces plates-formes prennent surtout en compte les problèmes d'interopérabilité verticale. Elles exploitent des mécanismes d'encapsulation des modules d'annotation, ce qui est une bonne solution si les annotations de ces outils sont fortement orthogonales.

Malheureusement, du point de vue de l'annotation sémantique, l'intersection entre les différents niveaux d'annotation n'est pas vide. Par exemple, dans le projet Alvis (Nazarenko *et al.*, 2006), l'annotation sémantique est en partie portée par le niveau de détection des entités nommées qui identifie ces entités et leur associe un premier type grossier, en partie par le niveau de détection des termes et en partie par un niveau spécifique d'annotation sémantique qui se limite à associer des catégories et des relations ontologiques aux éléments identifiés par les niveaux précédents. Ces trois niveaux sont donc fortement dépendants voire concurrents, ce qui pose d'importants problèmes de cohérence pour l'annotation sémantique.

Notre objectif ici est de résoudre ce problème d'interopérabilité horizontale non traité par les plates-formes d'annotation tout en tirant profit des solutions que ces dernières apportent au problème d'interopérabilité verticale. Il s'agit à la fois d'utiliser une ontologie pour rassembler en un lieu unique, cohérent et homogène ce qui est du ressort de la sémantique du document mais aussi d'étendre cette ontologie pour permettre son utilisation dans l'annotation sémantique, ce qui suppose de faire le lien entre le niveau conceptuel et le niveau linguistique qui sont par nature hétérogènes et non isomorphes.

[2]Le cas est fréquent : on sait que même dans les domaines spécialisés, les mots peuvent être ambigus.

Cette extension est généralement de nature lexicale (voir section 3) mais nous proposons une autre approche, à base de règles, qui préserve une articulation claire entre l'annotation linguistique (faite par les outils de TAL) et l'annotation sémantique qui est guidée par l'ontologie (section 4).

3 Les extensions lexicales

Il existe actuellement de multiples modèles pour représenter conjointement lexiques ou terminologies et ontologies. Dans cette optique, si OMV (*Ontology Metadata Vocabulary*) (Hartmann *et al.*, 2005) est une proposition de standardisation des méta-données descriptives d'une ontologie, LexOMV (Montiel-Ponsoda *et al.*, 2007) cherche à étendre le modèle OMV pour apporter des méta-données décrivant le niveau lexical des éléments (concepts ou propriétés) de l'ontologie. LexOMV n'est donc pas une extension lexicale mais permet de décrire, par des méta-données, l'extension lexicale d'une ontologie donnée, et donc de certains des formalismes que nous décrivons dans cette section. Dans un premier temps, nous nous intéressons aux formalismes, recommandés par le consortium W3C, que sont RDFS, SKOS et OWL. Ces trois formalismes sont finalement assez limités pour représenter correctement un niveau lexical riche. Nous décrivons donc ensuite quelques travaux qui cherchent à dépasser ces limitations avant d'avancer l'idée que l'annotation sémantique ne passe pas nécessairement par la représentation d'un niveau lexical riche au sein de l'ontologie.

RDFS est une recommandation du consortium W3C (W3C, 2009) qui permet de définir des étiquettes pour des classes via la propriété `rdfs:label`. Cette propriété peut être utilisée pour associer une information lexicale à une classe de l'ontologie. Mais le domaine de définition de `rdf:label` est le *Littéral*, ce qui limite l'expression d'informations lexicales complexes.

SKOS, actuellement en développement dans le cadre du consortium W3C, utilise les propriétés `skos:prefLabel` et `skos:altLabel` (sous-propriétés de `rdf:label`) et `xml:lang` pour associer des termes multilingues aux concepts. SKOS souffre des même limitations d'expression que RDFS pour rendre compte d'informations lexicales complexes. De plus, il faut noter que SKOS est conçu pour décrire une ressource conceptuelle et n'est pas adapté, contrairement à OWL, pour décrire la richesse structurelle des ontologies.

Toujours dans le cadre du consortium W3C, mais bien plus expressif que RDFS, OWL permet de formaliser des ontologies selon une syntaxe et une sémantique bien définies qui autorisent l'inférence. Même s'il se décline en trois sous-langages d'expressivité croissante (*OWL Light*, *OWL DL* et *OWL Full*), OWL n'offre pas plus de souplesse que RDFS pour la représentation du niveau lexical.

LingInfo (Buitelaar *et al.*, 2006), développé dans le cadre du projet SmartWeb, est un modèle de lexique basé sur une ontologie qui permet de représenter une terminologie multilingue. LingInfo associe les connaissances linguistiques aux classes (respectivement aux propriétés) de l'ontologie en définissant des meta-classes (resp. des meta-propriétés). La définition d'une meta-classe/propriété consiste en la donnée d'un terme, d'une langue et d'une décomposition morpho-syntaxique constituée d'un ensemble d'un ou plusieurs mots/syntagmes. Un syntagme est un mot/syntagme qui modélise

récursivement un syntagme nominal ou verbal complexe. Un mot est un mot/syntagme qui représente la structure morphologique de termes complexes comme *joueur de football*. A un mot peuvent être associées différentes propriétés comme le nombre ou le genre. Cette décomposition s'étend jusqu'à l'objet racine (*stem*) et permet donc d'associer aux éléments de l'ontologie une représentation lexicale très détaillée.

Szulman *et al.* (2008) proposent, avec le logiciel Terminae, une méthodologie de construction d'ontologies à partir de corpus. Le lien entre le corpus et les concepts est assuré par la construction d'une terminologie du domaine où chaque terme se traduit par une fiche terminologique. Terminae propose un export OWL et contourne les limitations de l'utilisation de la seule propriété `rdfs:label` par l'utilisation de la structure de propriété d'annotation (`owl:AnnotationProperty`) pour représenter les informations terminologiques associées à une classe.

Pour pallier la faiblesse d'OWL pour représenter le niveau lexical, (Reymonet *et al.*, 2007) et (Cimiano *et al.*, 2007) proposent une solution originale consistant à représenter le niveau lexical en utilisant toute l'expressivité d'OWL. Les termes sont ainsi réifiés sous forme de `owl:Class`. (Reymonet *et al.*, 2007) utilise un niveau d'abstraction supérieur pour distinguer les concepts des termes. Le lien entre concept et terme se fait par une propriété *denote* orientée du terme vers le concept. LexOnto (Cimiano *et al.*, 2007) va beaucoup plus loin et utilise une meta-ontologie pour mettre en relation le niveau lexical et le niveau ontologique. Ce formalisme permet de représenter des relations simples entre termes et concepts, mais également des structures linguistiques de type prédicat-argument comme les cadres de sous-catégorisation.

Les travaux décrits dans cette section présupposent la nécessité d'un niveau lexical ou terminologique pivot, entre la ressource ontologique et le texte. Ce niveau lexical pose deux problèmes importants, le premier est celui de sa représentation et le second celui de sa constitution. D'ailleurs, dans la plupart des travaux présentés ci-dessus, la constitution du niveau lexical est faite manuellement par un utilisateur assisté par un environnement logiciel et méthodologique adéquat. Pourtant, dans le cadre d'une plate-forme d'annotation linguistique, le niveau lexical existe déjà en grande partie. Le problème n'est pas tant son absence que la pluralité, l'hétérogénéité et l'incohérence des différents niveaux qui le supportent. Ainsi, plutôt que de reconstituer un nouveau niveau de représentation lexicale, nous pensons que notre extension d'ontologie doit plutôt s'appuyer sur les niveaux de représentation lexicale existants.

4 Une extension à base de règles d'annotation

Nous proposons de distinguer plus clairement les niveaux ontologique et lexical (ou linguistique) ainsi que le processus d'annotation sémantique des autres étapes d'annotation qui associent au texte des métadonnées linguistiques. Cela revient à établir une frontière claire et opératoire, même si elle est pour partie artificielle[3], entre l'interprétation du texte et son analyse linguistique.

En pratique, le processus d'annotation d'une plate-forme d'annotation linguistique se compose d'une série de modules d'annotation, chacun s'appuyant sur des ressources

[3]L'interprétation s'appuie sur l'analyse et celle-ci opère des choix qui sont en partie guidés par celle-là.

particulières (lexique, terminologies, etc.) pour ajouter une couche particulière d'annotations à partir des annotations des modules précédents, parfois en corrigeant ces dernières. L'ontologie étendue est exploitée comme une ressource exploitée par un composant d'annotation sémantique. L'extension de l'ontologie doit s'appuyer au maximum sur les niveaux inférieurs de la plate-forme d'annotation pour rester la plus simple possible et bénéficier du travail déjà réalisé par les composants inférieurs.

Cette extension doit être tout à la fois 1) compréhensible et modifiable par un être humain pour autoriser d'inévitables interventions manuelles sur une ontologie étendue, 2) interprétable par un ordinateur pour permettre l'annotation sémantique automatique dans le cadre d'une plate-forme d'annotation et 3) inférable par un ordinateur pour pouvoir être apprise (semi-)automatiquement à partir d'un corpus sémantiquement annoté selon l'ontologie à étendre. Il nous apparaît que la meilleure solution est l'utilisation de règles dont les prémisses sont constituées d'un ensemble de contraintes qu'un fragment de texte doit satisfaire pour être annoté par l'élément ontologique figurant dans la conclusion. L'application de la règle déclenche l'annotation de tous les fragments du texte qui satisfont les contraintes de la prémisse.

4.1 Définition de l'extension

Soit $O = \langle C, R, I?, RI?, A\rangle$ une ontologie composée d'un ensemble de concepts (C), de rôles (R), d'instances (I), de relations entre instances (RI) et d'axiomes (A)[4] et $\mathcal{R} = \langle \mathcal{R}_C, \mathcal{R}_R, \mathcal{R}_I, \mathcal{R}_{RI}, \mathcal{R}_A\rangle$, un ensemble de règles permettant d'annoter des fragments de textes en les reliant à des concepts ($\mathcal{R}_C$), des rôles ($\mathcal{R}_R$), des instances ($\mathcal{R}_I$), des relations entre instances ($\mathcal{R}_{RI}$) ou des axiomes ($\mathcal{R}_A$). Une règle est la donnée d'un couple (P, C) où P (*Prémisse*) décrit les conditions qu'un segment de texte doit vérifier pour être annoté et C (*Conclusion*) indique comment annoter le segment. Nous disons qu'une ontologie O^R est étendue ssi :

- pour chaque concept c de C il existe un couple de règles (ρ_c, ρ_{ci}) concluant sur c et telles que $\rho_c \in \mathcal{R}_C$ et $\rho_{ci} \in \mathcal{R}_I$;
- pour chaque rôle r de R il existe un couple de règles (ρ_r, ρ_{ri}) concluant sur r et telles que $\rho_i \in \mathcal{R}_I$ et $\rho_{ri} \in \mathcal{R}_{RI}$;
- pour chaque axiome a de A il existe une règle $\rho_a \in \mathcal{R}_A$ concluant sur a.

4.2 Types de règles

Comme indiqué ci-dessus, les règles sont différenciées selon le type de l'élément ontologique qui figure en conclusion de la règle mais il faut surtout distinguer deux types de règles :

- Les règles d'annotation ontologique à proprement parler ($\mathcal{R}_C$, $\mathcal{R}_R$, $\mathcal{R}_A$)concluent sur un concept, un rôle ou un axiome. Elles visent à identifier en corpus des fragments de texte qui dénotent des concepts, rôles ou axiomes et à les annoter en conséquence ;
- Les règles de peuplement ontologique ($\mathcal{R}_I$, $\mathcal{R}_{RI}$) concluent également sur des concepts ou des rôles mais elles visent à identifier des fragments de texte qui ren-

[4] Les éléments notés ? sont optionnels : ils n'apparaissent que dans les ontologies peuplées.

voient à des instances de ces concepts ou de ces rôles. Le fragment de texte est annoté comme une instance de concept ou de rôle et cette instance est ajoutée à l'ontologie sous le concept ou le rôle qui figure en conclusion de la règle.

Dans l'exemple de la figure 2, le segment *protein* est annoté par une règle du type ρ_c qui conclut sur le concept PROTEIN, tandis que le segment *The GerE protein* est annoté par une règle du type ρ_{ci} qui implique la création d'une instance GerE rattachée au concept PROTEIN.

Si l'objectif n'est pas de peupler l'ontologie, les règles de peuplement sont soit ignorées soit interprétées comme des règles conceptuelles. Dans ce cas, le fragment de texte qui renvoie à une instance est annoté comme le concept père de cette instance. Dans l'exemple de la figure 1, cela revient à annoter *Marie* non pas comme l'instance Marie mais comme le concept PERSONNE.

Concernant l'application des règles dédiées à l'identification d'occurrences d'instances, nous supposons que le nombre d'instances à créer correspond au nombre d'occurrences d'instances identifiées dans le texte. Le fait que deux occurrences dans le texte renvoient à un même individu est un problème de résolution de coréférence ou d'anaphore. Nous considérons que ce n'est pas du ressort de l'extension de l'ontologie de résoudre ce problème, qui doit trouver une solution à un autre niveau d'annotation de la plate-forme. Le fait de déléguer ce problème permet par ailleurs de centraliser les règles de peuplement au niveau du concept et de gagner en généricité. Dans le cas contraire, les règles devraient être réparties, spécialisées et attachées aux instances qui, contrairement aux concepts, ne sont pas des objets préexistants dans l'ontologie.

Ces extensions, associées aux concepts et aux rôles, ne sont pas des propriétés au sens où elles ne s'héritent pas. Le fait que le concept B hérite de A n'implique pas que les règles qui permettent d'identifier des fragments dénotant A puissent être utilisés pour identifier des fragments dénotant B.

4.3 Expression des règles

La prémisse d'une règle peut être représentée par un ensemble de patrons qui s'appliquent sur un corpus. S'il a été préalablement analysé par certains modules d'annotation (étiquetage morpho-syntaxique, reconnaissance d'entités nommées, étiquetage terminologique, par exemple) celui-ci porte déjà des annotations linguistiques. Un patron est une expression qui s'appuie sur ces différents niveaux d'annotations. L'application d'un patron sur un corpus est une opération qui retourne un ensemble de segments du corpus à annoter selon la conclusion de la règle.

A titre d'illustration, voici trois exemples distincts de patrons, écrits dans un pseudo langage pour en faciliter la compréhension, pour repérer dans le texte des occurrences du concept informatique *Système d'exploitation* :

1. `[string={système|d|'|exploitation}]`
2. `[lemme={système|de|exploitation}]`
3. `[terme={système d'exploitation}]`

string correspond à la forme brute du texte, *lemme* à la forme lemmatisée des mots et *terme* aux annotations de l'extracteur de termes. Ces trois patrons montrent l'intérêt de

l'utilisation des différents types d'annotations de la plate-forme. En effet, le premier patron n'est pas générique et ne peut pas reconnaître de simples variations comme *Système d'exploitation* ou *systèmes d'exploitation*. Le second est plus générique car insensible à la casse et au nombre. Le dernier est encore plus générique car, selon l'extracteur de termes utilisé, il peut reconnaître des chaînes comme *OS* pour lesquelles l'extracteur proposera la forme canonique *système d'exploitation*.

L'expression du patron ne peut pas toujours se réduire à la délimitation de la portion de corpus à annoter : il faut souvent exprimer des contraintes de désambiguïsation portant sur son contexte, comme pour les occurrences de *détention* qui sont, selon les cas, à annoter par un concept $\mathtt{D}_1$ correspondant au *fait d'être incarcéré ou enfermé*, ou par un concept $\mathtt{D}_2$ correspondant au *fait d'avoir en sa possession*. Pour englober ce cas de figure, un patron devrait plutôt s'écrire sous la forme d'un triplet $(LC?, T, RC?)$ où $LC?$ et $RC?$ sont des expressions facultatives pour contraindre le contexte gauche et droit, et où T (*Target*) est l'expression permettant d'identifier la portion de corpus qui doit supporter l'annotation.

Le langage utilisé pour l'écriture des patrons dépend de la façon dont le corpus et les annotations sont représentées. Dans le cas d'un document XML, un patron peut être la donnée d'un triplet d'expressions XPath ou d'une requête XQuery, l'expressivité de XQuery permettant de se passer de cette notion de triplet. Si les annotations et le corpus sont des objets Java persistants, les patrons peuvent, par exemple, se traduire par des requêtes JDOQL dans le cas d'une persistance gérée par JDO, ou des requêtes HQL dans le cas d'une persistance gérée par Hibernate.

4.4 Opérations sur une ontologie étendue

Étendre les ontologies suppose de redéfinir les opérations de mise à jour de l'ontologie, car toute opération effectuée sur l'ontologie a des répercussions sur son extension :

- La fusion de deux concepts est une opération simple. Soit c_1 et c_2 respectivement étendus par les couples de règles (ρ_{c1}, ρ_{i1}) et (ρ_{c2}, ρ_{i2}). On peut poser que le concept c résultant de la fusion de c_1 et c_2 a comme extension le couple de règles $(\rho_{c1} \vee \rho_{c2}, \rho_{i1} \vee \rho_{i2})$.
- La décomposition d'un concept impose la décomposition de son extension. Deux scénarios peuvent être envisagés. Le premier consiste à réaliser la décomposition de l'extension à la main par l'édition des règles associées à ce concept. Le second consiste à annoter un corpus représentatif avec le concept à décomposer en appliquant les règles qui lui sont attachées. L'utilisateur doit ensuite répartir les annotations sur les différents concepts qui remplacent le concept à décomposer. Le système peut ensuite reconstituer les extensions des différents concepts en inférant les patrons des règles à partir du corpus annoté.
- La suppression d'un concept peut se traduire par l'absorption de son extension par le ou les concepts qui le subsument, ce qui ramène au cas de la fusion. L'extension peut également être supprimée avec le concept : les fragments annotés cessent de l'être. Elle peut enfin être décomposée comme dans le cas précédent.
- L'ajout d'un nouveau concept implique de préciser son extension. Là encore deux scénarios sont envisageables : soit la proposition spontanée par l'utilisateur des

patrons associés, soit l'inférence de ces patrons à partir d'annotations manuelles des occurrences du concept créé dans un corpus représentatif.

5 Conclusion et perspectives

Cet article propose une solution au problème de l'annotation sémantique d'un texte au regard d'une ontologie. Nous nous interrogeons sur la manière dont des fragments de texte peuvent être mis en correspondance avec les éléments d'une ontologie. Nous avons proposé d'étendre les ontologies par des règles d'annotation sémantique plutôt que par l'ajout de (méta-)propriétés lexicales. Cette solution permet de tirer le meilleur parti des outils de TAL qui produisent chacun à leur niveau des annotations linguistiques. Elle a également le mérite de distinguer clairement le processus d'analyse et la tâche d'interprétation, qui seule met l'ontologie en jeu.

Étendre l'ontologie par des règles d'annotation qui s'expriment sous la forme de patrons présente, de notre point de vue, de nombreux intérêts. Tout d'abord, le pouvoir expressif des patrons est très grand. Il va de la simple représentation d'une liste de mots à des expressions complexes basées sur des annotations de haut niveaux (entités nommées, termes) et autorise l'expression de règles de flexion ou de désambiguïsation. Ensuite, les patrons sont compréhensibles et modifiables par une personne tout en étant interprétables, voire calculables, par un ordinateur, ce qui permet d'envisager leur acquisition automatique à partir d'un corpus annoté. Enfin, ils peuvent s'exprimer dans de nombreux formalismes largement connus voire standardisés comme les expressions régulières ou les chemins Xpath.

Ce travail appelle un double prolongement. Il s'agit tout d'abord de tester l'approche proposée en intégrant un module d'annotation sémantique dans une chaîne d'annotation existante. Nous projetons de le faire dans la plate-forme Ogmios et des expériences d'annotation sémantique doivent se faire dans le cadre du programme Quaero. La question de l'acquisition des règles est également une piste à explorer. Il est d'autant plus important de pouvoir apprendre (semi-)automatiquement les règles qu'elles varient d'une ontologie à l'autre mais aussi d'une plate-forme d'annotation à l'autre.

Références

AMARDEILH F., LAUBLET P. & MINEL J. L. (2005). Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques. In *Actes des 16èmes journées francophones d'Ingénierie des Connaissances*, p. 25–36.

BIRD S., DAY D., GAROFOLO J. S., HENDERSON J., LAPRUN C. & LIBERMAN M. (2000). Atlas : A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, p. 1699–1706.

BUITELAAR P., SINTEK M. & KIESEL M. (2006). A multilingual/multimedia lexicon model for ontologies. In *ESWC*, p. 502–513.

CIMIANO P., HAASE P., HEROLD M., MANTEL M. & BUITELAAR P. (2007). Lexonto : A model for ontology lexicons for ontology-based nlp. In *Proceedings*

of OntoLex - From Text to Knowledge : The Lexicon/Ontology Interface (workshop at the International Semantic Web Conference).

CIRAVEGNA F. (2000). Learning to tag for information extraction from text. In *Proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction*.

CUNNINGHAM H. (2002). Gate, a general architecture for text engineering. In S. NETHERLANDS, Ed., *Computers and the Humanities*, volume 36, p. 223–254.

DILL S., EIRON N., GIBSON D., GRUHL D., GUHA R., JHINGRAN A., KANUNGO T., MCCURLEY K. S., RAJAGOPALAN S., TOMKINS A., TOMLIN J. A. & ZIEN J. Y. (2003). A case for automated large scale semantic annotations. *Journal of Web Semantics*, **1**, 115–132.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de linguistique . Université de Paris VII.

FERRUCCI D. & LALLY A. (2004). UIMA : an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, **10**(3-4), 327–348.

GRUBER T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199–220.

HAMON T., DERIVIÈRE J. & NAZARENKO A. (2007). Ogmios : a scalable nlp platform for annotating large web document collections. In *Procedings of Corpus Linguistics 2007*, Birmingham, UK.

HANDSCHUH S. (2002). S-cream - semi-automatic creation of metadata. In *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, p. 358–372 : Springer Verlag.

HARTMANN J., SURE Y., HAASE P., PALMA R. & DEL CARMEN SUÁREZ-FIGUEROA M. (2005). OMV – Ontology Metadata Vocabulary. In C. WELTY, Ed., *ISWC 2005 - In Ontology Patterns for the Semantic Web*.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

KOGUT P. (2001). Aerodaml : Applying information extraction to generate daml annotations from web pages. In *First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation*.

MONTIEL-PONSODA E., DE CEA G. A., SUAREZ-FIGUEROA M., PALMA R., PETERS W. & GOMEZ-PEREZ A. (2007). Lexomv : an omv extension to capture multilinguality. In *n Proceedings of the OntoLex07*, p. pp. 118–127.

NAZARENKO A., NÉDELLEC C., ALPHONSE E., AUBIN S., HAMON T. & MANINE A.-P. (2006). Semantic annotation in the alvis project. In *Proceeding of IIIA-2006 : International Workshop on Intelligent Information Access*, Helsinki, Finland.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de ressources termino-ontologiques en OWL. In F. TRICHET, Ed., *Journées Francophones d'Ingénierie des Connaissances (IC), Grenoble*, p. 169–180.

SZULMAN S., AUSSENAC-GILLES N. & DESPRES S. (2008). The Terminae Method and Platform for Ontology Engineering from Texts. In *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, p. paru. IOS press.

W3C (2009). World wide web consortium. `http://www.w3.org/`.

Evaluation d'associations sémantiques dans une ontologie de domaine

Thabet Slimani[1], Boutheina Ben Yaghlane[2], Khaled Mellouli[2]

ISG de Tunis BP 41 - Rue de la liberté- Bardo 2000
IHEC Carthage, Carthage Présidence 2016, Tunisia
thabet.slimani@issatm.rnu.tn
{boutheina.yaghlane, khaled.mellouli}@ihec.rnu.tn

Résumé : Dans une ontologie de domaine, une association sémantique entre deux entités (concepts, attributs d'un concept) est une représentation d'un chemin ou d'un lien sémantique (LS) indirect entre elles. Un défi prometteur pour le Web sémantique est de développer des méthodes pour découvrir des données fortement liées dans un nombre important d'associations sémantiques rassemblées à partir des sources disparates. Dans ce contexte, cet article présente, en premier lieu, le degré d'un lien sémantique (DLS) pour mesurer une relation directe entre deux entités, et en deuxième lieu le degré d'une association sémantique (DAS) pour mesurer des associations sémantiques extraites à partir d'une ontologie de domaine. Les résultats expérimentaux montrent l'avantage des méthodes proposées et démontrent leur efficacité prometteuse.

Mots-clés : association sémantique, degré d'un lien sémantique, degré d'une association sémantique, ontologie de domaine, concept.

1 Introduction

Aujourd'hui, la croissance massive de données stockées dans le Web est contrôlée par un réseau des sources de données en corrélation, appelé réseau sémantique des liens (RSL), incluant des personnes, des compagnies, des connaissances de domaine, des publications scientifiques, des articles, etc. Un RSL est conçu pour établir des rapports sémantiques parmi diverses ressources dans le Web visant à prolonger le réseau WWW des liens hypertextes à un réseau sémantique riche (Zhuge, 2007). Formellement, un RSL est un réseau contenant des noeuds sémantiques et des liens sémantiques. Un noeud sémantique peut être un concept, un attribut de concept, un schéma d'ensemble de données, un URL, une entité, etc. Les liens entre les noeuds sémantiques (entités) dans une base de connaissance RDF fournissent le type, la signification ou l'interprétation des entités.

La découverte des liens sémantiques est un problème important pour des applications appréhendant des données gérées en réseau. Un lien sémantique (LS) se rapporte à une relation directe entre deux entités. Par contre, une association sémantique (AS) est un chemin connectant deux entités d'une manière indirecte.

L'extraction d'une association sémantique (AS) signifie l'extraction d'un chemin entre deux entités reliées indirectement par des relations concrètes (propriétés) contenues dans un graphe RDF (Aleman-Meza *et al.*, 2003) (Anyanwu *et al.*, 2005) (Aleman-Meza *et al.*, 2006) (Ning *et al.*, 2006). Une AS a un sens différent par rapport à la similarité sémantique. Une similarité sémantique permet de mesurer le degré de proximité de deux entités du point de vue sémantique, alors qu'une AS permet de mesurer le degré de connectivité de deux entités du point de vue sémantique. Par exemple, l'entité "author" a une valeur faible de similarité sémantique avec l'entité "Research-Area" parce qu'elles ne se réfèrent pas au même objet, alors que "author" possède une association sémantique forte avec l'entité "Research-Area".

Dans des applications analytiques telles que la sécurité nationale, la bioinformatique, etc, il est indispensable d'extraire des connaissances significatives à partir des liens et des associations sémantiques déjà extraits. Malheureusement, dans les travaux recensés dans la littérature, les avances remarquables aux niveau des développements actuels montrent un manque de satisfaction pour de nombreuses applications récentes telles que l'exploitation de données, la recherche documentaire, l'extraction des services Web qui exigent l'évaluation des associations sémantiques.

Dans ce papier, nous proposons certaines mesures qui combinent des fonctions sémantiques et statistiques pour mesurer les liens sémantiques directs entre deux entités et les associations sémantiques entre deux entités reliées à travers un chemin en évaluant le rapport entre elles.

Le reste de ce document est organisé comme suit. La section 2 présente les travaux liés et nos contributions principales. La section 3 décrit les spécifications d'un lien sémantique et d'une association sémantique. Les expérimentations réalisées sont présentées dans la section 4. Finalement, la section 5 donne un résumé et une perspective des travaux futurs.

2 Travaux liés

Dans le travail de (Aleman-Meza *et al.*, 2006), les auteurs proposent une approche pour la découverte d'une diversté de liens sémantiques entre les "reviewers" et les "authors" dans une ontologie complète pour déterminer un degré de conflit d'intérêt. La création de cette ontologie a été basée sur les entités et l'intégration des relations de deux réseaux sociaux : réseau foaf (friend-of-a-friend) et réseau "co-author". Dans cette même perspective, (Cao *et al.*, 2005) présentent une approche pour la fouille des communautés cachées par l'exploitation des réseaux sociaux hétérogènes. Peterson *et al.* discutent certaines approches pour exploiter le capital social pour créer un standard réseau sémantique riche (Peterson *et al.*, 2008). En plus, une autre approche qui discute la question de la désambigüisation d'URI dans le cadre des données liées est présentée dans le travail (Jaffri *et al.*, 2008).

Notre travail se situe dans le cadre des travaux permettant d'évaluer les liens et les associations sémantiques. Nous pouvons classifier les approches, dans la littérature, dans deux directions : (1) les approches orientées données (data-driven) qui essayent de capturer la dépendance entre les concepts (dérivés du corpus) par l'information statistique (Cao *et al.*, 2005). Cette approche utilise les co-occurrences des relations binaires

entre les concepts et (2) les approches orientées structures (structure-driven) (Watabe & Kawaoka, 2001) qui exploitent les caractéristiques de la structure d'une ontologie (les classes/concepts d'une ontologie et leurs relations).

Notre approche dérive des approches statistiques basées sur un modèle de langage statistique, puisque les mesures proposées sont basées sur la combinaison des formules de probabilité. Les approches statistiques peuvent être appliquées aux applications dans le domaine de recherche documentaire, de la reconnaissance de forme et la fouille de données. Ces approches utilisent la distribution de probabilité pour mesurer des données liées. Par exemple, dans le domaine de recherche documentaire, la pertinence d'un document avec une requête peut être évaluée par la probabilité de la génération du document vis-à-vis à une requête donnée (Song & Croft, 1999). D'une manière analogue, dans la base de connaissance, deux entités connexes (Classe/Concept, Instance de Classe) peuvent être évaluées, par l'intermédiaire des liens sémantiques, en utilisant des mesures statistiques. Dans le travail de (Tian *et al.*, 2007), les auteurs proposent une approche qui mesure les associations sémantiques dans une ontologie de domaine est bien lié à notre travail. La différence par rapport à notre travail réside au niveau des formules du degré d'un lien sémantique et au niveau de l'évaluation des associations sémantiques.

3 Spécification d'un lien et d'une association sémantique

3.1 Base de connaissance

Une ontologie O est une structure incluant deux ensembles disjoints C et R, dont les éléments s'appellent, respectivement, les classes/concepts et les relations (propriétés/attributs). Au niveau de la partie supérieure de la figure 1, l'ensemble de classes C est défini par : {"Professor", "University", "Course", "Project", "Publication", "Student"}. L'ensemble des relations R est défini par {"Author-Of", "DegreeFrom", "Offers", "Related-to", "Enrolled-In"}. Une base de connaissance est une structure de deux ensembles disjoints incluant les instances de classes et les instances des liens sémantiques (propriétés). Au niveau de la partie inférieure de la figure 1, l'ensemble des instances de classes est défini par : {"Author-Of", "DegreeFrom", "Offers", "Related-to"}. La relation entre l'entité "P0" de type "Professor" et l'entité "PR0" de type "Project", présentée dans la figure 1, constitue une association sémantique (AS) définie par l'expression suivante : $\text{P0} \xrightarrow{DegreeFrom} \text{U0} \xrightarrow{Offers} \text{C0} \xrightarrow{Related-to} \text{PR0}$.

3.2 Signature et degré d'un lien sémantique

Cette section présente des formules qui calculent le degré d'un lien sémantique représenté dans une ontologie, ou des éléments (ressources) représentés dans un schéma. Ces formules de calcul exploitent le fait que les entités (concepts/classes) qui sont comparées peuvent avoir des propriétés (sous format d'attributs) associées entre elles et qui prennent en considération le niveau de généralité (ou de spécificité) de chaque entité dans l'ontologie aussi bien que leurs rapports avec d'autres entités ou concepts.

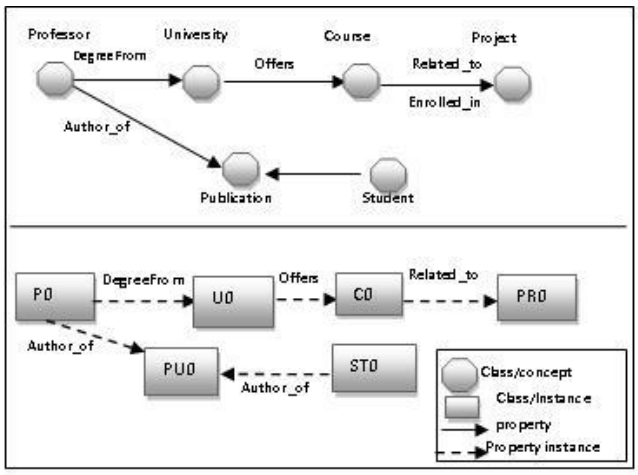


FIG. 1 – Un exemple de base de connaissance décrivant un réseau social.

La signature d'un lien sémantique entre deux entités est définie par les paires de concept/classe source et concept/classe cible. A titre d'exemple, le lien sémantique "DegreeFrom" admet comme signature l'ensemble (Professor, University) ; le lien sémantique "AuthorOf" admet deux signatures : (Student, Publication) et (Professor, Publication). La signature d'un lien sémantique n'est pas symétrique puisqu'il représente une propriété RDF qui se lit à partir d'un seul sens. Les associations sémantiques constituent un chemin qui combine, l'entité source, l'entité cible, les entités intermédiaires (instances de classe/concept) et les propriétés (instances de propriétés). Dans l'exemple de la figure 1, l'association sémantique entre l'entité source "P0" et l'entité cible "PR0" inclut deux entités intermédiaires {"U0", "C0"} et trois propriétés intermédiaires {"DegreeFrom", "Offers", "Related-to"}.

Soit A et B deux concepts/classes d'une ontologie de domaine, nous présentons les définitions suivantes :

- *Lien hiérarchique (LH)* : Si A est défini comme un "SuperClassof" de B, alors B est contenu dans A. Ce lien est représenté par l'expression A$\supset$B.
- *Lien d'équivalence (LE)* : Si A est défini comme "EquivalentClassof" de B, alors A est équivalent à B. Ce lien est représenté par l'expression A$\equiv$B.
- *Lien sémantique (LS)* : Si A est défini comme "PropertyOf" de B ou B est défini comme "PropertyOf" de A, on dit que A et B possèdent un lien sémantique. Ce lien est indiqué par l'expression A$\propto$B ou B$\propto$A.
- *Degré d'un lien sémantique (DLS) :* S'il y a un lien sémantique entre deux concepts/ classes A et B, nous définissons DLS comme le poids d'un lien sémantique, qui mesure le lien entre A et B. Par conséquent, nous pouvons définir l'équation (1) qui calcule $DLS_A(B, l)$, où l représente le lien qui relie les concepts A et B.

$$DLS_A(B, l) = P_A(B|l), l \in \{A \supset B, A \propto B\} \quad (1)$$

La probabilité conditionnelle $P_A(B|l)$ de A et de B avec le lien *l* définie dans l'équation (1) peut être décrite par l'équation (2) comme suit :

$$P_A(B|l) = P_A(B,l)/P_A(l) \tag{2}$$

Où $P_A(l)$ est la probabilité de A ayant *l* comme lien sémantique et $P_A(B,l)$ la probabilité d'apparition de A et de B avec le lien *l*.

En respectant les définitions du langage OWL[1], un concept/classe d'ontologie A peut avoir quelques instances (termes, ressources RDF). En conséquence, $P_A(B,l)$ devrait prendre en compte ces instances. Cependant, l'expression $P_A(B,l)$ dans l'équation (2) est enrichie par l'écriture de l'équation (3) :

$$P_A(B|l) = \sum_{A_i,B_j} P_{A_i}(B_j,l)/(P(B).P(l|B)) \tag{3}$$

Où A_i et B_j sont, respectivement, les instances des concepts A et B. ***P(l\|B)*** signifie la probabilité conditionnelle du lien *l* étant donnée l'entité *B* et ***P(B)*** désigne la probabilité d'apparition de l'entité B. Ensuite, la mesure $DLS_A(B,l)$ est changée pour estimer les 3 expressions : $P_{A_i}(B_j,l)$, ***P(B)*** et ***P(l\|B)***. L'évaluation de ces expressions sera basée sur l'estimation du maximum de vraisemblance. Cette évaluation prend en compte les co-occurrences de A et B avec *l* dans le corpus basé sur une ontologie de domaine.

3.2.1 Estimation de P$_{A_i}(B_j,l)$

Soit T l'ensemble de termes dérivés du corpus basé sur l'ontologie de domaine, et t le terme ayant un lien *l* avec A et/ou B. $P_{A_i}(B_j,l)$ désigne la probabilité d'apparition du concept A et B ensemble avec le lien *l*. Selon le type de *l*, $P_{A_i}(B_j,l)$ doit être prise en compte différemment. Si *l* est un lien sémantique (LS), il faut imposer une portion de texte qui ne dépasse pas une certaine limite (TL : Limite du texte en nombre de mots) et dans laquelle nous pouvons calculer les co-occurrences de A et B. L'estimateur de ce type de lien est obtenu par l'équation (4) :

$$\hat{\mathrm{E}}(P_{A_i}(B_j,l)) = \frac{count_{A_i}(B_j|TL)}{\sum_{t1\in(A_i,B_j)^t\in T} count_t(t1,TL) - count_{A_i}(B_j|TL)} \tag{4}$$

Si *l* est un lien hiérarchique (LH), nous définissons le modèle de co-occurrence TLS comme caractéristique lexique-syntaxique du lien sémantique se produisant dans le texte TL. Par exemple, l'expression "A inclus dans B" dans une portion de texte TL donne une indication d'un lien LH, qui doit être inclus dans la texte de TLS. L'estimateur de ce type de lien est obtenu par l'équation (5) :

$$\hat{\mathrm{E}}(P_{A_i}(B_j,l)) = \frac{count_{A_i}(B_j|TLS)}{\sum_{t1\in(A_i,B_j)^t\in T} count_t(t1,TLS) - count_{A_i}(B_j|TLS)} \tag{5}$$

[1]OWL Reference. http ://www.w3.org/TR/owl-ref/

3.2.2 Estimation de *P(B)*

La distribution de probabilité du concept B possède deux interprétations. De point de vue structure de l'ontologie, l'estimateur de P(B) peut être représenté par l'équation (6), où $|l_B|$ est le nombre des liens de B, c est un concept dans l'ontologie de domaine O, et $|l_c|$ est le nombre de liens du concept c. De point de vue occurrence de termes, P(B) est estimée par l'équation (7). Pour un concept c $\in$ O, fq(c)=$\sum_{c_i} count(c_i, C)$, où C est le corpus donné, c_i est le terme instance du concept c, alors *fq(c)* désigne la fréquence de l'apparition des instances du concept c dans le corpus C et fq_B désigne la fréquence de l'apparition du concept B.

$$P_st(B) = \frac{|l_B|}{Max_{c\in O}|l_c|} \tag{6}$$

$$P_ot(B) = \frac{|fq_B|}{Max_{c\in O}fq(c)} \tag{7}$$

L'estimateur de *P(B)* est représenté par le modèle mixte donné dans l'équation (8), dont λ est un coefficient permettant de combiner la structure de l'ontologie avec le corpus. Ce modèle mixte devient $P_ot(B)$ si λ=0 et $P_st(B)$ si λ=1. Ce coefficient est adopté pour optimiser la performance de recherche de l'information.

$$\hat{\mathrm{E}}(P(B)) \quad = \quad \lambda.P_st(B) + (1-\lambda).P_ot(B), 0 \leq \lambda \leq 1 \tag{8}$$

TL, TLS, et λ peuvent être différents lorsque les ontologies varient d'un domaine à un autre. Nous avons attribué une valeur constante, par intuition, comme valeur pour λ, ce qui ne pourrait pas être la meilleure valeur. Pour TL, nous avons obtenu sa valeur en analysant manuellement le corpus.

3.2.3 Estimation de *P(l|B)*

L'équation (9) donne l'estimation de *P(l|B)*, où $|l_B|$ est le nombre des liens B, et $|l_B|r$ le nombre des relations r ayant des liens avec B :

$$\hat{\mathrm{E}}(P(l|B)) = \frac{|l_B|r}{|l_B|} \tag{9}$$

Pour illustrer la méthode ci-dessus, nous énumérons différents types de liens dans l'ontologie MeSH (M*edical Subject Heading*)[2] comme présenté dans l'exemple du tableau 1.

[2]http ://www.nlm.nih.gov/mesh/

LS	Exemple dérivé de MeSH	Représentation graphique
A$\propto$B	L=Formalin Test$\propto${Pain, Intractable}	$A \xrightarrow{L}$B
A$\supset$B	Headache $\supset$Pain	$A \xleftarrow{is-a}$B
A$\equiv$B	Pain$\equiv$Postoperative	$A \leftrightarrow B$

TAB. 1 – Exemples de liens sémantiques extraits à partir de l'ontologie MeSH

3.3 Signature et degré d'une association sémantique

Une AS a trois types de signatures définis comme suit :

- *Signature des propriétés d'une association sémantique (SPAS)* définie par les propriétés contenues dans l'association sémantique. Dans l'exemple de la figure 1, la signature des propriétés d'une AS qui mène de l'objet source "P0" à l'objet cible "PR0" est définie par l'ensemble de propriétés/liens {"DegreeFrom", "Offers", "Related-to"}.
- *Signature d'instances d'une association sémantique (SIAS)* définie par les entités contenues dans un lien sémantique. A titre d'exemple, la signature d'instances de l'association sémantique reliant "P0" et "PR0" est définie par l'ensemble d'instances/ressources {U0, C0}.
- *Signature de classes d'une association sémantique (SCAS)* définie par les classes des entités intermédiaires dans une AS. Par exemple, la signature de concepts/classes de l'association sémantique entre "P0" et "PR0" est définie par l'ensemble des classes {university, course}.

Dans la section précédente, nous avons discuté comment évaluer un LS de deux concepts directement liées par l'évaluation du degré de connectivité entre elles. Dans cette section nous essayerons de formuler une mesure permettant d'évaluer le degré d'une association sémantique (DAS) entre deux classes reliées par un lien sémantique. L'évaluation à travers la mesure DAS sera basée sur la mesure DLS des classes intermédiaires reliées.

Supposons que A et B sont deux concepts/classes contenus dans l'ontologie et DAS(A, B) représente le degré d'association sémantique entre eux. L'expression qui calcule DAS (A, B) est obtenue par les formules conditionnelles suivantes :

- *Cas 1* : Si A$\equiv$B ou B$\equiv$A, alors DAS(A, B)=1
- *Cas 2* : Si A$\supset$B ou A$\propto$B, alors DAS(A, B) =DLS(A,*l*)
- *Cas 3* : Si Non(B$\subset$A) et Non(A$\propto$B), alors DAS(A, B)=0
- *Cas 4* : Autrement, A et B ont une association sémantique. L'expression DAS(A, B) est définie par l'équation (10) :

$$DAS(A,B) = log(n^2). \prod_{i=A\ldots X, j=Y\ldots B} DLS_i(j,l) \tag{10}$$

Où $l \in$(A$\supset$X), Y$\propto$A, B$\equiv$Y et n est le nombre de concepts/classes intermédiaires qui mènent de A à B (les entités dans l'ensemble SCAS).

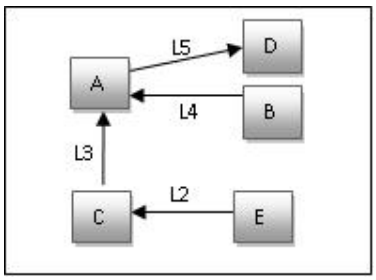


FIG. 2 – Exemples de liens et d'associations sémantiques dans l'ontologie MeSH.

S'il existe un chemin possible entre A et B, la valeur DAS de l'association sémantique entre A et B est calculée par la multiplication des DLS des paires de classes contenues dans ce chemin. Sinon la valeur de DAS est évaluée à 0.

Pour plus de clarification du principe de DAS et DLS, nous énumérons différents types d'associations sémantiques dans l'ontologie MeSH par les exemples de la figure 2 et du tableau 2.

	A	B	C	D	E
A	1	0	0	$\text{DLS}_A(D, L5)$	0
B	$\text{DLS}_B(A, L4)$	1	0	$\text{DLS}_B(A, L4)*$ $DLS_A(D, L5)*$ $Log(4)$	0
C	$\text{DLS}_C(A, L3)$	0	1	$\text{DLS}_C(A, L3)*$ $DLS_A(D, L5)*$ $Log(4)$	0
D	0	0	0	1	0
E	$\text{DLS}_E(C, L2)*$ $DLS_C(A, L3)*$ $Log(4)$	0	$\text{DLS}_E(C, L2)$	$\text{DLS}_E(C, L2)*$ $DLS_C(A, L3)*$ $DLS_A(D, L5)*$ $Log(9)$	1

TAB. 2 – Exemples d'évaluation en utilisant la formule DAS

Une entité dans une base de connaissance RDF est une instance d'une classe spécifique. Si A et B sont deux entités dans la base de connaissance, nous pouvons extraire des informations inattendues à travers l'extraction d'associations sémantiques entre une entité donnée et une autre non spécifiée au départ.

3.4 Exemple d'évaluation d'associations sémantiques : La mesure DAS basée sur la fréquence des propriétés entrantes (DAS-PE)

DAS-PE représente le degré maximum des propriétés entrantes qui relient une entité spécifique d'une association par un lien *l* déjà connu. Cet exemple de degré est men-

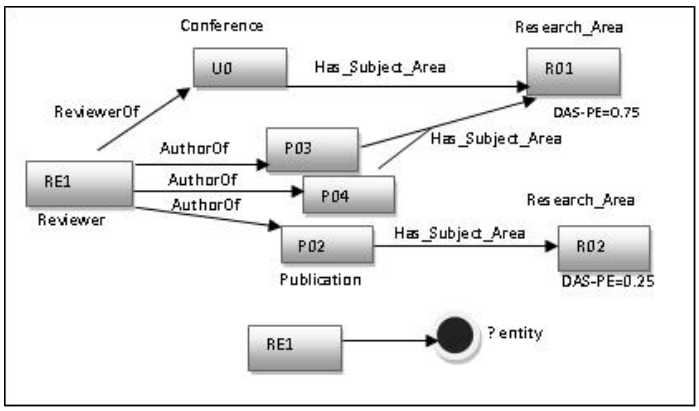


FIG. 3 – Un exemple fictif d'évaluation de DAS-PE.

tionné dans notre précédent travail qui décrit le langage PmSPARQL pour l'extraction des chemins à partir d'un graphe RDF (Slimani *et al.*, 2008). L'exemple présenté dans la figure 3 décrit une évaluation des liens sémantiques à travers la recherche du champ d'expertise d'un "Reviewer" pour une conférence déterminée. Les résultats peuvent être limités à l'identification d'un nouveau lien sémantique (Expert-in) entre l'objet source prédéfini et un objet initialement inconnu.

La valeur de DAS-PE de chaque association sémantique qui mène de A à B est obtenue par la valeur de la probabilité maximale $P_A(B_i|l)$ d'une entité spécifique (B_i) en respectant la théorie des probabilités.

$P_A(B_i|l)$ se réfère à la probabilité de l'entité A étant donnée l'entité B_i et le lien sémantique *l*. La formule de DAS-PE est obtenue par l'expression : $P_A(B_i|l)$=CB_i/NIR ; CB_i désigne le nombre de relations (propriétés) entrantes reliant l'entité spécifiée B_i et *NIR* désigne le nombre total des relations entrantes reliant les instances de B à travers le lien *l*. L'expression de la formule DAS-PE est spécifiée comme suit :

$$DAS - PE(AS) = Max\{P_A(B_1|l), P_A(B_2|l), ..., P_A(B_n|l)\} \quad (11)$$

Dans la figure 3, l'entité "R01" qui désigne l'instance de la classe "Research-Area", possède 3 relations entrantes parmi 4 relations reliant toutes les entités de type "Research-Area" avec le lien "Has-Subject-Area". La valeur de probabilité de "R01" =3/4 =0.75 est la valeur maximale qui sera affectée à DAS-PE. Et finalement le lien sémantique cherché est défini par : (RE1$\xrightarrow{Expert-in}$R01).

Intuition : L'évaluation d'associations sémantiques par DAS-PE est importante dans le cas où la recherche d'un genre de lien sémantique n'est pas explicitement existant dans la base de connaissance. Le nouveau lien sémantique découvert (expert-in) aide à l'enrichissement de la base de connaissance, qui facilite la découverte de nouveaux LS inattendus.

4 Résultats expérimentaux

Au niveau de cette section, nous discutons les évaluations expérimentales, ainsi que les tests appliqués sur la formule DAS discutée dans la section 3. Les tests d'évaluation sont réalisés sur une machine avec un processeur Intel (R) Core (TM) 2 Duo 2.2GHZ, une mémoire de 2GB et Windows Xp.

Nous avons adopté DragonToolkit[3] qui est un paquet de développement Java utile pour l'utilisation académique dans l'extraction sémantique pour évaluer l'efficacité de notre mesure DAS.

Nom du concept		Pain
Liens	Related-To	Intractable
	SubClassOf	Headache
	SynonymOf	Postoperative

TAB. 3 – Les informations de l'ontologie MeSH utilisées dans nos expérimentations.

Pour ce faire, nous avons appliqué le principe de l'expansion des requêtes (Grootjen & T.P.v.d, 2006) appliqué sur l'ontologie MeSH et l'ensemble des données CFC[4] incluant des résumés documentaires. Dans les expériences réalisées, le terme "classe" discutée plus haut est remplacé par le terme "concept" pour des raisons de convenance avec le contenu de l'ontologie MeSH. Les informations utilisées à partir de l'ontologie MeSH sont essentiellement présentées dans le tableau 3. Ce tableau représente le concept "Pain" en terme des liens décrits dans la section 3.2 ("Related-To" c'est un lien LS, "SubClassOf" est un LH et "SynonymOf" est un LE).

Nous avons appliqué une recherche basée sur la sémantique de l'expansion des requêtes pour comparer l'efficacité de la méthode DAS proposée par rapport à DAS dérivé de l'intuition (manuellement). Dans l'étude comparative nous avons utilisé la même valeur proposée dans le travail de (Tian *et al.*, 2007) pour faciliter la comparaison avec d'autres approches. La valeur de DLS avec notre méthode a été fixée à 0.8 (analyse manuelle du corpus) et celle de l'intuition a été fixée à 0.5 (valeur de λ décrite dans la section 3.2.2).

La méthode MAP (*Mean Average Precision*) est l'outil d'évaluation traditionnel qui est adopté dans le domaine de la recherche documentaire. Elle calcule la précision moyenne de toutes les requêtes. MAP (n) est employée dans notre travail pour évaluer les n documents recherchés. La formule *précision (n)* est donc adoptée pour mesurer la précision des n documents cherchés. La mesure de précision est bien connue pour l'examination de la qualité des documents appropriés.

La précision moyenne PM (*Average Precision*) tient en compte la moyenne des scores de précision des documents appropriés parmi top-k documents recherchés par une requête simple. La précision moyenne d'une requête i est définie par la formule suivante :

[3]http ://www.ischool.drexel.edu/dmbio/dragontool/

[4]http ://www.ischool.berkeley.edu/ hearst/irbook/cfc.html

	Document approprié	Précision(30)	MAP(30)
Intuition	22	81.32%	85.12%
DAS	29	91.34%	93.48%
DLS	27	95.33%	97.24%

TAB. 4 – Comparaison de l'efficacité de la recherche sémantique basée sur la mesure DAS avec celle basée sur DLS et l'intuition

$$PM_i(k) = \frac{\sum_{j=1...n_i} j/R_{i,j}}{n_i} \tag{12}$$

Où n_i est le nombre des documents appropriés de l'ième requête, k est le nombre des documents extraits, et R_{ij} est le rang du jème document approprié de l'ième requête.

La fonction MAP(k) est utilisée ici pour calculer le rang des k-premiers documents recherchés avec une précision (n). L'expression de MAP(k) est obtenue par la formule suivante :

$$MAP(k) = \frac{\sum_{i=1...q_n} PM_i(k)}{q_n} \tag{13}$$

Où q_n est le nombre de requêtes exécutées. Les résultats qui figurent dans le tableau 4 montrent l'amélioration apportée par notre méthode DAS permettant d'évaluer les AS par rapport à la méthode DAS discutée par Tian et al (Tian *et al.*, 2007).

L'approche adoptée est appliquée sur une ontologie de grande taille (MeSH) et montre une lenteur au niveau de l'évaluation de la précision des documents. Cependant, un travail conséquent reste à faire d'une part en amont sur l'analyse théorique de ces mesures, et d'autre part sur leur implémentation à grande échelle.

5 Conclusion

Dans cet article nous avons proposé une approche qui se préoccupe par l'évaluation des associations sémantiques dans le cadre de la recherche de l'information basée sur une ontologie de domaine. À ce propos, nous avons présenté une mesure DLS pour mesurer les liens sémantiques directs entre des entités contenues dans une base de connaissance. Nous avons présenté une deuxième mesure qui évalue le degré d'une association sémantique (DAS) entre deux entités qui sont liées d'une manière indirecte. Les mesures proposées pour l'évaluation des liens et des associations sémantiques ont été appliquées sur une ontologie de domaine et du corpus sous format des résumés. Les résultats obtenus montrent une amélioration remarquable au niveau de la précision dans le domaine de l'extraction des documents appropriés.

Références

ALEMAN-MEZA B., HALASCHEK-WIENER C., ARPINAR I. B. & SHETH A. P. (2003). Context-aware semantic association ranking. In *Proceedings of SWDB'03, The first International Workshop on Semantic Web and Databases, Co-located with VLDB 2003, Humboldt-Universität, Berlin, Germany, September 7-8*, p. 33–50.

ALEMAN-MEZA B., NAGARAJAN M., RAMAKRISHNAN C., DING L., KOLARI P., SHETH A. P., ARPINAR I. B., JOSHI A. & FININ T. (2006). Social networks : Semantic analytics on social networks : Experiences in addressing the problem of conflict of interest detection. In *Proceedings of the 15th international conference on World Wide Web, WWW 06*, p. 407–416 : ACM Press.

ANYANWU K., MADUKO A. & SHETH A. (2005). Sem-rank : Ranking complex relationship search results on the semantic web. In *International World Wide Web Conference*, volume 14, p. 117–127, New York, NY, USA : ACM Press.

CAO G., NIE J.-Y. & BAI J. (2005). Integrating word relationships into language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 298–305.

GROOTJEN F. & T.P.V.D W. (2006). Conceptual query expansion. In *Data & Knowledge Engineering*, p. 174–193.

JAFFRI A., GLASER H. & MILLARD L. (2008). Uri disambiguation in the context of linked data. In *Proceedings of the 17th international conference on World Wide Web, WWW08* : ACM Press.

NING X., JIN H. & WU H. (2006). Semrex : Towards large-scale literature information retrieval and browsing with semantic association. In *Proceedings of IEEE International Conference on e-Business Engineering (ICEBE 06)*, p. 602–609 : IEEE Computer Society.

PETERSON D., CREGAN A., ATKINSON R. & BRISBIN J. (2008). Exploiting social capital to create a standards-rich semantic network. In *Proceedings of the 17th international conference on World Wide Web, WWW08* : ACM Press.

SLIMANI T., BEN YAGHLANE B. & MELLOULI K. (2008). Pmsparql : Extended sparql for multiparadigm path extraction. *International Journal of Computer, Information, and Systems Science, and Engineering.*, **2 (3)**, 179–190.

SONG F. & CROFT W. B. (1999). A general language model for information retrieval. In *In CIKM 99 : Proceedings of the eighth international conference on Information and knowledge management*, p. 316321, New York, NY, USA : ACM Press.

TIAN X., LI H. & DU X. (2007). Measuring semantic association in domain ontology. In *IEEE Third International Conference on Semantics, Knowledge and Grid*, p. 515–518 : IEEE Computer Society.

WATABE H. & KAWAOKA T. (2001). The degree of association between concepts using the chain of concepts. In *Systems, Man, and Cybernetics*, volume 2, p. 877–881, Tucson, AZ, USA : Computer Society.

ZHUGE H. (2007). Autonomous semantic link networking model for the knowledge grid. In *Concurrency and Computation : Practice & Experience*, volume 19, p. 1065 – 1085 : John Wiley and Sons Ltd.

SEMIOSEM : une mesure de similarité conceptuelle fondée sur une approche sémiotique

Xavier Aimé[1,3], Frédéric Fürst[2], Pascale Kuntz[1], Francky Trichet[1]

[1] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissance & Décisions
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
{pascale.kuntz,francky.trichet}@univ-nantes.fr
[2] MIS - Modélisation, Information et Systèmes
Université de Picardie - Jules Verne
33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
[3] Société TENNAXIA
37 rue de Châteaudun - 75009 Paris
xaime@tennaxia.com

Abstract : Cet article propose une nouvelle mesure de similarité conceptuelle baptisée SEMIOSEM (*Semiotic-Based Similarity Measure*). La première originalité de cette mesure est de prendre en compte les trois dimensions sémiotiques de la conceptualisation sous-jacente à une ontologie de domaine : l'*intension* (*i.e.* les propriétés utilisées pour définir les concepts et la structure de la hiérarchie de subsomption), l'*extension* (*i.e.* les instances des concepts) et l'*expression* (*i.e.* les termes utilisés pour dénoter à la fois les concepts et leurs instances). Ainsi, SEMIOSEM vise à aggréger et enrichir des mesures existantes de types intensionnel et extensionnel. La seconde originalité de cette mesure est d'être sensible au contexte dans lequel l'utilisateur met en œuvre SEMIOSEM. Ce contexte s'exprime au moyen d'un corpus, d'un ensemble d'instances et d'une valeur caractérisant son état émotionnel. Ainsi, SEMIOSEM s'avère être plus flexible, plus robuste et plus proche du jugement de l'utilisateur que les autres mesures de similarité, lesquelles sont généralement fondées sur un seul aspect d'une conceptualisation et ne prennent pas en compte le contexte d'utilisation.
Mots-clés : Mesure de similarité, Sémiotique, Mesure sémantique, Proximité conceptuelle.

1 Introduction

À l'heure actuelle, la notion de similarité est mise en avant dans plusieurs domaines d'activités liés à l'ingénierie des ontologies tels que l'apprentissage, l'alignement ou en-

core le peuplement d'ontologies. Ces dernières années, de nombreuses mesures dédiées à la définition de la (dis-)similarité entre concepts ont été proposées. Ces mesures peuvent être classées suivant deux approches : (i) les mesures de type extensionnel telle que Resnik, Lin, Jiang et Conrath ou d'Amato et (ii) les mesures de type intensionnel telle que Rada, Leacock et Chodorow ou Wu et Palmer. La plupart de ces mesures se focalisent sur un seul aspect de la conceptualisation sous-jacente à une ontologie de domaine, soit l'*intension* – au travers de la structure de la hiérarchie de subsomptions, soit l'*extension* – au travers des instances de concepts ou des occurences de termes dénotant les concepts au sein d'un corpus. De plus, ces mesures sont majoritairement sensibles à la structure de la hiérarchie de subsumptions (par l'utilisation du subsumant commun le plus spécifique) et, par conséquent, dépendantes des choix de modélisation. Enfin, ces mesures ne prennent pas en compte la perception du domaine par l'utilisateur de l'ontologie.

Cet article présente SEMIOSEM, une mesure de similarité définie dans le cadre d'une approche sémiotique permettant de combiner ces différentes approches. La première originalité de SEMIOSEM est de prendre en compte les trois dimensions d'une conceptualisation : (1) le *signifié*, *i.e.* le concept défini en intension, (2) le *signifiant*, *i.e.* les termes désignant le concept, et (3) le *référent*, *i.e.* le concept défini en extension. SEMIOSEM est ainsi une mesure issue de l'agrégation et l'enrichissement de travaux existants, avec pour particularité d'être indépendante de la structure de la hiérarchie de subsomptions. La seconde originalité de SEMIOSEM est d'être sensible au contexte, et en particulier aux particularités de chaque utilisateur. En effet, SEMIOSEM est fondé sur l'exploitation de multiples sources d'informations : (1) un corpus textuel fourni par l'utilisateur et reflétant les particularités de conceptualisation de ce dernier, (2) un ensemble d'instances propres à l'utilisateur, (3) une ontologie enrichie par la perception de l'utilisateur de l'importance de chaque propriété associée à un concept dans la définition même de ce dernier et enfin (4) l'état émotionnel de l'utilisateur. L'importance de chacune de ces ressources peut être modulée suivant le contexte d'usage et SEMIOSEM reste efficient même si une des sources est absente.

La suite de cet article est structurée comme suit. La section 2 introduit brièvement les mesures de similarité les plus connues. La section 3 décrit en détail SEMIOSEM : les fondements, les définitions formelles, les paramètres liés à l'utilisateur et leurs interactions. La section 4 présente des résultats expérimentaux et compare notre mesure avec les travaux existants dans le contexte d'un projet dédié à la veille juridique sur des documents réglementaires relatifs au domaine "Hygiène, Sécurité et Environnement" (HSE).

2 Mesures de similarité existantes

2.1 Mesures de type intensionnel

Les mesures de type intensionnel sont fondées sur l'analyse et l'exploitation de la structure des réseaux sémantiques. Une hiérarchie de concepts est considérée comme un graphe orienté (où les arcs correspondent à des liens *is-a* et les noeuds à des concepts) au sein duquel des indices (par exemple la profondeur ou la densité) sont utilisés pour

comparer les noeuds. Intuitivement, tous ces travaux sont fondés sur le principe suivant : un objet A est jugé plus similaire à un objet B qu'à un objet C, si la distance de A à B au sein du graphe est plus courte que celle de A à C.

Rada *et al.* (1989) considère cette distance, notée $dist_{edge}(c_1, c_2)$, comme étant la longueur du plus court chemin entre deux concepts. La similarité entre c_1, c_2 est définie par :

$$Sim_{Rad}(c_1, c_2) = \frac{1}{dist_{edge}(c_1,c_2)}$$

Resnik (1995) complète cette définition en utilisant la profondeur maximale de la hiérarchie. La similarité entre c_1, c_2 est définie par :

$$Sim_{Res}(c_1, c_2) = \frac{2*prof_{max}}{dist_{edge}(c_1,c_2)}$$

Leacock & Chodorow (1998) normalisent cette distance de la façon suivante :

$$Sim_{Lea}(c_1, c_2) = -log\left(\frac{dist_{edge}(c_1,c_2)}{2*max}\right)$$

Wu & Palmer (1994) proposent une autre mesure de similarité, laquelle prend en compte la profondeur des concepts dans la hiérarchie. La similarité entre c_1, c_2, avec $prof(c_i)$ la profondeur du concept c_i dans la hiérarchie et c le Plus Petit Père Commun (PPPC) à c_1 et c_2, est définie par :

$$Sim_{Wu}(c_1, c_2) = \frac{2*prof(c)}{prof(c_1)+prof(c_2)}$$

Ces mesures n'exploitent que les liens *isa* et laissent de coté toute la richesse sémantique de l'intension des concepts, ce qui les rend parfois incorrectes (des concepts ayant une mesure de similarité élevée peuvent ne pas être sémantiquement proches) et souvent incomplètes (des concepts sémantiquement similaires mais non fortement reliés dans la hiérarchie auront une mesure de similarité faible).

Une autre approche de type intensionnel consiste à analyser et comparer les propriétés des concepts. Nous pouvons dire que deux concepts sont proches si le cardinal de l'intersection de leurs caractéristiques communes est plus grand que celui des caractéristiques qui les différencient[1]. Tversky (1977) propose la mesure de similarité suivante (avec α, β, γ des constantes) :

$$Sim_{Tversky}(c_1, c_2) = \alpha.comm(c_1, c_2) - \beta.diff(c_1, c_2) - \gamma.diff(c_2, c_1)$$

2.2 Mesures de type extensionnel

Les premières mesures de type extensionnel furent directement inspirées de celle de Jaccard (1901), *i.e.* le ratio entre le nombre d'instances communes et le nombre total d'instances de deux concepts. I_c étant l'ensemble des instances du concept c, cette mesure est définie par :

$$Sim_{Jaccard}(c_1, c_2) = \frac{|I_{c_1} \cap I_{c_2}|}{|I_{c_1}|+|I_{c_2}|-(|I_{c_1} \cap I_{c_2}|)}$$

[1]Dans la formule ci-après, $comm$ représente le nombre de propriétés communes à c_i de c_j, et $diff$ le nombre de propriétés qui différencient c_i de c_j.

Selon d'Amato *et al.* (2008), cette approche n'est pas réellement appropriée aux ontologies, car deux concepts peuvent être similaires sans pour autant avoir d'instances en commun. d'Amato *et al.* (2008) propose en conséquence une nouvelle mesure basée non pas sur l'intersection des extensions, mais sur la variation de la cardinalité des extensions pour les concepts considérés par rapport à leur plus petit père commun (*i.e. PPPC*), où I l'ensemble des instances de l'ontologie.

$$Sim_{Ama}(c_1, c_2) = \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{PPPC(c_1,c_2)}|}(1 - \frac{|I_{PPPC(c_1,c_2)}|}{|I|})(1 - \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{PPPC(c_1,c_2)}|})$$

La plupart des mesures de type extensionnel sont fondées sur la notion de Contenu Informationnel (CI) d'un concept, introduite par Resnik (1999), et basée sur la probabilité $p(c)$ d'avoir ce concept dans un corpus donné.

$$\Psi(c) = -log(p(c)) \text{ où } p(c) = \frac{\sum_{n \in words(c)} count(n)}{N}$$

où N représente le nombre total d'occurrences des termes de tous les concepts dans le corpus et $words(c)$ représente l'ensemble des termes possibles pour dénoter le concept c, ou un de ses descendants dans la hiérarchie. Ceci suppose au départ que chaque terme est attribué de manière unique à un concept, autrement dit qu'il n'existe aucune ambiguïté. Sanderson & Croft (1999) corrige ce problème de la façon suivante (où $nbc(n)$ est égal au nombre de concepts dont le terme n est label) :

$$p(c) = \frac{\sum_{n \in words(c)} \frac{count(n)}{nbc(n)}}{N}$$

La mesure de similarité proposée par Resnik (1999) est fondée sur le subsumant commun de c_1 et de c_2 ayant le CI le plus élevé (ce subsumant commun n'est pas forcément le PPPC). La similarité entre c_1, c_2, où $S(c_1, c_2)$ est l'ensemble des concepts qui subsument à la fois c_1 et c_2, est définie par :

$$Sim_{Res2}(c_1, c_2) = max_{c \in S(c_1,c_2)} \Psi(c)$$

Lin (1998) propose une mesure fondée sur le CI commun aux deux concepts. La similarité entre c_1, c_2 avec ppc le concept de $S(c_1, c_2)$ qui minimise $p(c)$, est définie par :

$$Sim_{Lin}(c_1, c_2) = \frac{2 * \Psi(ppc)}{\Psi(c_1) + \Psi(c_2)}$$

Fondée sur cette même approche, Jiang & Conrath (1997) proposent la mesure suivante (où $TC(c_i, c_j)$ pondère l'arc reliant c_i à c_j) :

$$Sim_{Jiang}(c_1, c_2) = \sum_{c \in path(c_1,c_2) - PPPC(c_1,c_2)} [\Psi(c) - \Psi(pere(c))] * TC(c, pere(c))$$

3 SEMIOSEM : une mesure de similarité sémiotique

Construire une ontologie O d'un domaine D consiste à spécifier une conceptualisation consensuelle de connaissances individuelles. Nous appelons endogroupe l'ensemble

des personnes qui partagent la conceptualisation capturée dans l'ontologie. Pour un même domaine, plusieurs ontologies peuvent être définies par différents endogroupes. Nous qualifions ces ontologies d'*Ontologies Vernaculaires du Domaine* (OVD), le terme vernaculaire étant utilisé au sens de relatif à une communauté d'usages, et non au sens de populaire (Aimé *et al.* (2008)). Nous définissons une *Ontologie Vernaculaire de Domaine* (OVD), pour un domaine D donné et un endogroupe G donné, par le tuple suivant :

$$O_{(D,G)} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\right\} \text{ où}$$

- $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont les ensembles de concepts, de propriétés et d'instances des concepts ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés[2] ;
- $dom\ : \mathcal{P} \rightarrow \mathcal{C}$ et $codom\ : \mathcal{P} \rightarrow (\mathcal{C} \cup Datatypes)$ associent à chaque propriété son domaine et éventuellement son co-domaine ;
- $\sigma\ : \mathcal{C} \rightarrow \mathscr{P}(\mathcal{I})$ associe à chaque concept ses instances ;
- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D où :
  - L_C, L_P et L_I sont les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$;
  - les fonctions $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ associent aux primitives conceptuelles les termes qui les désignent.

Cependant, une telle ontologie (1) ne capture pas la totalité des connaissances que les membres de l'endogroupe ont sur le domaine, et (2) ne tient pas compte du contexte dans lequel elle est utilisée. Une OVD peut donc être pragmatisée, c'est-à-dire personnalisée et contextualisée au moyen de ressources additionnelles représentant des connaissances particulières à l'utilisateur et son contexte d'utilisation. Cette pragmatisation ne remet pas en cause la sémantique (formelle) de l'OVD, mais consiste à ajouter une couche de connaissances, et conduit à une *Ontologie Personnalisée Vernaculaire du Domaine* (OPVD). Cette approche est également qualifiée par E. Rosch d'*écologique* (Gabora *et al.* (2008)), dans le sens où elle est fonction de l'endogroupe, mais également du contexte. SEMIOSEM est une mesure de similarité, personnalisée et contextualisée, et donc définie sur une OPVD.

Notre approche est fondée sur les trois dimensions introduites par Morris et Peirce dans leurs théories de la sémiotique : (1) le *signifié*, *i.e.* le concept défini en intension, (2) le *signifiant*, *i.e.* les termes désignant le concept, et (3) le *référent*, *i.e.* le concept défini en extension. Nous pragmatisons donc une OVD au moyen de ressources propres à l'utilisateur et fournies par lui : (1) des pondérations des propriétés des concepts

[2] $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

de l'OVD, (2) des instances et (3) un corpus supposé représentatif de l'univers cognitif de l'utilisateur (ou du groupe d'utilisateurs). Aussi, SEMIOSEM correspond à une agrégation de trois composantes pondérées selon le contexte et l'utilisateur[3] :

- une composante *intensionnelle* fondée sur la comparaison des propriétés des concepts dans l'OPVD ;
- une composante *extensionnelle* fondée sur la comparaison des instances des concepts dans l'OPVD ;
- une composante *expressionnelle* fondée sur la comparaison entre les termes désignant les concepts et leurs instances dans le corpus.

SEMIOSEM : $\mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$SemioSem(c_1, c_2) = [\alpha * intens(c_1, c_2) + \beta * extens(c_1, c_2) + \gamma * express(c_1, c_2)]^{\frac{1}{\delta}}$$

Les sections 3.1, 3.2 et 3.3 présentent respectivement les fonctions $intens$, $extens$ et $express$ et la section 3.4 donne le sens des paramètres α, β, γ et δ et propose une méthode pour en fixer les valeurs.

3.1 Composante intensionnelle

Le calcul de cette composante *intensionnelle* s'inspire de Au Yeung & Leung (2006) et s'appuie sur la représentation des concepts par des vecteurs dans l'espace des propriétés de l'ontologie. Formellement, à tout concept $c \in \mathcal{C}$, est associé le vecteur $\overrightarrow{v_c}$= $(v_{c1}, v_{c2}, ..., v_{cn})$ avec $n = |\mathcal{P}|$ et $v_{ci} \in [0,1], \forall i \in [1,n]$. v_{ci} est la pondération fixée par l'utilisateur pour le concept c par rapport à la propriété i (v_{ci} vaut 1 si l'utilisateur n'a pas fixé ces pondérations)[4]. L'ensemble des concepts forme ainsi un nuage de points dans un espace à $|\mathcal{P}|$ dimensions.

Nous calculons un vecteur prototype de c_p, qui a été originellement introduit dans Au Yeung & Leung (2006) comme une moyenne des vecteurs des concepts fils de c_p. Cependant, Au Yeung & Leung (2006) ne prend en compte dans sa moyenne que les concepts qui héritent directement de c_p. Pour notre part, nous étendons le calcul à tous les concepts de la descendance. En effet, des propriétés qui apparaissent uniquement sur des descendants indirects du concept père peuvent apparaitre dans le prototype du père, en particulier si l'aspect intensionnel est important. Le vecteur prototype p_{c_p} est donc un vecteur dans l'espace des propriétés, où l'importance de la propriété i est la moyenne des importances des propriétés des concepts de la descendance de c_p possédant i. Si pour $i \in \mathcal{P}$, $S_i(c) = \{c_j \leq^C c, c_j \in dom(i)\}$ alors :

[3] Ainsi, un zoologue aura tendance à conceptualiser en intension les connaissances du domaine des espèces animales (par des propriétés biologiques), alors que la plupart des personnes utilisent davantage des conceptualisations extensionnelles (basées sur les animaux rencontrés au cours de leur vie).

[4] La méthode que nous proposons pour fixer ces pondérations est la suivante. Pour chaque propriété p, l'utilisateur classe tous les concepts possédant p, afin de refléter sa perception de l'importance de p pour définir c en comparaison avec les autres concepts possèdant p. Cela conduit à ordonner les concepts possédant une même propriété (par exemple – pour la propriété *peut flotter* – l'ordre sera (*bateau* $>$ *tronc d'arbre* $>$ *canard*) car la propriété est plus importante pour un *bateau* ; bien sûr, un *canard* peut flotter mais ce n'est pas une propriété fondamentale pour ce concept.

$$\overrightarrow{p_{c_p}}[i] = \frac{\sum_{c_j \in S_i(c_p)} \overrightarrow{v_{c_j}}[i]}{|S_i(c_p)|}$$

D'un point de vue *intensionnel*, plus les prototypes respectifs de c_1 et c_2 sont proches, *i.e.* plus leurs propriétés sont proches, plus ces concepts sont similaires. La composante intensionnelle $intens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est donc calculée comme la distance entre les vecteurs prototypes des deux concepts. Cette fonction est définie par :

$$intens(c_1, c_2) = 1 - dist(\overrightarrow{p_{c_1}}, \overrightarrow{p_{c_2}})$$

3.2 Composante extensionnelle

D'un point de vue *extensionnel*, nos travaux sont fondés sur la mesure de similarité de Jaccard (1901). La fonction $extens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$extens(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{|\sigma(c_1)| + |\sigma(c_2)| - (|\sigma(c_1) \cap \sigma(c_2)|)}$$

Cette fonction est définie par le ratio entre le nombre d'instances communes et le nombre total d'instances moins le nombre d'instances en commun. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre d'instances en commun et très peu d'instances distinctes.

3.3 Composante expressionnelle

D'un point de vue *expressionnel*, plus les termes respectifs de chaque concept sont présents ensemble dans les mêmes documents, plus les concepts c_1 et c_2 sont jugés similaires. La composante expressionnelle $express : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$express(c_1, c_2) = \sum_{t_1, t_2} \left(\frac{min(count(t_1), count(t_2))}{N_{occ}} * \frac{count(t_1, t_2)}{N_{doc}} \right)$$

Où (1) $t_1 \in terms(c_1)$ et $t_2 \in terms(c_2)$ et $terms(c)$ l'ensemble des termes désignant le concept c ou un de ses descendants (direct ou non), (2) $count(t_i)$ est le nombre d'occurrences du terme t_i dans les documents du corpus, (3) $count(t_1, t_2)$ est le nombre de documents du corpus où les termes t_1 et t_2 apparaissent simultanément, (4) N_{doc} est le nombre total de documents du corpus, et (4) N_{occ} est la somme de tous les nombres d'occurrences de tous les termes du corpus.

3.4 Paramètres de SEMIOSEM

α, β et γ sont des coefficients (positifs ou nuls) de pondération des trois composantes SEMIOSEM. Dans un souci de normalisation, nous imposons que les composantes varient dans l'intervalle $[0, 1]$, et que $\alpha + \beta + \gamma = 1$. Les valeurs de ces trois coefficients peuvent être fixées arbitrairement, ou calibrées par expérimentations. Nous proposons une méthode pour en calculer automatiquement des approximations. Comme le montre la figure 1, nous considérons que le triplet (α, β, γ) caractérise les coordonnées cognitives de l'utilisateur dans le triangle sémiotique. Pour fixer les valeurs de α, β et γ, nous proposons de calculer les ratio γ/α et γ/β, les valeurs des coefficients étant déduites de

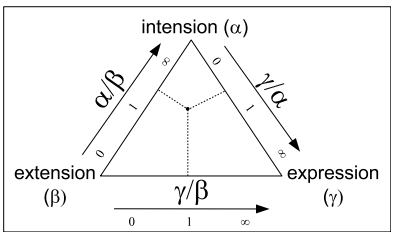


Figure 1: Les coefficients de pondération des composantes de SEMIOSEM comme coordonnées dans le triangle sémiotique. γ/α proche de 0 indique que l'utilisateur a une approche beaucoup plus intensionnelle qu'expressionnelle du domaine, le même rapport proche de l'infini indique le contraire, et le même rapport égal à 1 indique un équilibre entre les approches intensionnelle et extensionnelle. La même interprétation est adoptée pour les autres rapports. Quand les trois approches sont équilibrées, on a $\alpha = \beta = \gamma = 1/3$, les trois rapports sont égaux à 1 et les coordonnées cognitives de l'utilisateur correspondent au barycentre du triangle sémiotique.

l'équation $\alpha + \beta + \gamma = 1$. γ/α (resp. γ/β) est approximé par le taux de couverture des concepts (resp. des intances) de l'ontologie par le corpus. Ce taux est égal au nombre de concepts (resp. d'instances) dont au moins un des termes apparait dans le corpus divisé par le nombre total de concepts (resp. d'instances).

Le facteur $\delta \geq 0$ a pour objectif de tenir compte de l'état émotionnel de l'utilisateur. De multiples travaux ont été réalisés en Psychologie Cognitive sur le lien entre émotions et cognition, émotions et jugements (Bluck & Li (2001)). La conclusion de ces travaux peut être résumée ainsi : quand nous sommes dans un état émotionnel négatif (par exemple stress, colère), nous avons tendance à nous concentrer sur ce qui nous semble être le plus important, le plus caractéristique, le plus familier, ou le plus chargé émotionnellement dans nos souvenirs. Inversement, dans un état émotionnel positif (par exemple joie, amour), nous avons un jugement plus ouvert et nous acceptons plus facilement les éléments considérés comme non-caractéristiques. Selon Mikulincer *et al.* (1990), un état émotionnel négatif engendre une diminution dans les valeurs de représentation, et inversement pour un état émotionnel positif. Dans SEMIOSEM, nous caractérisons (1) un état émotionnel *négatif* par une valeur de $\delta \in]1, +\infty[$, (2) un état émotionnel *positif* par une valeur de $\delta \in]0, 1[$, et (3) un état émotionnel *neutre* par une valeur de 1. Ainsi, une très faible valeur de δ, qui caractérise un état émotionnel positif, va avoir pour effet d'augmenter la valeur de similarité des concepts qui, initialement, ne seraient pas considérés comme similaires. Inversement, une forte valeur de δ, qui caractérise un état émotionnel négatif, va avoir pour effet de diminuer ces valeurs.

4 Expérimentation

SEMIOSEM est actuellement expérimentée dans le contexte d'un projet porté par la

société Tennaxia[5]. Dans le cadre de ce projet, une ontologie du domaine HSE[6] a été développée. Cette ontologie couvre entre autre le domaine des *substances dangereuses*, sous la forme d'un treillis de 3.776 concepts (profondeur=11, largeur=1300), et 15 propriétés telles que *est cancérigène* ou *est radioactif*. Afin de pouvoir évaluer notre mesure et comparer les résutats avec les travaux existants, considérons la hiérarchie présentée en figure 2. L'objectif est de calculer la similarité entre le concept *Carbone* et les sous-concepts de *Halogène*. Les experts de Tennaxia ont évalué ces similarités comme suit : *Fluor*=0,6 ; *Chlore*=0,6 ; *Brome*=0,3 ; *Iode*=0,3 et *Astate*=0,1. Les calculs suivants sont effectués à l'aide d'un corpus spécifique composé d'environ un millier de textes réglementaires relatifs au domaine HSE (principalement des lois, décrets, directives, etc.).

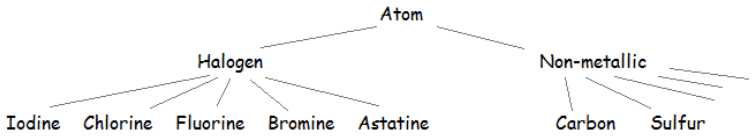


Figure 2: Extrait d'une hiérarchie de concepts.

Le tableau 1 présente les valeurs de similarité obtenues avec trois mesures de type intensionnel (Rada, Leacock et Wu) et trois mesures de type extensionnel (Lin, Jiang et Resnik). Nous pouvons noter que toutes les valeurs données par les mesures intensionnelles sont égales, car elles dépendent seulement de la structure de la hiérarchie.

Halogen	Rada	Leacock	Wu	Lin	Jiang	Resnik
Fluorine	0,25	0,097	0,6	0.31	0.14	1.43
Chlorine	0,25	0,097	0,6	0.28	0.12	1.43
Bromine	0,25	0,097	0,6	0.23	0.09	1.43
Iodine	0,25	0,097	0,6	0.22	0.09	1.43
Astatine	0,25	0,097	0,6	0	0	1.43

Table 1: Similarités avec le Carbone.

Le tableau 2 présente les valeurs de similarité obtenues avec SEMIOSEM dans le cadre de 6 contextes définis par les paramètres suivants : A ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 1$), B ($\alpha = 0.2$, $\beta = 0.7$, $\gamma = 0.1$, $\delta = 1$), C ($\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.7$, $\delta = 1$), D ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.33$, $\delta = 1$), E ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 0.1$) et F ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 5.0$).

Nous pouvons tout d'abord remarquer que quelque soit le contexte, SEMIOSEM fournit le même ordre de similarité que les autres mesures. Dans un contexte où la priorité est donnée à la composante intensionnelle (cf. contexte A), SEMIOSEM est meilleure

[5]Tennaxia est une société de service et de conseils en veille juridique et réglementaire dans le domaine Hygiène, Sécurité, Environnement et Développement Durable (HSE-DD) - www.tennaxia.com.

[6]Propriété Tennaxia - tous droits réservés – dépôt INPI N°322.408, 13 juin 2008 – dépôt *Scam Vélasquez* N°2008090075, 16 septembre 2008.

Halogen	A	B	C	D	E	F
Fluorine	*0.40*	0.14	0.32	0.27	0.91	0.025
Chlorine	*0.36*	0.12	0.29	0.25	0.90	0.017
Bromine	*0.29*	0.10	0.23	0.20	0.88	0.007
Iodine	*0.28*	0.10	0.23	0.19	0.88	0.006
Astatine	*0.01*	2.10^{-4}	2.10^{-4}	3.10^{-4}	0.63	1.10^{-8}

Table 2: Similarité avec le Carbone (SEMIOSEM).

que les autres mesures. Dans le contexte B qui donne la priorité à la composante extensionnelle (resp. le contexte C qui donne la priorité à la composante expressionnelle), SEMIOSEM est plus proche de la mesure de Jiang (resp. de la mesure de Lin). Dans un contexte qui ne donne aucune priorité spécifique (cf. contexte D), SEMIOSEM est entre la mesure de Lin et la mesure de Jiang. Deuxièmement, les contextes E et F montrent clairement l'influence du facteur émotionnel : un état mental positif (cf. contexte E) augmente très clairement les valeurs de similarité et un état mental négatif (cf. contexte F) diminue tout aussi clairement ces valeurs. Enfin, le concept *Astatine* n'est ni évoqué dans le corpus, ni représenté par des instances. Aussi, il n'est pas considéré comme similaire par les mesures de Lin et de Jiang, alors même que l'expert considère qu'une similarité existe. SEMIOSEM trouve une valeur de similarité grâce à sa composante intensionnelle.

5 Conclusion

Étant donné que l'utilisation d'une ontologie s'inscrit dans un contexte déterminé par une communauté d'usage et une application, nous soutenons qu'une mesure de similarité doit dépendre de ce contexte. Alors qu'une ontologie capture des connaissances consensuelles pour un endogroupe, nous préconisons de contextualiser les ontologies à l'aide de connaissances subjectives, variables d'un utilisateur à l'autre, et qui complètent les connaissances exprimées dans l'ontologie, sans remettre en cause leur sémantique. Basée à la fois sur l'ontologie et sur ces connaissances contextuelles, SEMIOSEM est ainsi une mesure particulièrement pertinente dès lors que la perception par l'utilisateur du domaine considéré peut avoir une large influence sur l'évaluation de la similarité entre les concepts.

Formellement, SEMIOSEM respecte les propriétés des mesures de similarité définies par d'Amato *et al.* (2008) : ***positivité*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) \geq 0$), ***reflexivité*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) \leq SemioSem(x, x)$) et ***symétrie*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) = SemioSem(y, x)$). Mais, SEMIOSEM n'est pas une distance de similarité car elle ne vérifie pas simultanément la propriété *strictness* ($\forall x, y \in \mathcal{C} : SemioSem(x, y) = 0 \Rightarrow x = y$) et l'***inégalité triangulaire*** ($\forall x, y, z \in \mathcal{C} : SemioSem(x, y) + SemioSem(y, z) \geq SemioSem(x, z)$).

Nous avons choisi de rendre SEMIOSEM aussi indépendante que possible de la structure de l'ontologie, et en particulier indépendante de l'utilisation du *PPPC*. C'est pourquoi nous avons choisi d'utiliser la mesure de Jaccard pour la composante extension-

nelle et non la mesure de d'Amato *et al.* (2008) qui est certes plus précise, mais profondément dépendante de la structure de la hiérarchie. Pour la composante expressionnelle, notre approche est similaire aux travaux de Resnik, si ce n'est que (1) nous n'utilisons pas le *PPPC* et (2) nous ne considérons pas le corpus comme étant composé d'un seul et unique document – nous tenons compte de la granularité des muliples documents. Ce choix est justifié par le principe suivant : deux concepts fréquemment associés dans peu de documents sont moins similaires que s'ils étaient associés moins souvent, mais d'une manière uniforme dans la majorité des documents du corpus. Enfin, pour la composante intensionnelle, notre approche peut être chronophage (si l'utilisateur décide de pondérer chaque propriété[7]), mais elle s'avère totalement novatrice et présente des résultats prometteurs.

Pour résumer, SEMIOSEM est plus flexible (elle tient compte de plusieurs sources d'information), plus robuste (car elle fournit des résultats pertinents pour des cas atypiques comme celui de l'*Astatine* dans les résultats expérimentaux) et plus centré sur l'utilisateur que toutes les méthodes actuelles, car fondé sur sa perception du domaine et son état émotionnel.

Cependant, SEMIOSEM présente quelques limites. Tout d'abord, la pondération des propriétés peut s'avérer impraticable pour des ontologies de très grande taille. D'autre part, le temps de calcul du nombre d'occurences de termes dans les textes devient conséquent si le corpus est de très grande taille (cependant, ce calcul ne se fait qu'une seule fois). Enfin, SEMIOSEM est dépendante de l'imprécision des calculs d'occurences liés aux limites du TALN. En effet, nos calculs se fondent sur la fréquence d'apparition de termes dans les documents. Il s'agit d'une donnée statistique purement syntaxique et nullement sémantique. Elle prend en compte l'apparition d'un ensemble de lettres juxtaposées formant un mot, mais nullement l'environnement qui va en influer le sens, et donc la sémantique. Il en est ainsi de syntagmes comme " l_1 *mais surtout pas* l_2 ", " l_1 *et* l_2 *n'ont rien à voir* ", ou encore " l_1 *et* l_2 *sont incompatibles* ". Il en est de même avec la présence d'anaphores (par exemple, " *Paul n'avait pas de voiture, je lui ai prêté la mienne* ") où les reprises sémantiques des précédents segments ne sont pas comptabilisées. Une manière de palier cet inconvénient serait d'étiqueter au prélable tout le corpus. Pour finir, fixer la valeur du coefficient de l'état émotionnel de l'utilisateur n'est pas trivial. Cependant, la mesure de cet état émotionnel peut se faire, soit en impliquant directement l'utilisateur au moyen d'un questionnaire qu'il devra remplir, soit de manière indirecte par la mesure de la vitesse de balayage de sa souris ou de la pression sur les touches du clavier, ou encore une analyse de son faciès, du clignement de ses yeux, etc.

References

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2008). Conceptual and lexical prototypicality gradients dedicated to ontology personalisation. In S. V. . HEIDELBERG.,

[7]Par défaut, toutes les pondérations sont égales à 1 si le concept possède la propriété, et la fonction *Intens* demeure valide. Dans le cas de notre expérimentation, les résultats obtenus dans ces contextes pour le concept Fluor sont : A - 0,59 ; B - 0,19 ; C - 0,38 ; D - 0,37 ; E - 0,95 ; F - 0,12.

Ed., *7th International Conference on Ontologies Databases and Applications of Semantics (ODBASE'2008 - Monterrey, Mexique). Lecture Notes in Computer Science (LNCS)*, volume 5332, p. 1423–1439. ISBN 978-3-540-88872-7.

AU YEUNG C. M. & LEUNG H. F. (2006). Ontology with likeliness and typicality of objects in concepts. In S. B. . HEIDELBERG, Ed., *Proceedings of the 25th International Conference on Conceptual Modeling - ER 2006*, volume 4215/2006. ISSN 0302-9743 (Print).

BLUCK S. & LI K. (2001). Predicting memory completeness and accuracy: Emotion and exposure in repeated autobiographical recall. *Applied Cognitive Psychology*, (15), 145–158.

D'AMATO C., STAAB S. & FANIZZI N. (2008). On the influence of description logics ontologies on conceptual similarity. In *EKAW 2008, International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns*, p. 48–63.

GABORA D. L. M., ROSCH D. E. & AERTS D. D. (2008). Toward an ecological theory of concepts. *Ecological Psychology*, **20**(1-2), 84–116.

JACCARD P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise de Sciences Naturelles*, **37**, 241–272. (in french).

JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxinomy. In *International Conference en Research in Computationnal Linguistics*, p. 19–33.

LEACOCK C. & CHODOROW M. (1998). *WordNet: an electronic lexical database*, chapter Combining local context and Wordnet similarity for word sense identification, p. 265–283. Cambridge, MA, The MIT Press.

LIN D. (1998). An information-theoric definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, p. 296–304.

MIKULINCER M., KEDEM P. & PAZ D. (1990). Anxiety and categorization-1, the structure and boundaries of mental categories. *Personnality and individual differences*, **11**(11), 805–814.

RADA R., MILI H., BICKNELL E. & M.BLETTNER (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man and Cybernetics*, **19**(1), 17–30.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, volume 1, p. 448–453, Montréal.

RESNIK P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, **11**, 95–130.

SANDERSON M. & CROFT W. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd International ACM SIGIR Conference*, p. 206–213.

TVERSKY A. (1977). Features of similarity. In *Psychological Review*, volume 84, p. 327–352.

WU Z. & PALMER M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, p. 133–138.

Gradients de prototypicalité appliqués à la personnalisation d'ontologies

Xavier Aimé[1,3], Frédéric Fürst[2], Pascale Kuntz[1], Francky Trichet[1]

[1] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissance & Décisions
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
{pascale.kuntz,francky.trichet}@univ-nantes.fr
[2] MIS - Modélisation, Information et Systèmes
Université de Picardie - Jules Verne
33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
[3] Société TENNAXIA
37 rue de Châteaudun - 75009 Paris
xaime@tennaxia.com

Abstract : Cet article présente une méthode originale de personnalisation des ontologies principalement dédiée à la personnalisation des SI à base d'ontologie. Cette méthode s'appuie sur l'ajout, à l'ontologie, de connaissances supplémentaires propres à l'utilisateur mais respectant la sémantique exprimée dans l'ontologie. Ces connaissances expriment des prototypicalités, c'est-à-dire des représentativités entre deux concepts ou entre un terme et le concept qu'il désigne. Nous proposons de calculer ces prototypicalités à partir des connaissances présentes dans l'ontologie et communes à tous les utilisateurs, et à partir de ressources propres à l'utilisateur, à savoir des instances de concepts, un corpus de textes et des pondérations fixées par l'utilisateur et exprimant l'importance des propriétés dans la définition des concepts. Les premières expérimentations, menées à l'aide d'un outil dédié appelé TooPrag, confirment l'intérêt de notre approche.
Mots-clés : Personnalisation, Prototypicalité, Sémiotique

1 Introduction

La personnalisation d'un système d'information (SI) vise à adapter son fonctionnement au profil et à l'activité de l'utilisateur, afin de lui permettre d'accèder aux informations les plus pertinentes à la mise en œuvre de cette activité. Ce processus de personnalisation est devenu de plus en plus crucial du fait de l'accroissement incessant du volume d'informations géré par des SI de plus en plus ouverts, c'est le cas en particulier des applications opérant sur le Web [Brusilovsky & Kobsa (2007)]. La personnalisation est souvent basée sur la représentation des préférences de l'utilisateur au

niveau de l'interface du système, qui peut, en fonction des préférences, réinterpréter des requêtes, les étendre au besoin, et/ou filtrer les résultats et adapter leur présentation.

Cependant, certains SI intègrent déjà une représentation des connaissances du domaine couvert par le système, et de plus en plus sous forme d'ontologies. Notre approche suppose que les utilisateurs du SI conceptualisent le domaine en accord avec l'ontologie considérée. Une personnalisation du SI est donc possible à travers cette ontologie, du fait qu'elle ne spécifie pas de façon complète la sémantique du domaine et ne capture que ce qui est consensuel dans les conceptualisations des utilisateurs. Nous proposons de faire de ces ontologies elles-mêmes le support de la personnalisation du SI, en ce sens qu'elles représentent un fond cognitif commun à tous les utilisateurs du système, et qu'il est possible de les adapter en y ajoutant des connaissances supplémentaires, variables selon les utilisateurs. Nous proposons d'utiliser comme connaissances additionnelles les degrés de prototypicalité entre deux entités cognitives, c'est-à-dire des degrés de représentativité d'une entité par rapport à l'autre [Harnad (2003)].

Nous introduisons ces prototypicalités, d'une part, entre deux concepts liés hiérarchiquement (prototypicalité conceptuelle) et, d'autre part, entre un concept et un terme le dénotant (prototypicalité lexicale), ce qui nous permet de personnaliser l'ontologie sur le plan conceptuel et sur le plan terminologique. Ces prototypicalités sont représentées, pour la prototypicalité conceptuelle, par des pondérations des liens hiérarchiques et des propriétés, et, pour la prototypicalité terminologique, par des pondérations des termes. Nous proposons également plusieurs méthodes permettant de calculer ces pondérations de façon automatique ou semi-automatique, en nous reposant (1) sur la structure formelle de l'ontologie, (2) sur une population d'instances des concepts de l'ontologie et (3) sur un corpus de textes relatifs au domaine couvert par l'ontologie.

Une première application de ce travail est effectuée dans le cadre du projet REDENE-10 (*REcherche Documentaire Ecologique Neurale et Émotionnelle*), développé au sein de l'entreprise Tennaxia et dédié à la recherche sémantique et personnalisée d'information dans le domaine de la législation HSE[1].

La suite de l'article est structurée comme suit. La section 2 présente notre approche de la personnalisation des ontologies, et les différents types de prototypicalité utilisés. La section 3 détaille les méthodes de calcul des prototypicalités conceptuelle et lexicale. La section 4 introduit quelques résultats expérimentaux et la section 5 compare nos travaux à d'autres approches.

2 Personnalisation des ontologies et prototypicalité

Les SI exploitent depuis des années les ontologies, définies comme des représentations conceptuelles des connaissances d'un domaine donné et reposant sur un consensus partagé par un endogroupe[2]. Classiquement, une ontologie est composée d'ensembles hiérarchisés de concepts et de propriétés[3], enrichis à l'aide d'axiomes affinant la repré-

[1] Tennaxia est une société de service et de conseils en veille juridique et réglementaire dans le domaine Hygiène, Sécurité, Environnement et Développement Durable (HSE-DD), *www.tennaxia.com*.

[2] Le terme endogroupe, issu des sciences cognitives, désigne ici l'ensemble des personnes qui partagent la conceptualisation exprimée dans l'ontologie, et non uniquement celles qui ont participé à sa construction.

[3] Le terme propriété est pris au sens large et inclut les relations unaires (attributs) et binaires.

sentation de la sémantique du domaine.

Cependant, une telle ontologie ne capture pas la totalité des connaissances que les membres de l'endogroupe possèdent sur le domaine. Ainsi, une ontologie ne dit rien quant à la représentativité d'un concept par rapport à son (ou ses) sur-concept(s). Cette notion, nommée *prototypicalité* en psychologie cognitive, est pourtant sous-jacente à toute catégorisation conceptuelle [Rosch (1975)]. Par exemple, en Europe, si les perroquets, les poules et les moineaux sont tous considérés comme des sortes d'oiseaux, le concept de moineau est cependant plus proche conceptuellement de celui d'oiseau que ne le sont ceux de poule ou de perroquet. En d'autres termes, penser à un oiseau nous conduira bien plus volontiers à penser à un moineau qu'à un perroquet ou une poule.

La prototypicalité, comme toute connaissance, est subjective, et peut varier d'un individu à l'autre. Il est cependant possible de bâtir une ontologie au sein d'un endogroupe où il existe un consensus, non seulement sur les hiérarchies de concepts et les propriétés, mais également sur les prototypicalités entre concepts. Mais nous proposons d'exploiter cette notion de prototypicalité pour la personnalisation des ontologies, en considérant que le consensus sur lequel est basé l'ontologie ne porte que sur les concepts, les propriétés, les liens hiérarchiques et les connaissances axiomatiques. Au sein de l'endogroupe, les prototypicalités peuvent donc varier d'un individu à l'autre, ce qui va permettre d'adapter l'ontologie à chaque utilisateur, ou groupe d'utilisateurs. Dans le cadre d'une recherche d'information, par exemple, ces prototypicalités pourront servir à l'extension de requête (la requête est étendue aux concepts les plus prototypiques de ceux qui y apparaissent déjà) ou la personnalisation de la présentation des résultats (les résultats les plus prototypiques sont présentés en premier).

Nous basons en outre nos travaux sur un modèle d'ontologie étendu à l'aspect terminologique. En effet, notre méthode de personnalisation est appliquée principalement dans le cadre de travaux visant à la recherche d'information dans des documents juridiques dédiés à la législation concernant le domaine HSE, où la terminologie métier est riche en synonymies (c'est le cas par exemple des substances chimiques). Cette richesse terminologique doit être représentée dans l'ontologie de manière à permettre aux utilisateurs d'utiliser différents termes pour mener leurs recherches. Nous définissons une *Ontologie Vernaculaire de Domaine* (OVD), pour un domaine D donné et un endogroupe G donné (d'où le qualificatif de vernaculaire), par le tuple suivant :

$$O_{(D,G)} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\right\} \text{ où}$$

- $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont les ensembles de concepts, de propriétés et d'instances des concepts reconnus par tous les membres de l'endogroupe ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés[4] ;
- $dom\ : \mathcal{P} \rightarrow \mathcal{C}$ et $codom\ : \mathcal{P} \rightarrow (\mathcal{C} \cup Datatypes)$ associent à chaque propriété son domaine et éventuellement son co-domaine ;
- $\sigma\ : \mathcal{C} \rightarrow \mathscr{P}(\mathcal{I})$ associe à chaque concept ses instances ;

[4] $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D avec L_C, L_P et L_I les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$, et $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ les fonctions qui associent à chaque concept, propriété ou instance les termes qui les désignent.

Au-dessus des ontologies ainsi définies, nous ajoutons, pour les personnaliser, une couche de connaissances pragmatiques[5], qui varient selon la personne, ou le groupe de personnes, qui utilisent l'ontologie. Ce processus de personnalisation, peut, à partir d'une même OVD, conduire à plusieurs *Ontologies Personnalisées Vernaculaires de Domaine* (OPVD) chacune adaptée à un utilisateur ou groupe d'utilisateurs (cf. figure 1). Ce processus est fondé sur l'apport de ressources supplémentaires :

- un ensemble d'**instances** supposées représentatives de l'univers cognitif de l'utilisateur (dans le cas d'un SI commercial, par exemple, ces instances seront les clients traités par l'utilisateur, les produits qu'il leur vend, etc);

- un **corpus** fourni par l'utilisateur et supposé représentatif de son univers cognitif (ce corpus peut, par exemple, être tiré de documents numériques écrits par l'utilisateur sur un blog ou un wiki sémantique) ;

- des **pondérations portant sur les propriétés** de chaque concept et qui expriment l'importance que l'utilisateur accorde aux propriétés dans la définition du concept.

Ces pondérations sont fixées par l'utilisateur de la façon suivante : pour chaque propriété p de $\mathcal{P}$, l'utilisateur ordonne sur une échelle de 0 à 1 les concepts qui font partie du domaine de p, selon qu'il associe plus ou moins ce concept à cette propriété. Par exemple, pour la propriété "a un auteur", on mettra en premier le concept d'article scientifique, puis celui d'article de presse (pour lequel cette propriété est moins importante), puis celui de mode d'emploi d'aspirateur (pour lequel cette propriété est encore moins importante). D'une certaine façon, l'utilisateur classe les concepts en fonction de leur prototypicalité par rapport à une propriété qu'ils partagent : pour une propriété donnée, il s'agit de savoir quel concept cette propriété évoque le plus souvent.

Bien entendu, ces ressources de personnalisation ne sont pas forcément toujours disponibles, mais la méthode proposée fonctionne même si aucune autre ressource que l'ontologie n'est disponible (dans ce cas, le calcul des prototypicalités ne permet plus une personnalisation, mais constitue simplement un enrichissement de l'ontologie). Ainsi, si l'utilisateur ne souhaite pas pondérer les propriétés, les poids sont tous fixés à 1. Utiliser trois ressources différentes offre ainsi plusieurs façons de personnaliser l'ontologie.

[5] Au sens de la pragmatique linguistique, qui considère le contexte comme indispensable à l'interprétation des textes.

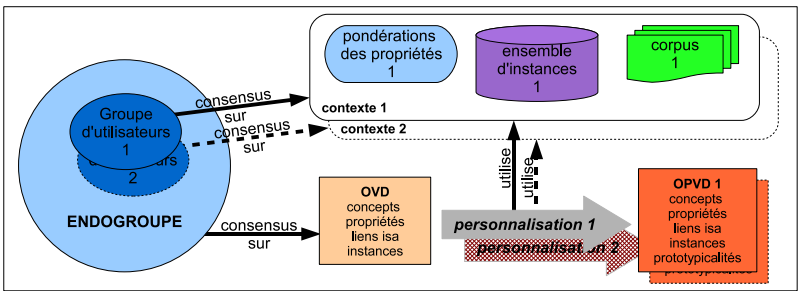


Figure 1: Processus de personnalisation d'une ontologie.

3 Gradients de prototypicalité

Les prototypicalités s'expriment par des gradients numériques qui pondèrent les liens *isa* entre concepts mais également les termes de l'ontologie. Nous distinguons :

- la **prototypicalité conceptuelle** : deux concepts liés hiérarchiquement peuvent être plus ou moins proches sémantiquement. Plus précisément, au sein d'une fratrie de concepts, certains seront plus prototypiques de leur père commun que les autres. Par exemple, parmi tous les modèles d'avion, le modèle le plus représentatif, celui auquel on pense le plus volontiers lorsqu'on pense à un avion, sera plutôt du type des avions commerciaux modernes que du type des premiers biplans ou d'un avion mu par la force musculaire.

- la **prototypicalité lexicale** : pour un concept donné (resp. une propriété) pouvant être désigné par plusieurs termes, certains termes sont utilisés plus volontiers que d'autres. Par exemple, de nos jours, on utilise plus souvent le terme *avion* que les termes *aéroplane* ou *plus lourd que l'air*.

3.1 Gradient sémiotique de prototypicalité conceptuelle

Les ontologies que nous souhaitons personnaliser comportent les trois dimensions introduites par Morris dans sa sémiotique, à savoir le *signifié* (l'intension du concept), le *signifiant* (les termes désignant le concept) et le *référent* (l'extension du concept) [Morris (1938)]. Le gradient de prototypicalité conceptuelle, baptisé *Semiotic-Based Conceptual Prototypicality Gradient* (SPG), exploite ces trois dimensions et comporte (1) une **composante intensionnelle** basée sur la comparaison des intensions des concepts, c'est-à-dire des propriétés attachées aux deux concepts, (2) une **composante extensionnelle** basée sur la comparaison des instances des concepts et (3) une **composante expressionnelle** basée sur la comparaison des expressions des deux concepts au sein d'un corpus, c'est-à-dire le fait que les termes désignant les concepts y soient plus ou moins présents.

Chaque composante du SPG est pondérée de façon à pouvoir moduler, dans le calcul, l'importance des aspects intensionnel, extensionnel et expressionnel au sein de la conceptualisation des utilisateurs. Ces différences d'importance sont conditionnées par

le domaine traité, l'univers cognitif des utilisateurs et le contexte d'utilisation. Ainsi, dans le domaine des mathématiques, les concepts sont plutôt manipulés en intension. Dans le domaine des espèces animales, un zoologue aura tendance à les conceptualiser en intension (par des propriétés biologiques), alors que la plupart des gens utilisent davantage des conceptualisations extensionnelles (basées sur les animaux rencontrés au cours de leur vie).

Formellement, le SPG est une fonction $spg : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ qui, à tout couple de concepts $(c_f, c_p) \in \mathcal{C} \times \mathcal{C}$ tel que $c_f \leq^C c_p$ associe la valeur :

$$spg(c_f, c_p) = \alpha * intension(c_f, c_p) + \beta * extension(c_f, c_p) + \gamma * expression(c_f, c_p)$$

Les fonctions $intension$, $extension$ et $expression$ sont détaillées plus loin. α, β et γ sont des coefficients positifs ou nuls de pondération des composantes. Dans un souci de normalisation, nous imposons que le SPG varie de 0 (représentativité nulle) à 1 (représentativité maximale), que les 3 composantes varient elles aussi entre 0 et 1 et que $\alpha + \beta + \gamma = 1$. Les valeurs de ces 3 coefficients peuvent être fixées arbitrairement ou calibrées par expérimentations. Mais nous proposons une méthode pour les évaluer automatiquement, méthode basée sur le principe suivant. Les rapports entre α, β et γ expriment d'une certaine façon les coordonnées cognitives de l'utilisateur dans le triangle sémiotique (cf. figure 2). De ce fait, il n'est pas possible de fixer en même temps les valeurs des trois rapports (le système d'équations peut être insoluble). Nous avons choisi de calculer les valeurs de γ/α et γ/β, les valeurs α, β et γ étant déduites de ces rapports[6] et de l'équation $\alpha + \beta + \gamma = 1$.

Le rapport γ/α représente le rapport entre ce qui est conceptualisé par l'utilisateur et ce qui est exprimé dans le corpus. En toute généralité, il s'agit donc du rapport entre ce qui est purement intensionnel, c'est-à-dire les concepts de l'ontologie non exprimés dans le corpus, et ce qui est purement expressionnel, c'est-à-dire les termes du corpus désignant des concepts non présents dans l'ontologie. Cependant, nous considérons que l'ontologie couvre bien tout le corpus, c'est-à-dire que tous les termes du corpus renvoient bien à des concepts, relations ou instances de l'ontologie. Aussi, le rapport γ/α va t-il évoluer entre 0 et 1 et il est approximé par le taux de couverture des concepts de l'ontologie par le corpus. Ce taux est égal au nombre de concepts dont au moins un des termes apparait dans le corpus, divisé par le nombre total de concepts. De même, γ/β est approximé par le taux de couverture des instances de l'ontologie par le corpus, qui est égal au nombre d'instances dont au moins un des termes apparait dans le corpus, divisé par le nombre total d'instances.

3.1.1 Composante intensionnelle

La composante intensionnelle du SPG mesure la représentativité d'un concept par rapport à son père en comparant les propriétés qui leur sont rattachées. Il est possi-

[6]Dans le cas où aucun corpus n'est disponible, seul le rapport α/β peut être calculé. Il est évalué par la moyenne, sur l'ensemble des concepts, des rapports entre l'importance des propriétés portées par un concept et le nombre des instances de ce concept. En effet, un concept portant des propriétés marquantes, mais ayant peu d'instances (par exemple un dragon) est conceptualisé davantage de façon intentionnelle (on se souvient des propriétés), alors qu'un concept portant des propriétés banales mais ayant de nombreuses instances (par exemple une voiture) est conceptualisé de façon extensionnelle (on se souvient de la voiture qu'on rencontre le plus souvent).

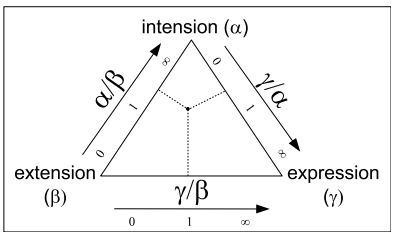


Figure 2: Les coefficients de pondération des composantes du SPG comme coordonnées dans le triangle sémiotique. γ/α proche de 0 indique que l'utilisateur a une approche beaucoup plus intensionnelle qu'expressionnelle du domaine, le même rapport proche de l'infini indique le contraire, et le même rapport égal à 1 indique un équilibre entre les approches intensionnelle et extensionnelle. La même interprétation est adoptée pour les autres rapports. Quand les trois approches sont équilibrées, on a $\alpha = \beta = \gamma = 1/3$, les trois rapports sont égaux à 1 et les coordonnées cognitives de l'utilisateur correspondent au barycentre du triangle sémiotique.

ble de calculer la composante intensionnelle comme rapport entre le nombre de propriétés ajoutées par le concept fils et le nombre de propriétés totales du fils [Aimé *et al.* (2008)] (un concept est ainsi vu comme d'autant plus représentatif de son père qu'il ajoute moins de propriétés à son intension). Mais la méthode de calcul utilisée ici s'inspire de [Au Yeung & Leung (2006)] et s'appuie sur la représentation des concepts par des vecteurs dans l'espace des propriétés de l'ontologie. Le principe consiste à calculer dans cet espace un vecteur prototype du concept père c_p et la prototypicalité du concept fils c_f est la distance euclidienne entre le vecteur représentant le fils et le vecteur prototype du père. Cependant, [Au Yeung & Leung (2006)] donne comme coordonnées des vecteurs des valeurs de vérité floues, alors que nos coordonnées sont des valeurs mesurant l'importance de la propriété pour le concept. Formellement, à tout concept $c \in \mathcal{C}$, est associé le vecteur $\overrightarrow{v_c}= (v_{c1}, v_{c2}, ..., v_{cn})$ avec $n = |\mathcal{P}|$ et $v_{ci} \in [0, 1], \forall i \in [1, n]$. v_{ci} est la pondération fixée par l'utilisateur pour le concept c par rapport à la propriété i (v_{ci} vaut 1 si l'utilisateur n'a pas fixé ces pondérations).

Le vecteur prototype d'un concept c_p a été originellement introduit dans [Au Yeung & Leung (2006)] comme une moyenne des vecteurs des concepts fils de c_p. Cependant, [Au Yeung & Leung (2006)] ne prend en compte dans la moyenne que les concepts qui héritent directement de c_p, alors que nous étendons le calcul à tous les concepts de la descendance. En effet, des propriétés qui apparaissent uniquement sur des descendants indirects du concept père peuvent pourtant apparaitre dans le prototype du père, en particulier si l'aspect intensionnel est important. Par exemple, dans le cas du concept *chercheur*, le fait d'avoir une blouse blanche n'est pas une propriété du concept, mais peut très bien apparaître dans le prototype du concept. Le vecteur prototype p_{c_p} est donc un vecteur dans l'espace des propriétés, où l'importance de la propriété i est la moyenne des importances des propriétés des concepts de la descendance de c_p possédant i. Si pour $i \in \mathcal{P}$, $S_i(c) = \{c_j \leq^C c, c_j \in dom(i)\}$ alors

$$\overrightarrow{p_{c_p}}\,[i] = \frac{\sum_{c_j \in S_i(c_p)} \overrightarrow{v_{c_j}}[i]}{|S_i(c_p)|}$$

La composante intensionnelle est donc $intension(c_f, c_p) = 1 - d(\overrightarrow{v_{c_f}}, \overrightarrow{p_{c_p}})$ où d est la distance euclidienne normée dans l'espace des propriétés.

3.1.2 Composante extensionnelle

La composante extensionnelle du SPG mesure la représentativité d'un concept par rapport à son père en évaluant la place relative occupée par les instances de ce concept dans l'extension du concept père : plus l'extension du fils a d'importance au sein de l'extension de son père, plus le fils est prototypique de son père. Par exemple, quelqu'un qui possède une douzaine de chats trouvera ce félin plus prototypique du concept d'animal domestique que quelqu'un qui possède un poisson rouge. Le calcul de cette composante suppose que les concepts considérés possèdent des instances. Pour le calcul, toutes les instances des concepts sont prises en compte, celles de l'OVD et celles ajoutées par l'utilisateur. Nous utilisons une forme logarithmique, de manière à obtenir un comportement de la composante proche de l'évaluation humaine (les prototypicalités des concepts ayant très peu d'instances ne sont pas trop proches de 0). La composante extensionnelle est ainsi donnée par :

$$extension(c_f, c_p) = 1 / \Big(1 - \log\big(\tfrac{|\sigma(c_f)|}{|\sigma(c_p)|}\big)\Big)$$

3.1.3 Composante expressionnelle

La composante expressionnelle du SPG mesure la représentativité d'un concept c_f par rapport à son père c_p en comparant leurs expressions : plus un concept est exprimé, plus il sera prototypique de son père. Une première mesure de l'expression d'un concept est donnée par le nombre de termes qui le désignent. Ainsi, plus le nombre de termes désignant un concept est grand, plus ce concept occupe de place dans l'univers cognitif de l'utilisateur. Par exemple, le concept de cheval, possédant de nombreux synonymes (bourrin, canasson, dada, ...), est plus prototypique du concept animal que le concept de raton-laveur, qui n'a pas de synonyme. Cette première mesure de l'expression de c_f relativement à ses frères ne dépend que de l'OVD et est donnée par le rapport entre le nombre de termes désignant c_f et le nombre maximum de termes désignant les fils directs de c_p :

$$expression_{OVD}(c_f, c_p) = |term_c(c_f)| \,/ \max_{c_i \leq^C c_p, \nexists c_j, c_i \leq^C c_j \leq^C c_p} (|term_c(c_i)|)$$

Cette mesure repose sur les termes fixés dans l'OVD et est donc la même pour tous les utilisateurs. Si l'utilisateur fournit un corpus, il est possible de l'utiliser pour personnaliser le calcul de cette composante expressionnelle selon le principe suivant : plus les termes de c_f ou de ses descendants sont présents dans le corpus, plus c_f est exprimé dans l'univers cognitif de l'utilisateur, et plus il est prototypique de c_p. La prégnance d'un concept dans le corpus dépend du nombre d'occurences des termes désignant le concept ou un de ses fils, rapporté au nombre total de termes du corpus. Les occurences

sont de plus pondérées en fonction de la structure du document où elles apparaissent. Par exemple, une occurence apparaissant dans un titre ou dans une liste de mots-clés aura plus de poids qu'une occurence située à l'intérieur d'un paragraphe. Nous voulons également tenir compte dans le calcul de la prégnance du nombre de documents dans lesquels les occurences apparaissent, car un terme qui apparait souvent mais dans un nombre très réduit de documents doit avoir une prégnance moins élevée qu'un terme présent peu de fois dans chaque document mais de façon uniforme dans la majorité des documents du corpus. La fonction $pregnance_t(t) : L_C \rightarrow [0,1]$ donnant la prégnance d'un terme est définie comme suit :

$$pregnance_t(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}}$$

où $count_{occ}(t)$ est le nombre d'occurrences pondérées de t dans les documents du corpus, $count_{doc}(t)$ est le nombre de documents du corpus où *t* apparaît, N_{occ} est la somme de tous les nombres d'occurrences pondérées de tous les termes contenus dans le corpus et N_{doc} est le nombre total de documents du corpus.

Finalement, la fonction $pregnance_c(c)$ définie sur $\mathcal{C}$ et donnant la prégnance d'un concept est définie comme suit ($S_{term}(c)$ est l'ensemble des termes désignant c ou un des concepts de sa descendance) :

$$pregnance_c(c) = \sum_{t \in S_{term}(c)} pregnance_t(t)$$

La composante expressionnelle vaut $expression(c_f, c_p) = expression_{OVD}(c_f, c_p) \times \frac{pregnance_c(c_f)}{pregnance_c(c_p)}$ ou bien $expression(c_f, c_p) = expression_{OVD}(c_f, c_p)$ si aucun corpus n'est fourni par l'utilisateur.

3.2 Gradient de prototypicalité lexicale

Le gradient de prototypicalité lexicale, noté LPG (*Lexical Prototypicality Gradient*), évalue, pour un concept donné et un terme le désignant, la représentativité de ce terme pour désigner ce concept, dans l'univers cognitif du groupe d'utilisateurs pour lequel on veut adapter l'ontologie. Le calcul du LPG repose, comme celui de la composante expressionnelle du SPG, sur l'utilisation d'un corpus représentatif du groupe en question. Le principe du calcul est que plus le rapport entre le nombre d'apparitions du terme et le nombre d'apparitions d'un des termes utilisés pour désigner le concept est proche de 1, plus le terme est prototypique, au sens lexical, de ce concept. Comme pour le calcul de la composante expressionnelle du SPG, les occurences des termes sont pondérées selon la place qu'ils occupent dans les documents et leur comptage dépend de leur répartition dans les documents. La fonction $lpg(t,c) : L_C \times \mathcal{C} \rightarrow [0,1]$, définie pour tout couple (t,c) où c est le concept désigné par t, est donnée par :

$$lpg(t,c) = 1/\Big(1 - \log\big(\frac{pregnance_t(t)}{\sum_{m \in term_c(c)} pregnance_t(m)}\big)\Big)$$

3.3 Facteur émotionnel

Des travaux en psychologie cognitive ont montré que l'état émotionnel d'une personne influe sur sa perception des catégories d'objets : plus on est stressé, plus notre esprit est concentré sur les objets les plus proches cognitivement de ceux qui nous occupent, et inversement [Mikulincer *et al.* (1990)]. Nous introduisons donc un facteur émotionnel qui nous permet de moduler les gradients eux-mêmes en fonction de l'état d'esprit de l'utilisateur. Ce facteur émotionnel est modélisé par un coefficient δ qui peut varier entre 0 et 1 pour un état d'esprit ouvert et entre 1 et ∞ pour un état d'esprit fermé. Les valeurs des SPG et des LPG sont élevées à la puissance $1/\delta$, ce qui a pour effet, en cas d'état mental fermé de l'utilisateur, de réduire fortement les prototypicalités qui sont déjà faibles, et en cas d'état mental ouvert, d'augmenter les prototypicalités faibles.

En recherche d'information, l'extension de requête par ajout des concepts les plus prototypiques de ceux spécifiés par l'utilisateur constitue un exemple d'utilisation de ce facteur émotionnel. Dans le cas où l'utilisateur effectue une recherche très ouverte (par exemple, il ne sait pas exactement ce qu'il cherche), le coefficient δ sera positionné à une valeur proche de 0 et le nombre de concepts ajoutés sera important. Au contraire, s'il souhaite limiter les résultats de sa requête (par exemple, s'il est pressé) le coefficient δ sera positionné à une valeur élevée, ce qui restreindra le nombre de concepts ajoutés à la requête.

4 Expérimentations

TOOPRAG (*A Tool dedicated to the Pragmatics of Ontology*) est un outil dédié au calcul automatique de nos gradients. Cet outil, implémenté en Java 1.5, utilise les bibliothèques Lucène (librairie d'indexation et de recherche full-text, *lucene.apache.org*) et Jena (framework permettant la prise en charge d'ontologies OWL et incluant un moteur d'inférence, *jena.sourceforge.net*). Il prend en entrée (1) une ontologie représentée en OWL 1.0, où chaque concept et propriété est associé à un ensemble de termes et (2) un corpus composé de fichiers au format texte. Le corpus est indexé à l'aide de Lucène, puis TOOPRAG calcule les valeurs de SPG des liens *is-a* entre concepts et les valeurs des LPG de tous les termes utilisés pour dénoter les concepts et les propriétés. L'OPVD résultante est stockée dans un format OWL étendu par rapport aux spécifications de OWL 1.0. Une valeur de LPG est représentée par un nouvel attribut *xml:lpg*, directement associé à la primitive *rdfs:label* et une valeur de SPG est représentée par un nouvel attribut *xml:spg*, directement associé à la primitive *rdfs:subClassOf*.

Une ontologie du domaine HSE a été réalisée dans le cadre du projet REDENE10[7]. Elle comprend 10000 concepts (organisés au sein d'un treillis de profondeur 12 et de largeur maximale 1500) et 20 relations (formant un arbre de profondeur 3). Le corpus utilisé est composé de 1100 textes réglementaires, avec un taux de couverture global (intentionnel + extensionnel) de 13%. La valeur moyenne des SPG sur les liens *isa* de l'ontologie est de 0.128 et 30,2% des valeurs de SPG sont non nulles avec la distribu-

[7]Cette ontologie est propriété de Tennaxia - tous droits réservés – dépôt INPI N°322.408, 13 juin 2008 – dépôt *Scam Vélasquez* N°2008090075, 16 septembre 2008.

tion suivante :

$[0, 0.01[$	$[0.01, 0.125[$	$[0.125, 0.25[$	$[0.25, 0.5[$	$[0.5, 0.75[$	$[0.75, 1[$	1
63.23%	15.73%	5.11%	4.97%	3.22%	2.96%	3.34%

Les valeurs obtenues ont été validées par les experts HSE de Tennaxia et reflète bien les représentativités des concepts et des termes dans l'ontologie. Ces mesures peuvent donc être exploitées dans le système de recherche personnalisée d'information qui est en cours de développement.

5 Conclusion

Nous proposons dans cet article de baser la personnalisation d'un SI utilisant une ontologie sur l'ajout d'une couche de connaissances pragmatiques au-dessus de l'ontologie. Ces connaissances pragmatiques sont exprimées par des gradients de prototypicalité qui pondèrent liens *isa* et les termes associés aux primitives conceptuelles. D'autres travaux ont introduit la notion de prototypicalité en ingénierie des ontologies, en particulier [Au Yeung & Leung (2006)], qui calcule les prototypicalités conceptuelles à partir des propriétés des concepts. Cependant, ils pondèrent les propriétés avec des valeurs de vérité floues, ce qui n'est pas cohérent avec la sémantique de la plupart des ontologies. Nos pondérations des propriétés respectent cette sémantique, en ajoutant simplement aux ontologies existantes des connaissances exprimant des prototypicalités. D'autre part, nous étendons la méthode de calcul de [Au Yeung & Leung (2006)] en prenant en compte les propriétés de toute la descendance des concepts considérés, et non uniquement celles des sous-concepts directs, ce qui est plus fidèle aux phénomènes cognitifs.

Concernant le calcul de la composante expressionnelle du SPG, la formule que nous proposons est proche de la mesure conceptuelle introduite par Resnik et fondée sur la notion de *contenu en information* [Resnik (1995)]. Cependant, alors que Resnik considère le corpus comme un tout, nous exploitons la granularité du corpus en tenant compte du nombre de documents où apparaîssent les termes. En effet, nous considérons que si beaucoup de documents contiennent quelques occurences d'un terme, le concept désigné par ce terme est davantage exprimé dans le corpus que si un faible nombre de documents contiennent de nombreuses occurences du terme. D'autre part, nous tenons compte de la structure des documents en pondérant les occurences suivant qu'elles apparaîssent dans un titre, un résumé, etc.

Notre principale contribution consiste à faire reposer le calcul des gradients, et donc la personnalisation, sur plusieurs types de ressources : l'ontologie elle-même, mais également un corpus de textes et un ensemble d'instances. Les calculs peuvent se faire à partir de la seule ontologie (mais on ne peut alors parler de personnalisation, uniquement de prototypicalisation), ou n'utiliser en plus que le corpus ou que les instances, ce qui offre plus de possibilités en terme d'applications. Par exemple, dans le cas d'une personnalisation au sein d'un site web collaboratif, sur lequel chaque utilisateur dépose des textes, il est possible d'utiliser ces textes pour la personnalisation, sans disposer d'instances propres à chaque utilisateur. Par contre, dans le cas du projet REDENE10,

un ensemble de textes réglementaires peut être commun à un groupe de consultants, mais chaque consultant a en charge ses propres clients et ses propres installations industrielles, qui sont autant d'instances permettant la personnalisation.

La personnalisation par les prototypicalités peut être exploitée en recherche d'information pour étendre des requêtes aux concepts les plus prototypiques de ceux qui apparaîssent dans la requête. Il est également possible de présenter les résultats de la recherche par ordre de prototypicalité, et de filtrer les résultats pour en éliminer ceux qui sont trop peu prototypiques. Une autre application possible de la prototypicalité est liée à la validation d'ontologie. Le calcul des gradients peut révéler des valeurs de prototypicalité très faibles pour un concept et son père, ce qui peut indiquer que le lien *isa* entre les deux n'est pas fondé, ou qu'il conviendrait d'introduire un concept supplémentaire entre les deux pour structurer davantage l'ontologie.

Le calcul des gradients peut en outre être généralisé au cas de deux concepts se trouvant sur une même branche de la hiérarchie mais non directement liés par *isa*. Une valeur élevée de prototypicalité indiquerait alors que le concept le plus bas est mal placé, et doit être remonté dans la hiérarchie. De façon encore plus générale, il est possible de calculer la prototypicalité entre deux concepts qui n'héritent pas du tout l'un de l'autre (la composante intensionnelle ne change pas, mais les composantes extensionnelle et expressionnelle seront différentes). Dans ce cas, des valeurs de prototypicalité élevées peuvent indiquer qu'un lien *isa* aurait dû être spécifié entre les concepts. De manière générale, les gradients de prototypicalité représentent une mesure sémantique de parentalité entre concepts, qui peut être exploitée pour la classification.

Nos travaux sur les gradients de prototypicalité trouvent également une autre application avec la définition d'une mesure de similarité entre concepts fondée sur la prototypicalité et sur notre méthode de calcul à base sémiotique.

References

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2008). Conceptual and lexical prototypicality gradients dedicated to ontology personalisation. In *Proceedings of the 7th International Conference on Ontologies Databases and Applications of Semantics (ODBASE'2008)*, volume 5332, p. 1423–1439: Springer Verlag - LNCS.

AU YEUNG C. M. & LEUNG H. F. (2006). Ontology with likeliness and typicality of objects in concepts. In S. B. . HEIDELBERG, Ed., *Proceedings of the 25th International Conference on Conceptual Modeling (ER'2006)*, volume 4215/2006.

BRUSILOVSKY P. & KOBSA A. (2007). *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer.

HARNAD S. (2003). Categorical perception. *Encyclopedia of Cognitive Science*, **LXVII**(4).

MIKULINCER M., KEDEM P. & PAZ D. (1990). The impact of trait anxiety and situational stress on the categorization of natural objects. *Anxiety Research*, **2**, 85–101.

MORRIS C. (1938). *Foundations of the Theory of Signs*. Chicago University Press.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, volume 1, p. 448–453.

ROSCH E. (1975). Cognitive reference points. *Cognitive Psychology*, (7), 532–547.

Connaissances opérationnelles pour la conception automatique de légendes de cartes

Catherine Dominguès, Sidonie Christophe, Laurence Jolivet

Laboratoire COGIT – Institut Géographique National, 73 avenue de Paris
F-94165 Saint-Mandé cedex
http://recherche.ign.fr/labos/cogit/
{catherine.domingues, sidonie.christophe,
laurence.jolivet}@ign.fr

Résumé : La réalisation d'une légende de carte topographique est un processus qui fait appel à de nombreux domaines de connaissances. L'objectif du travail présenté est de permettre à des utilisateurs de créer des cartes sur mesure, en particulier de les aider à concevoir des légendes cartographiquement correctes et adaptées à leurs goûts et à leurs besoins. Pour cela, il est nécessaire de lister et de formaliser les connaissances opérationnelles intervenant dans la représentation de données géographiques. L'utilisation de la couleur dans la conception de la symbolisation est en particulier étudiée. L'ontologie Ontocarto décrit des concepts cartographiques ; elle est notamment liée à des contraintes cartographiques. Sont présentées deux applications d'aide à la conception automatique de légendes qui exploitent cette ontologie et les contraintes associées à ses concepts.

Mots-clés : modélisation des connaissances, cartographie, sémiologie graphique, ontologie, dialogue coopératif

1 Introduction

Un randonneur qui place, sur des données topographiques, un itinéraire et des étapes personnalisés (restauration, hôtellerie, sites remarquables) pour créer sa carte de randonnée ; un bureau d'études qui ajoute à des données topographiques les bassins versants et des informations sur les risques d'inondation : l'explosion des outils cartographiques sur Internet[1] ainsi que la profusion des systèmes d'information géographique offrent à tous des moyens techniques pour réaliser des cartes. Cependant la qualité des cartes produites n'est souvent pas bonne : problèmes de lisibilité, piètre rendu esthétique, problèmes de compréhension du message censé être délivré par la carte sont les critiques majeures relevées par les cartographes professionnels. En effet, la majorité de ces outils superposent des symbolisations associées par défaut aux données de l'utilisateur sans vérifier la cohérence de ces superpositions, ou bien proposent des palettes de couleurs définies sans tenir compte du contexte d'utilisation de la carte. La conception d'une carte reste un processus

[1] Géoportail : http://www.geoportail.fr/ ; Google Maps : http://maps.google.fr/, …

complexe à la fois *technique*, soutenu par une théorie cartographique particulièrement riche (Bertin 1967, Béguin & Pumain 2000, Cuenin 1972, Robinson 1952, Monmonnier 1991), et *créatif*, requérant des aptitudes artistiques (Krygier 1995, Rekacewicz 2006).

Dans ce contexte, nos travaux s'attachent à proposer un modèle de conception cartographique en vue de créer des outils d'aide à la conception automatique de *carte sur mesure*, quelque soit le niveau d'expertise des utilisateurs. La réalisation d'une carte par un utilisateur répond à son besoin, en vue d'un objectif, de visualiser des informations géoréférencées. Cette réalisation intègre donc le besoin de l'utilisateur (l'objectif de sa carte), ses goûts (préférences esthétiques et culturelles) et son ressenti quant à l'aspect visuel de la carte, tout en respectant les règles de sémiologie graphique.

Le processus de réalisation cartographique respecte un processus logique (Bucher 2007) que traduit la figure 1.

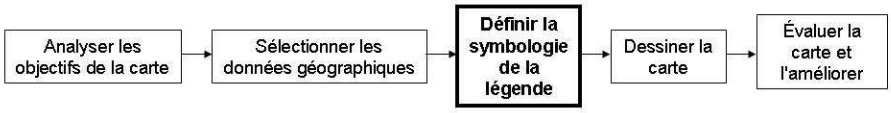


Fig. 1 – Étapes du processus de construction d'une carte

Dans cet article nous nous intéressons à l'étape de définition de la symbologie de la légende. En effet, la légende synthétise les contraintes et les choix concernant la représentation cartographique des données sélectionnées. Il s'agit dans un premier temps d'expliciter les connaissances opérationnelles pour la définition de la symbologie. Nous proposons ensuite de les formaliser et d'organiser les concepts sous la forme d'une ontologie cartographique et de contraintes associées. Enfin, nous présentons des applications de conception de cartes sur mesure qui exploitent ces connaissances de façon automatique.

2 Connaissances pertinentes pour la construction de la légende

La construction de la légende s'appuie sur des connaissances pléthoriques et partiellement formalisées. Après avoir précisé la définition de la légende d'une carte, nous justifions le choix de l'étude de la couleur dans la symbolisation.

2.1 La définition de la légende d'une carte géographique

La légende est vue ici comme un ensemble structuré de lignes de légendes où un symbole est associé à des types d'objets géographiques à représenter. Son contenu et son organisation sont régis par les principes de la sémiologie graphique formalisés par Bertin (1967). La sémiologie graphique dicte les règles de construction d'un système de signes qui garantit que la traduction graphique du message censé être délivré par la carte est compréhensible par tous. Cette traduction graphique s'appuie sur les

variables visuelles qui sont des variations de figurés telles la *taille*, la *forme*, la *teinte*, la *valeur* des symboles utilisés dans la légende (Robinson 1952).

L'influence des choix de symbolisation est illustrée par la figure 2 : les deux extraits de cartes ont été réalisés en appliquant des symbolisations différentes aux mêmes données géographiques. Les rendus visuels diffèrent pourtant selon l'intensité des couleurs et la nature des contrastes notamment entre les routes et le fond de carte.

Fig. 2 – Symbolisations de cartes topographiques autrichienne et estonienne appliquées aux mêmes données géographiques (Jolivet 2009).

Plus généralement, l'appréciation des teintes et valeurs, de leurs contrastes et de leur harmonie varient selon les compétences physiologiques, mais surtout la culture, l'expérience et les goûts de l'utilisateur. Le choix des dessins des symboles, des polices de caractères, des épaisseurs de traits, des aplats de couleur relèvent aussi des sensibilités et des compétences individuelles. L'ensemble de ces impressions forment le ressenti de l'utilisateur qui rend compte finalement de sa satisfaction quant à la carte qu'il a construite.

2.2 Utilisation de la couleur dans la construction de la légende

Parmi les variables visuelles définies par la sémiologie graphique, la couleur joue un rôle particulier et prédominant. Pour Bertin (1967) elle "*exerce une indéniable attraction psychologique. ... [Elle] retient l'attention, multiplie le nombre de lecteurs, assure une meilleure mémorisation et en définitive augmente la portée du message*". Nous avons donc choisi de nous concentrer sur l'utilisation de la couleur, i.e. des seules variables visuelles *teinte* et *valeur*, dans le choix de la symbologie. Cependant, Cuenin (1972) insiste sur *"[ses] règles d'utilisation complexes et controversées car elles ne réalisent généralement qu'un compromis plus ou moins bien équilibré entre divers facteurs souvent contradictoires qui sont d'ordre physique, physiologique, subjectif, symbolique et esthétique*".

Les teintes et les valeurs permettent de structurer la légende. Comme le montre la figure 3, la légende est ordonnée en *thèmes* (par exemple *bâti*) et les *thèmes* en *lignes de légende* (dans cet exemple : *bâtiments administratifs*, *équipements sportifs*, *logements sociaux*, etc). Le regroupement en thèmes met en évidence les relations entre les objets. Les variables visuelles teinte et valeur permettent de traduire, entre les symboles de la légende, les relations existant entre les objets géographiques. Ceux appartenant au même thème sont liés par une relation d'association : ils sont

représentés par des symboles de teintes proches : ici, la gamme de bleus est dédiée aux symboles d'hydrographie. Ceux liés par une relation de différenciation sont représentés par des symboles de teintes éloignées. Ceux liés par une relation d'ordre sont représentés par des symboles ordonnés : variation d'intensité des bleus pour les symboles de l'aléa torrentiel.

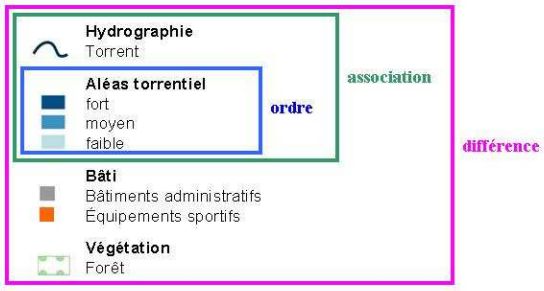


Fig. 3 – Traduction dans la légende des relations existant entre les objets cartographiés (Chesneau 2006)

Notre objectif initial, aider un utilisateur à construire une légende correcte et adaptée à ses besoins, est alors reformulé de la manière suivante : associer une famille de couleurs à chaque thème de la légende.

Des travaux ont été menés récemment sur la perception de la couleur et son utilisation dans les cartes. Ils sont fondés sur les travaux antérieurs d'Itten (1985) qui a étudié les différents types de contrastes colorés. Brewer (2003) a proposé des palettes de couleurs adaptées à la traduction des relations entre les objets géographiques. Chesneau (2006) a créé un cercle chromatique (représenté en fig. 8) qui s'appuie sur les variables visuelles de *teinte* et *valeur* et les contrastes entre teintes et entres valeurs *(*Buard & Ruas 2007) pour organiser une palette finie et cohérente de 156 couleurs. Dans nos applications (cf. §4), nous nous appuyons sur ce système de couleurs de référence pour modéliser nos raisonnements.

3 Organisation des connaissances

Nous présentons dans ce paragraphe l'ontologie Ontocarto associée à un ensemble de contraintes que nous avons créées afin de formaliser les connaissances opérationnelles nécessaires à la définition de la symbologie de la légende.

3.1 L'ontologie Ontocarto

Des ontologies géographiques ont été proposées afin de décrire les besoins spécifiques en termes de modélisation et d'interrogation de l'information

géographique. Ces ontologies doivent intégrer des relations et des concepts spatiaux afin de permettre des raisonnements spécifiques (Cullot et al. 2003).

A l'instar de Iosifescu-Enescu & Hurni (2007), nous proposons de distinguer les concepts pertinents pour la représentation cartographique des règles d'utilisation de ces concepts. Nous souhaitons construire des légendes qui respectent les règles de sémiologie graphique. Dans cet objectif, Dao (2004) a proposé un système automatique qui vise à interdire les représentations non valides. Nos propositions de légendes sont fondées sur l'organisation de la légende en thèmes et le respect des relations entre thèmes.

Le consensus préalable à la construction de l'ontologie et la formalisation induite par sa construction garantissent la cohérence de l'ensemble des concepts et leur évolution. En particulier, la formalisation des concepts liés à la description des différents points de vue sur la couleur est un point d'intégration crucial des différentes contributions. La formalisation dans l'ontologie du cercle chromatique et de l'organisation de la légende est ensuite fondée sur cette notion fondamentale. Ces concepts sont présentés dans les paragraphes suivants.

3.1.1 Différents points de vue sur la couleur

L'acception courante du mot *couleur* est trop polysémique et imprécise pour être utilisable dans l'ontologie. Cependant, les connaissances esthétiques et symboliques concernant les couleurs utilisent ce terme en ignorant son imprécision. Par exemple, la littérature utilise des termes simples comme *vert* ou composés comme *couleur jaune* pour définir des propriétés : *la couleur jaune est associée au mensonge* (Pastoureau 2005) ou *le vert représente la chance, et aussi la malchance* (Pastoureau 2005) sans préciser l'intensité concernée. Le concept *CouleurGénérique* rend compte de cette acception courante et permet de traduire ces propriétés dans l'ontologie. Les instances de *CouleurGénérique* sont les couleurs désignées par (Mollard-Desfour 1998) : *bleu*, *blanc*, *brun* [syn. *: marron*], *gris*, *jaune*, *noir*, *orange*, *rose*, *rouge*, *vert*, *violet.*

En sémiologie graphique, la notion de couleur renvoie à deux variables visuelles dont les concepts sont retenus dans l'ontologie : *Teinte* et *Valeur* sont subsumés par *VariableVisuelle*. Selon les corpus, le terme *couleur* renvoie à la teinte ou à sa valeur. D'autres codages de couleurs sont référencés dans l'ontologie : *CodageRGB*[2], *CodageTSL*[3], *CodageCIELab*4 ainsi que les algorithmes de passage d'un codage à l'autre pour les couleurs descriptibles dans plusieurs codages.

Le concept *CouleurPhysique* renvoie à la définition physique d'une couleur, c'est à dire une longueur d'onde, un intervalle de longueurs d'ondes (pour une couleur monochromatique) ou une union d'intervalles de longueurs d'ondes (pour les couleurs usuelles non monochromatiques).

---

[2] PASCALE D. A review of RGB color spaces ... from xyY to R'G'B'
http://www.babelcolor.com/download/A%20review%20of%20RGB%20color%20spaces.pdf

[3] MUNSELL Color Co (1943). Munsell book of color, Baltimore. Edité par Luneau Industrie. Paris

[4] http://www.fho-emden.de/~hoffmann/cielab03022003.pdf

3.1.2 Cercle chromatique

Le cercle chromatique de Chesneau (2006), représenté dans la figure 8, est composé de douze teintes (concept *CercleChromatiqueTeinte* équivalent au concept *CercleChromatiqueQuartier)* dont la valeur varie en sept paliers (concept *CercleChromatiqueValeur* équivalent au concept *CercleChromatiqueSecteur)*. A ce cercle principal sont ajoutés deux cercles de couleurs grisées et de gris colorés adaptées à la cartographie de données d'importance secondaire. Le concept *CercleChromatique* est relié à *CercleChromatiqueQuartier* par une relation *estComposéDe* ; une même relation *estComposéDe* relie *CercleChromatiqueQuartier* et *CercleChromatiqueSec-teur*. Afin de pouvoir leur associer les mêmes propriétés qu'à une *CouleurGénérique*, *CercleChromatiqueQuartier* et *CercleChromatiqueCouleur* sont subsumés par *CouleurGénérique*. La figure 4 montre l'organisation permettant d'articuler les concepts liés à la couleur et ceux du cercle chromatique.

Une famille de couleurs est une construction fondée sur l'acception du langage courant, par exemple : la *famille des bleus*. Elle permet de regrouper sous le même terme différentes intensités d'une même teinte ou de teintes visuellement proches. Dans l'ontologie, le concept *FamilleDeCouleurs* est relié au concept *Cercle-ChromatiqueTeinte* par une relation *EstComposéDe*.

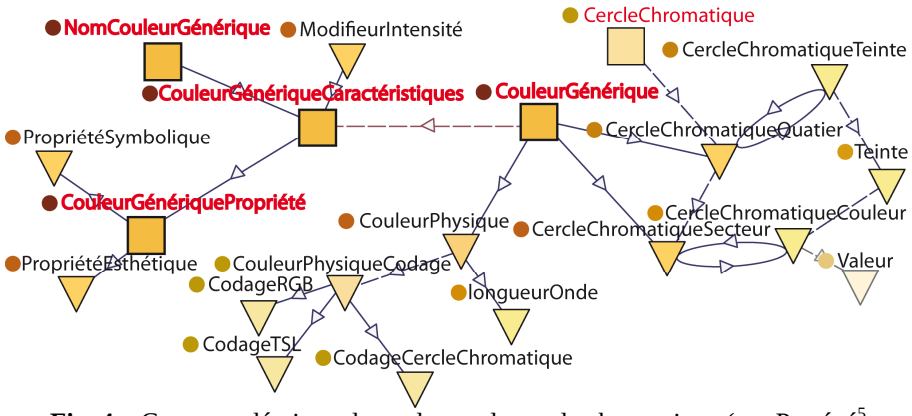


Fig. 4 – Concepts décrivant la couleur et le cercle chromatique (par Protégé[5]-Jambalaya)

3.1.3 Organisation de la légende

La description de la légende dans l'ontologie reprend l'organisation décrite en 2.2. La ligne de légende (concept *LégendeLigne*) associe un symbole (concept *Légende-Caisson*) à la description d'un objet géographique (concept *LégendeLibellé*). La figure 5 montre cette organisation.

[5] http://protege.stanford.edu/

Les relations d'association, différenciation et ordre, qui existent entre thèmes ou à l'intérieur d'un thème, sont traduites par des contrastes entre teintes et/ou valeurs. Par exemple, le concept *ContrasteDeTeinte* (subsumé par *Constraste*, un type particulier de *ConceptPerceptionVisuelle*) qui a une intensité *fort* est relié par la relation *traduit* au concept *Différenciation* (subsumé par *RelationLégende* puis *ConceptSémiologie-Graphique*). La figure 6 montre l'organisation des différents contrastes.

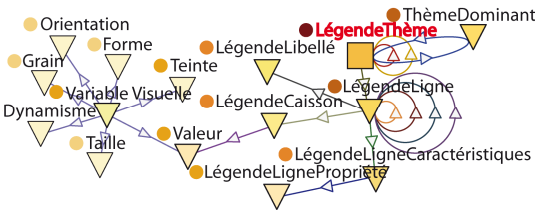


Fig. 5 – Concepts décrivant l'organisation de la légende

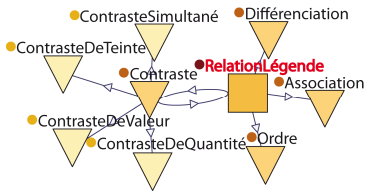


Fig. 6 – Concepts décrivant l'organisation des contrastes

Comme (Kurokawa & Ota, 2007), nous avons associé des qualificatifs que nous avons étudiés (Dominguès 2006) à des lignes de légende. Par exemple, la *LégendeLigne* qui associe l'instance de *LégendeLibellé*, *bâti,* à une couleur de la *FamilleDeCouleursDesJaunes* est qualifiée de *originale* et peut donner une impression de *gaîté* (concept *Qualificatif* subsumé par *CarteCaractéristiques* relié à *Carte* par la relation *admetPourCaractéristiques*) à la carte. Une *LégendeLigne* par laquelle un objet du *Thème hydrographie* est cartographié par une couleur de la *FamilleDeCouleursDesJaunes* est qualifiée de *très original*, voire *impossible*.

3.2 Les contraintes associées à l'ontologie Ontocarto

La théorie cartographique telle qu'elle est définie par les auteurs cités précédemment (cf. § 2) décrit des recommandations que nous traduisons par des contraintes sur les concepts d'Ontocarto. Dans un premier temps, nous avons choisi de travailler sur les trois catégories de contraintes suivantes :

- contraintes traduisant l'utilisation conventionnelle des couleurs (concept *CouleurGénérique*). Les objets géographiques appartenant au thème *hydrographie* sont cartographiés avec un bleu ("couleur" définie par une teinte et une valeur et appartenant à la *FamilleDeCouleursDesBleus*). De même, les objets du thème *végétation* sont cartographiés avec un vert de la *FamilleDeCouleursDesVerts* ;
- contraintes traduisant les relations entre thèmes à l'aide des teintes. Les objets appartenant au même thème sont cartographiés avec des couleurs de même teinte ou formant un *ContrasteDeTeinte faible*. Les objets appartenant à des thèmes différents sont cartographiés avec des couleurs formant un *ContrasteDeTeinte fort* ;

- contraintes exploitant les contrastes de valeurs. Ces contrastes permettent de traduire des relations d'ordre entre données du même thème. Les superficies relatives des objets influent également sur les choix des valeurs. Par exemple, le fond cartographique, qui occupe en général la plus grande superficie de la carte, doit être contrasté avec les objets qui s'y superposent. Pour cela, la couleur du fond doit être de faible valeur (par exemple les deux premiers niveaux de valeur des quartiers du cercle chromatique) et les objets de valeur plus foncée (les intensités 5, 6 et 7 du cercle chromatique).

Nos systèmes automatiques intègrent ces différentes catégories de contraintes pour faire des propositions de légendes.

4 Exploitation de l'ontologie à travers des applications

L'organisation des connaissances nécessaires à la représentation cartographique dans notre ontologie Ontocarto permet d'utiliser ces connaissances dans des systèmes automatisés de création de légendes de cartes sur mesure. Nous illustrons cette utilisation à travers deux applications que nous avons développées. La première interprète le besoin et les goûts de l'utilisateur pour concevoir automatiquement une légende personnalisée, et est déployée sous forme de services Web. La seconde est un système coopératif de conception de légendes innovantes. Ces applications reposent sur la gestion des contraintes associées à l'ontologie Ontocarto.

4.1 Service web de conception automatique de légendes

Permettre à un utilisateur de créer une carte sur mesure sur le Web implique de générer une légende adaptée au besoin et aux goûts de l'utilisateur de manière automatique, interactive et rapide. Notre première application (Jolivet 2009) repose sur des services Web traduisant les étapes de réalisation d'une carte (cf. figure 6).

Fig. 6 – Services Web reprenant les étapes de réalisation d'une carte et permettant d'interpréter le besoin et les goûts de l'utilisateur en carte sur mesure

Le service 1 a pour vocation d'écrire une description générale formalisée de la carte. Il s'agit pour l'utilisateur de définir le contexte de réalisation de la carte. En particulier, il précise la zone géographique à visualiser, l'échelle, le type de carte. Ces précisions guident l'organisation de la légende en thèmes et les choix de symbolisation (cf. § 2) ; elles permettent la proposition de légendes prédéfinies adaptées.

Le service 2 affine les propositions précédentes en prenant en compte, tout en respectant les contraintes liées à l'ontologie, les préférences de l'utilisateur quant à la couleur des symboles. L'utilisateur peut attribuer des qualificatifs qui sont associés à des couleurs ; il peut également déterminer une couleur qu'il ressent comme plus adaptée ou plus lisible pour des objets qu'il juge importants.

Le service 3 affiche la carte : la légende et les données auxquelles a été appliquée la symbolisation construite.

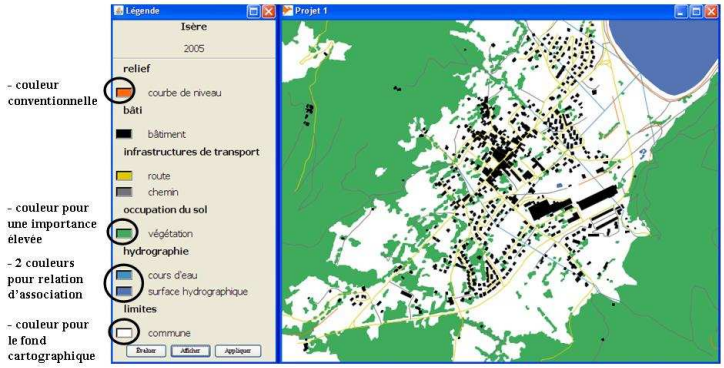


Fig. 7 – Exemple d'une conception de légende de carte topographique à partir des choix d'un utilisateur

La figure 7 illustre la carte au 25:000 obtenue en appliquant la symbolisation, générée automatiquement, à des données topographiques de l'Isère. Cette symbolisation prend en compte une préférence exprimée par l'utilisateur sur la visualisation prioritaire de la zone arborée. Les couleurs de symbolisation sont donc adaptées (ici une valeur forte choisie pour la forêt) tout en respectant les contraintes cartographiques.

Les légendes prédéfinies sont variées, construites pour tenir compte de la diversité des données et des utilisateurs : l'application présentée permet d'obtenir une légende personnalisée, en excluant les erreurs dues à la non-observation des règles de sémiologie graphique.

4.2 Conception coopérative de légendes à l'aide de palettes de peintres

Notre deuxième application (Christophe 2008) propose un modèle de conception coopérative de légende : il s'agit d'aider un utilisateur à concevoir des légendes cartographiquement correctes et innovantes en s'appuyant sur des techniques de dialogue homme-machine. Le modèle proposé gère à la fois des contraintes cartographiques formalisées et les préférences que l'utilisateur exprime au cours du dialogue. Une des stratégies implémentées consiste à proposer à l'utilisateur des palettes harmonieuses extraites de toiles de maîtres dans lesquelles il peut sélectionner des couleurs qui lui plaisent et qui seront utilisées dans sa légende. Les préférences de l'utilisateur sont enregistrées comme contraintes sur la légende en construction. Nous utilisons ici la « palette Derain », extraite du tableau « Montagnes à Collioure » de

Derain (1905). Dans notre exemple, les objets à représenter sont structurés en cinq thèmes : *commune, bâti, réseau routier, forêt, mer*. (Christophe 2008) raisonne sur le cercle chromatique de Chesneau (2006) pour associer les couleurs de la peinture aux thèmes de la légende : ces couleurs sont donc rapprochées des couleurs du cercle par mesure de la plus faible distance et représentées par des ronds noirs sur la figure 8.

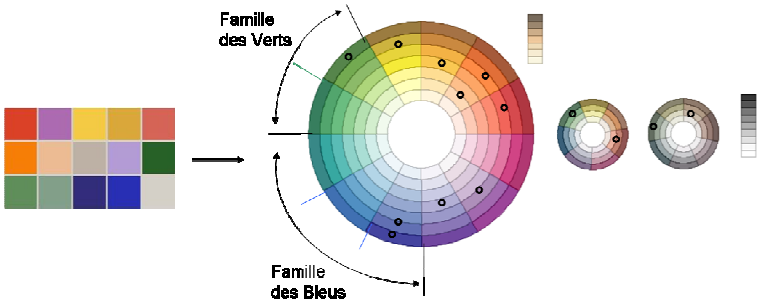


Fig. 8 – Rapprochement de notre « palette Derain » et des couleurs du cercle

Nous satisfaisons d'abord les règles d'utilisation conventionnelle de la couleur en s'appuyant sur les familles de couleurs du cercle : deux couleurs de la peinture appartiennent à la famille des verts et sont donc utilisables pour le thème *forêt*. De même trois couleurs de la peinture appartiennent à la famille des bleus et sont utilisables pour la *mer*. Une seule couleur suffisamment claire est disponible dans la palette pour le fond cartographique (qui correspond au thème commune). Les autres familles de couleurs sont utilisables pour les thèmes restants : bâti, routes, autoroutes. Il s'agit alors d'un simple problème de satisfaction de contraintes : un ensemble de variables (les thèmes restants), un ensemble de valeurs (les couleurs restantes dans la palette), des contraintes (les petits objets plus foncés ; des teintes différentes pour des thèmes différents). L'ensemble de ces contraintes associées aux préférences utilisateur permettent de proposer à celui-ci plusieurs légendes où les relations entre thèmes sont respectées. La figure 9 en présente deux où l'utilisateur a choisi le rouge de la palette pour les autoroutes.

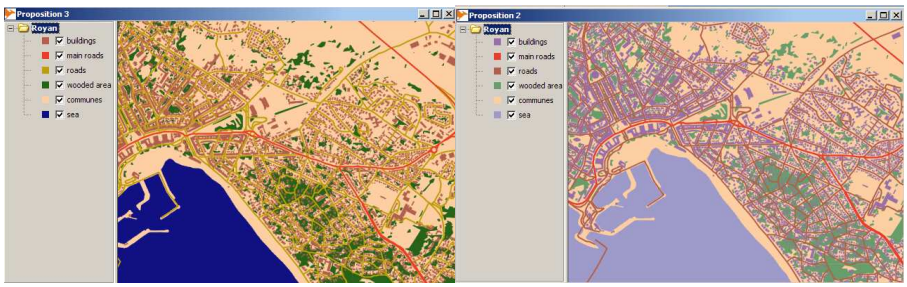


Fig. 9 – Propositions de légende à partir de la palette Derain

Il faut noter selon ses préférences, l'utilisateur peut relaxer des contraintes cartographiques.

5 Conclusions et perspectives

Pour aider un utilisateur à construire lui-même une carte adaptée à son besoin et prenant en compte ses préférences, nous nous sommes intéressées à l'étape de définition de la symbologie de la légende. Nous avons extrait des connaissances opérationnelles et les avons intégrées dans des outils qui mettent ainsi à disposition des utilisateurs, novices ou experts, une expertise en cartographie. Les outils que nous avons proposés permettent de réaliser une carte sur mesure parce qu'ils donnent la possibilité à l'utilisateur de personnaliser la représentation de ses données (ce que n'autorisent pas les logiciels existants) et prennent en compte ses préférences et sa satisfaction. Enfin, parce qu'ils intègrent à la fois les règles cartographiques et les préférences des utilisateurs, ces outils peuvent être proposés à un large éventail de profils : à des novices, ils apporteront d'abord l'expertise en cartographie qui leur fait défaut pour concevoir des légendes correctes ; aux experts, ils permettront d'affiner leur utilisation de la couleur en proposant des palettes colorées riches et éprouvées.

Des travaux complémentaires s'intéressant à la traduction de l'harmonie et du contraste dans la carte toute entière sont en cours de développement. Un travail entamé sur la recherche à l'aide de patrons morphosyntaxiques dans un corpus spécialisé, ayant pour sujets la cartographie et la couleur, a montré que d'autres termes liés à l'expression du ressenti des utilisateurs devraient être rajoutés à l'ontologie. En outre, celle-ci devra être élargie à l'utilisation des autres variables visuelles de la sémiologie graphique pour proposer des légendes innovantes plus variées. Enfin, la définition de règles décrivant plus spécifiquement le savoir-faire et l'expertise des professionnels pourraient aussi augmenter la variété des propositions de légende.

Ontocarto vise à décrire l'utilisation de la couleur dans la conception de la légende de la carte, et plus précisément l'attribution au sein d'une ligne de légende d'un symbole coloré à un objet cartographique représentant un objet géographique. Cet objet géographique permettra de faire le lien entre Ontocarto et l'ontologie géographique GeOnto (Abadie & Mustière, 2008). GeOnto vise à aider l'utilisateur à sélectionner dans des bases de données géographiques (BDG) les objets pertinents par rapport à son besoin de carte, en comparant les spécifications des BDG à sa hiérarchie d'objets.

Références

ABADIE N. & MUSTIERE S. (2008). Création d'une taxonomie géographique à partir des spécifications de bases de données. In SAGEO'08, Montpellier.

BEGUIN M. & PUMAIN D. (2000). La représentation des données géographiques. Armand Colin, 2nde édition.

BERTIN J. (1967). Sémiologie graphique: les diagrammes, les réseaux, les cartes. Rééditions en 1973, puis en 1998. Paris. Editions de l'EHESS.

BUARD E. & RUAS A. (2007). Evaluation of colour contrasts by means of expert knowledge for on-demand mapping. 23rd ICA conference, 4-10 August 07, Moscow, Russia.

BUCHER, B. (2007). La Carte à la carte sur le Web, dans Le Monde des Cartes, revue du comité français de cartographie, n°193.

BREWER C. (2003). A Transition in Improving Maps: The ColorBrewer Example. *Cartography and Geographic Information Science*. Vol.30, n°2, pp. 159-162.

CHESNEAU E. (2006). Modèle d'amélioration automatique des contrastes de couleurs en cartographie. Thèse de doctorat, Université de Paris Est-Marne La Vallée.

CHRISTOPHE S. (2008). Creative cartography based on dialogue. In the proceedings of the Conference AutoCarto 2008, 8-10 September.

CUENIN R. (1972). Cartographie générale (tome 1) Notions générales et principes d'élaboration. p. 109-179. Paris. Editions Eyrolles.

CULLOT N., PARENT C., SPACCAPIETRA S., VANGENOT C. (2003). Des SIG aux ontologies géographiques. *Revue internationale de géomatique. Les SIG sur le web*. Vol 13 n°3/2003 pp. 285-306.

DAO H (2004). Les principes de la représentation cartographique de données géographiques. Une approche ontologique et sémiologique. *Géomatique. Les ontologies spatiales*. 14/2004 pp. 259-283.

DOMINGUES C., BUCHER B., (2006). Application d'aide à la conception de légende, actes du colloque SAGEO, Strasbourg.

IOSIFESCU-ENESCU I. & HURNI L. (2007). Towards cartographic ontologies or "how computers learn cartography". 23rd ICA conference, 4-10 August 07, Moscow, Russia.

ITTEN J. (1985). Art de la couleur. Réédition en 2004. Dessain et Tolra.

JOLIVET L. (2009) Characterizing maps to improve on-demand cartography - the example of European topographic maps. Actes de la conference GISRUK, Royaume-Uni, Durham.

KRYGIER J. (1995). Cartography as an art and a science? *The Cartographic Journal* 32: 6. pp. 3-10.

KUROKAWA C & OTA M. (2007). Portrayal schema design and mechanism for the map personalization. 23rd ICA conference, 4-10 August 07, Moscow, Russia.

MOLLARD-DESFOUR A. (1998). Le dictionnaire des mots et expressions de couleur du XXè siècle. Le bleu. Rééd. 2004. CNRS Editions. Paris.

MONMONIER M. (1991). How to lie with maps. University of Chicago Press.

PASTOUREAU M. & SIMONNET D. (2005). Le petit livre des couleurs. Ed. du Panama. Paris.

REKACEWICZ Ph. (2006). La cartographie, entre science, art et manipulation. A propos de L'Atlas 2006 du Monde diplomatique. *Le Monde Diplomatique.*

ROBINSON A.H. (1952). *The Looks of Maps*. Madison. University of Wisconsin Press.

De l'analyse d'un corpus de texte à la conception d'une interface graphique facilitant l'accès aux connaissances sur le médicament

Jean-Baptiste Lamy, Catherine Duclos, Alain Venot

Laboratoire d'Informatique Médicale et de Bioinformatique (LIM&BIO), UFR SMBH, Université Paris 13, 74 rue Marcel Cachin, 93017 Bobigny cedex, France
jiba@soyaproject.org, catherine.duclos@avc.ap-hop-paris.fr, avenot@smbh.univ-paris13.fr

Abstract :

Pour faciliter l'accès aux contre-indications, interactions médicamenteuses et effets indésirables des médicaments par les professionnels de santé, nous avons conçu une interface graphique s'appuyant sur un langage iconique. Dans cet article, nous présentons la méthode de conception de cette interface, laquelle repose sur des techniques de visualisation d'information, et sur l'analyse distributionnelle d'un corpus de textes de référence sur le médicament. Cette analyse a permis de déterminer les principaux axes pour organiser les icônes sur l'interface.

Mots-clés : Interface graphique, Interface iconique, Accès aux connaissances textuelles, Analyse d'un corpus de texte, Connaissances sur le médicament.

1 Introduction

Les erreurs de prescriptions sont responsables d'un nombre important d'hospitalisations (Moore *et al.*, 2007) ; mais beaucoup d'entre elles auraient pu être évitées si les propriétés des médicaments avaient été prises en compte lors de la prescription. Ces connaissances sont disponibles sous forme textuelle dans le RCP (Résumé des Caractéristiques Produits, le texte de référence sur un médicament décrivant notamment contre-indications, interactions médicamenteuses et effets indésirables). Cependant, le volume de texte et le manque de temps empêchent souvent la lecture du RCP en consultation.

Les capacités de la vision humaine (Paivio, 1990) étant sous-utilisées par la présentation textuelle, l'accès aux connaissances peut être facilité par une présen-

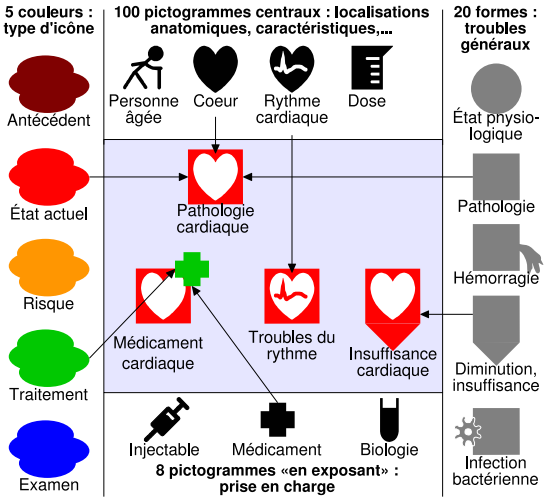


Figure 1: Construction des icônes du langage VCM à partir des primitives.

tation graphique. Les approches graphiques peuvent être classées en deux catégories : les *langages iconiques* (Chang, 1986) qui permettent de représenter des informations ou des connaissances en combinant pictogrammes, couleurs, formes,..., selon une grammaire graphique (Meunier, 1998), par exemple les panneaux routiers, et la *visualisation d'information* (Andrews, 2002) qui propose plusieurs techniques pour présenter graphiquement la structure de l'information, par exemple l'arborescence d'un système de fichiers. Bien que complémentaires, ces deux approches ont rarement été utilisées de manière conjointe.

Pour faciliter l'accès aux connaissances médicales, nous avons mis au point le langage iconique VCM (Visualisation des Connaissances Médicales) (Lamy *et al.*, 2008a; Lamy *et al.*, 2008b) qui permet de représenter par des icônes les principaux concepts des connaissances médicales : pathologies, risques, médicaments, examens, habitudes de vie,... Il se compose d'un jeu de primitives graphiques incluant plusieurs couleurs et une centaine de pictogrammes, et d'une grammaire graphique décrivant les règles pour combiner les primitives et construire les icônes. Par combinatoire, il est ainsi possible de créer plusieurs milliers d'icônes (figure 1). La grammaire du langage VCM définit aussi des relations *est-un* entre les icônes : par exemple entre l'icône signifiant "trouble du rythme" et celle signifiant "pathologie cardiaque". VCM a été conçu à destination des professionnels de santé (médecins, pharmaciens...). Il demande un court apprentissage, qui est facilité par l'utilisation de nombreux symboles et pictogrammes évocateurs.

Nous nous intéressons dans cet article à la conception d'une interface graphique pour faciliter et accélérer la consultation des contre-indications, des interactions médicamenteuses et des effets indésirables décrits dans un RCP, en s'appuyant sur le langage VCM et sur des techniques de visualisation d'information. Nous

avons travaillé sur deux cas d'utilisation mis au point par un expert du domaine :

1) le médecin prescrit un médicament complexe ou qu'il connaît mal, et souhaite vérifier l'ensemble des contre-indications et interactions médicamenteuses par rapport à son patient. Il a alors besoin d'une vue d'ensemble des propriétés du médicament, ce que ne fournit pas le texte.

2) le patient présente une pathologie ou un traitement inhabituel, et le médecin veut vérifier si le médicament qu'il prescrit est contre-indiqué avec cette pathologie ou ce traitement ; ou alors le patient présente un symptôme et le médecin veut vérifier s'il s'agit d'un effet indésirable d'un médicament donné. Si la contre-indication, l'interaction médicamenteuse ou l'effet indésirable recherché n'existe pas, cette connaissance est implicite dans le texte du RCP ; par exemple si le RCP ne mentionne pas de contre-indication avec l'asthme, il est sous-entendu qu'il n'y en a pas. Le médecin doit alors parcourir la totalité de la section correspondante du RCP avant de pouvoir déduire son absence, ce qui prend du temps. Un second besoin du médecin est donc d'accéder rapidement aux absences de contre-indication, d'interaction médicamenteuse ou d'effet indésirable.

L'objectif de cet article est de présenter la méthode de conception d'une interface répondant à ces cas d'utilisation. Tout d'abord, nous présenterons une analyse d'un corpus de RCP qui nous a permis de choisir les axes pour classer et organiser les contre-indications, les effets indésirables et les interactions médicamenteuses sur une vue d'ensemble. Ensuite, nous verrons comment les résultats de cette étude ont été utilisés pour construire une interface graphique s'appuyant sur des techniques de visualisation d'information. Enfin, nous donnerons brièvement les résultats obtenus lors d'une évaluation de cette interface, et nous discuterons la méthode utilisée.

2 Analyse d'un corpus de RCP

2.1 Matériel et méthodes

Nous avons travaillé sur les 278 RCP des médicaments disponibles en France et présents dans la liste des médicaments essentiels de l'OMS, donnant lieu à 3 corpus comprenant les sections contre-indications (constituées d'énumérations et comprenant 25672 mots), effets indésirables (constituées d'énumérations et de phrases, et comprenant 534376 mots) et interactions médicamenteuses (constituées d'énumérations et comprenant 492351 mots) de chaque RCP.

Cette analyse distributionnelle s'est appuyée sur des outils de Traitement Automatique de la Langue (Bouillon & Vandooren, 1998). Elle a consisté à extraire les candidats termes automatiquement avec Cordial et Lexter (Bourigault, 1995), puis à extraire les termes décrivant des contre-indications pour le premier corpus, des effets indésirables pour le second, et des médicaments interagissant pour le troisième, et enfin à extraire les attributs qui caractérisent chacun de ces termes, de manière semi-automatique, selon les étapes suivantes (figure 2) :

(1) Nous appellerons CT1 la liste des candidats-termes. À partir de CT1, extraire la liste de tous les noms et adjectifs, sans doublon, que nous appellerons W1.

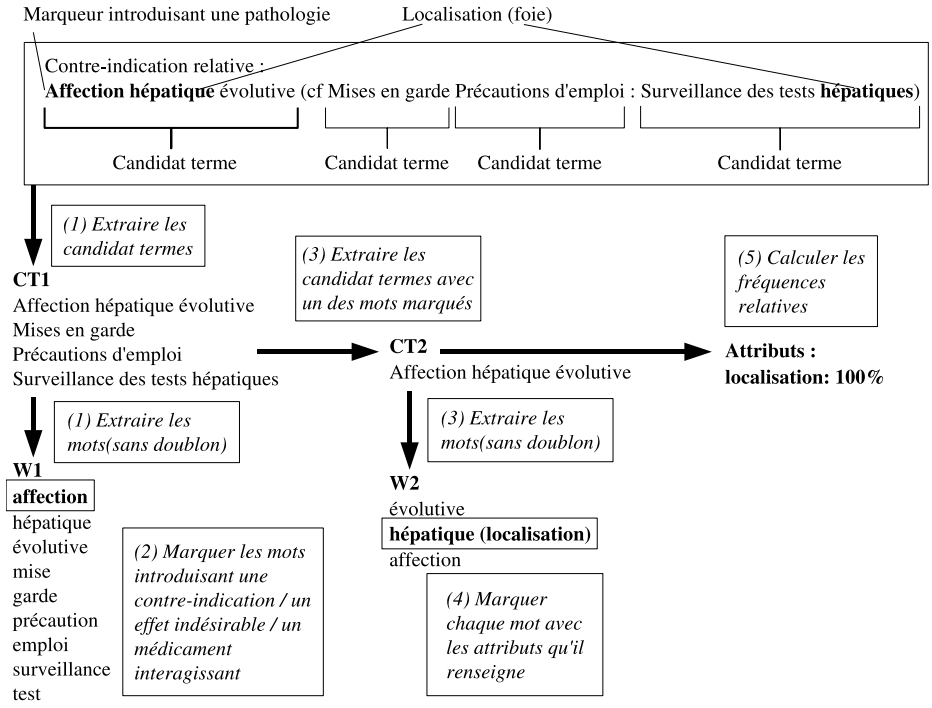


Figure 2: Méthode utilisée pour l'analyse des corpus de textes. La méthode est ici appliquée à une contre-indication.

(2) Un expert marque dans la liste W1 quels sont les mots indiquant les pathologies contre-indiquées pour le premier corpus (par exemple "maladie", "cardiopathie", "hypertension",... indiquent des pathologies), les effets indésirables pour le second, et les médicaments interagissant pour le troisième.

(3) À partir de CT1, extraire la liste de tous les candidats termes comprenant au moins un mot marqué à l'étape précédente dans W1 ; nous l'appellerons CT2. Cette liste contient donc tous les termes qui nous intéressent. Ensuite, à partir de CT2, extraire la liste de tous les noms et adjectifs, sans doublon, que nous appellerons W2.

(4) Un expert marque dans la liste W2 chaque mot avec les attributs qu'il renseigne, y compris les attributs qui ne sont pas indiqués de manière explicite mais dont un médecin a connaissance. Par exemple, "cardiopathie" renseigne l'attribut "localisation anatomo-fonctionnelle" (valeur : cardiaque), et "tuberculose" renseigne les attributs "localisation anatomo-fonctionnelle" (valeur : pulmonaire) et "étiologie" (valeur : bactérienne).

(5) Pour chaque attribut, la fréquence relative de l'attribut est donnée par la proportion de candidat-termes dans la liste CT2 qui inclut au moins un mot marqué avec cet attribut dans la liste W2.

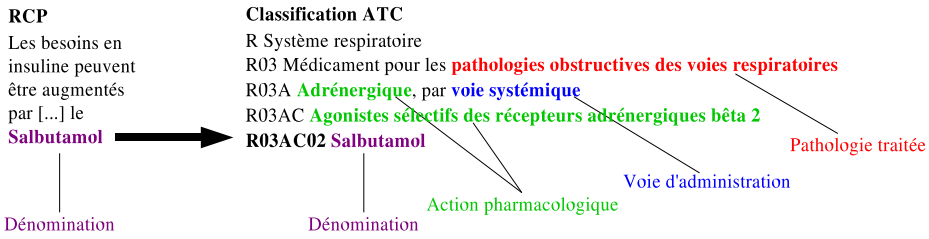


Figure 3: Exemple d'utilisation de la classification ATC pour caractériser les dénominations des médicaments. Ici, la dénomination chimique "Salbutamol" est enrichie par des attributs concernant l'activité pharmacologique, la voie d'administration et la pathologie traitée.

Les étapes 2 et 4 ont été réalisées à la main par un expert du domaine, pharmacien de formation. Les autres étapes ont été automatisées à l'aide de scripts.

Pour les médicaments, les dénominations chimiques des principes actifs sont fréquemment employées, et il est difficile, même pour un expert, de connaître la totalité des attributs se rapportant à chaque dénomination. Nous avons résolu ce problème avec la classification ATC (*Anatomical Therapeutical Chemical drug classification*, classification Anatomique, Thérapeutique et Chimique des médicaments). Dans un premier temps, les libellés des termes ATC ont été marqués avec les attributs qu'ils définissaient. Dans un second temps, nous avons relié chaque dénomination de médicament au terme ATC correspondant, et nous lui avons associé les attributs marqués dans le libellé, ainsi que dans ceux de tous les termes parents dans la classification (figure 3).

2.2 Résultats

L'analyse du corpus de RCP a permis de déterminer les fréquences relatives des attributs caractérisant les contre-indications et les effets indésirables (tableau 1) et les médicaments (tableau 2).

Pour les contre-indications et les effets indésirables, une localisation anatomo-fonctionnelle (par exemple cardiaque ou pulmonaire) peut être associée à 69% des contre-indications et 76% des effets indésirables, un trouble (c'est à dire le problème présent, par exemple une insuffisance dans la fonction d'un organe) à 46% des contre-indications et 70% des effets indésirables, et une étiologie (c'est à dire la cause de la pathologie, par exemple virale ou bactérienne) à 30% des contre-indications et 10% des effets indésirables. Les autres attributs, gravité de la pathologie, niveau de certitude du diagnostic, et nom propre associé, sont plus rarement présents.

Pour les médicaments, les dénominations (chimiques ou noms de marque) ont été remplacées par les attributs correspondant en s'appuyant sur la classification ATC. Un terme ATC a pu être trouvé pour 86% des dénominations, les autres correspondant à des familles chimiques très larges ou sans équivalent ATC. Cette

	Contre-indications	Effets indésirables	Pathologies traités par les médicaments
Localisation anatomo-fonctionnelle	69%	76%	39%
Troubles	46%	70%	34%
Étiologie	30%	10%	31%
Gravité	10%	5%	-
Incertitude	2%	0,2%	-
Nom propre	2%	2%	-
Localisation et / ou étiologie	94%	82%	67%
Localisation et / ou trouble	73%	84%	40%

Table 1: Fréquences relatives d'utilisation des attributs pour les pathologies et effets indésirables.

	Médicaments
Dénomination chimique	71%
Pathologie traitée	67%
Effet	26%
Voie d'administration	26%

Table 2: Fréquences relatives d'utilisation des attributs pour les médicaments.

correspondance a été réalisée de façon automatique pour 56% des termes (qui correspondent aux dénominations chimiques bien orthographiées), et de façon manuelle pour le reste (noms de marque, familles chimiques ou dénominations chimiques comportant une faute d'orthographe). L'activité des médicaments est caractérisée par la pathologie qu'il traite (par exemple "anti-hypertenseur") dans 67% des cas, et par l'effet (opposé à la pathologie traitée, par exemple "hypotenseur") dans 26%. Pour les médicaments décrits par la pathologie traitée, nous avons indiqué dans le tableau 1 les attributs caractérisant la pathologie traitée.

L'utilisation conjointe des attributs localisation anatomo-fonctionnelle et étiologie permet de couvrir un maximum de termes rencontrés dans les RCP : pour 94% des contre-indications, 82% des effets indésirables et 67% des médicaments il est possible d'associer une localisation ou une étiologie (contre 73%, 84% et 40% si l'on retenait les attributs localisation et trouble). Les termes associés ni à une localisation ni à une étiologie sont principalement des candidats termes non interprétables médicalement (par exemple des démonstratifs : "cette pathologie") ou mal extraits par l'analyseur syntaxique, ou, pour les médicaments, des classes pharmacologiques ou chimiques très larges ("les sulfamides").

En conclusion de cette analyse, nous avons retenu les deux axes localisation anatomo-fonctionnelle et étiologie pour organiser les contre-indications, les effets indésirables et les interactions médicamenteuses sur l'interface interactive.

3 Conception de l'interface

3.1 Matériel et méthodes

Afin de présenter une vue d'ensemble des contre-indications, interactions médicamenteuses et effets indésirables d'un RCP, nous avons utilisé une technique de visualisation d'information appelée *Fisheye* (Furnas, 1986; Hascoët & Beaudouin-Lafon, 2001). Cette technique consiste à séparer les informations à visualiser en deux parties : le *contexte*, qui donne une vue d'ensemble de la totalité des informations avec un faible niveau de détails, et le *focus*, qui présente en détail la partie des informations qui a été sélectionnée par l'utilisateur dans le contexte. Dans notre interface, le contexte sera l'ensemble des icônes VCM associées aux contre-indications, effets indésirables et interactions médicamenteuses du RCP, et le focus, les passages textuels du RCP associés aux icônes sélectionnées par l'utilisateur.

L'analyse du corpus de RCP a conduit à organiser le contexte selon deux axes anatomique et étiologique. Un schéma anatomique peut être soit être réaliste, soit simplifié et organisé en "cases" fixes et semi-arbitraires, correspondant chacune à une localisation (figure 4). Nous avons préféré le schéma simplifié pour 3 raisons : (1) sur un schéma réaliste, il est difficile de représenter simultanément les organes de certaines régions comme l'abdomen à cause de leur superposition, (2) les cases ont toutes la même taille ce qui évite des erreurs d'interprétations, par exemple d'accorder plus d'importance aux organes de plus grande taille, et (3) il est plus facile de placer des icônes VCM sur un schéma simplifié. Le schéma anatomique a ensuite été enrichi en ajoutant des cases pour les étiologies ; ces cases ont été placées à l'extérieur du personnage, car beaucoup d'étiologies correspondent à des éléments exogènes (virus, bactéries,...). Afin de répondre au second cas d'utilisation vu en introduction (permettre au médecin de trou-

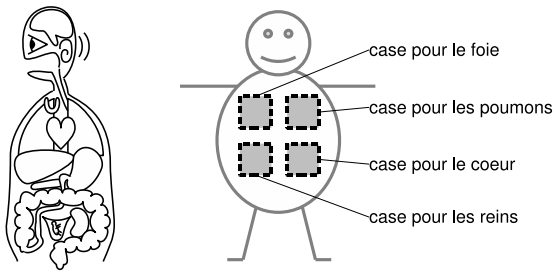


Figure 4: Exemples de schémas anatomiques. Le schéma de gauche est réaliste, tandis que celui de droite est simplifié et découpé en cases.

ver rapidement l'absence d'une contre-indication, interaction médicamenteuse ou effet indésirable) nous avons présenté l'absence d'item pour une localisation anatomo-fonctionnelle ou une étiologie de manière explicite sur le contexte, en grisant la case correspondante.

Un prototype a été évalué sur un groupe de 11 médecins généralistes formés à l'usage de VCM par l'utilisation d'un didacticiel qui leur a été remis un mois à l'avance, et auquel ils ont consacré en moyenne 4h. L'évaluation avait pour objectif de comparer la vitesse de lecture et le nombre d'erreurs obtenus par des médecins utilisant tantôt une interface textuelle standard, tantôt l'interface graphique. Elle a porté sur des RCP chimériques générés de manière aléatoire à partir de vrais RCP issus de la base médicamenteuse Thériaque, et des questions de la forme "Ce médicament peut-il être prescrit sans précaution particulière chez un patient souffrant de la pathologie X ?", "Ce médicament peut-il être prescrit sans précaution particulière chez un patient prenant le médicament Y ?", ou "Ce médicament peut-il provoquer l'effet indésirable Z ?", tirées au hasard elles-aussi et correspondant au second cas d'utilisation donné en introduction.

3.2 Résultats

Le schéma anatomique que nous avons conçu a la forme d'un bonhomme stylisé que nous avons appelé "Monsieur VCM". La figure 5 montre les différentes localisations et étiologies, chacune étant représentée par le pictogramme correspondant en langage VCM. Seuls la tête, les pensées, le corps et un bras du bonhomme sont représentés, délimitant ainsi 5 zones : la tête (avec des localisations comme les yeux, les oreilles,...), les pensées (psychiatrie, psychologie,...), le corps (système digestif, coeur, organes sexuels,...), le bras (système nerveux périphérique, os, peau,...) et un espace en dehors du bonhomme où sont représentées les étiologies (virale, bactérienne,...). Les jambes et le second bras sont ébauchés, afin de compléter le dessin du bonhomme.

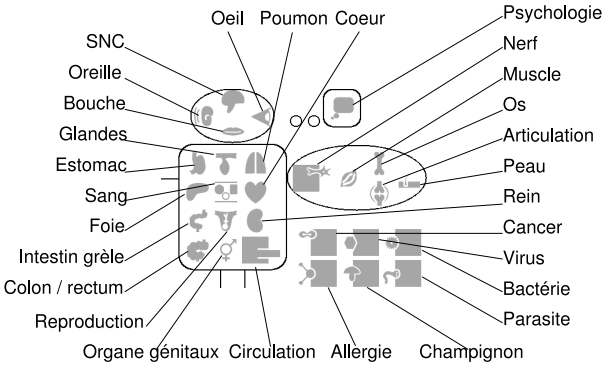


Figure 5: La répartition des différentes localisations et étiologies sur "Monsieur VCM".

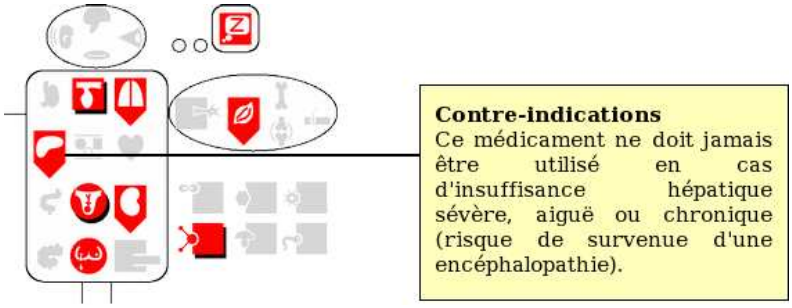


Figure 6: Exemple d'interface "Monsieur VCM" présentant les contre-indications du Stilnox® (un somnifère), après que l'utilisateur ait cliqué sur l'icône "insuffisance hépatique" placée sur la case du foie.

"Monsieur VCM" peut accueillir les icônes VCM correspondant aux contre-indications, aux effets indésirables ou aux interactions médicamenteuses (figure 6, un "Monsieur VCM" pour chacune des trois sections du RCP). Lorsque plusieurs icônes devraient occuper la même case du schéma, les relations *est-un* existant entre les icônes VCM sont utilisées pour les remplacer par l'icône du plus petit parent commun. Par exemple les icônes pour "insuffisance cardiaque" et pour "angor" se placent toutes deux au niveau du coeur ; elles seront remplacées par l'icône plus générale "pathologie cardiaque". Une ombre est ajoutée à cette icône, pour indiquer qu'elle résulte d'une combinaison. Lorsqu'une contre-indication, un effet indésirable ou un médicament est associé à la fois à une localisation et une étiologie, son icône est présente deux fois. Au contraire, si une pathologie ou un médicament n'a ni localisation ni étiologie, son icône est placée en dessous du bonhomme ; ce cas reste cependant exceptionnel, comme nous l'avons montré lors de l'analyse des RCP. Lorsqu'une case ne reçoit aucune icône, celle-ci est occupée par un pictogramme grisé. Enfin, il suffit de cliquer sur une icône pour faire apparaître le ou les passages de texte correspondant dans le RCP.

L'accès à un RCP dans l'interface graphique se fait par une recherche portant sur le nom du médicament. Les trois sections contre-indications, effets indésirables et interactions médicamenteuses du RCP peuvent être représentées par trois "Messieurs VCM". Il est aussi possible de regrouper les contre-indications et les interactions médicamenteuses sur le même "Monsieur VCM", car le langage VCM définit des relations *est-un* entre les icônes de pathologies et de médicaments. En effet, ces icônes représentent en fait, du point de vue d'un médecin, un patient souffrant d'une pathologie et un patient prenant un médicament. Comme les médicaments sont représentés à l'aide de la pathologie qu'il traite, nous avons bien la relation est-un suivante : un patient prenant un médicament traitant la pathologie X est un patient souffrant de la pathologie X.

Les résultats de l'évaluation ont montré que l'interface utilisant "Monsieur VCM" avait été lue par les médecins en moyenne 2,2 fois plus vite ($p < 10^{-16}$,

test t apparié) et avec 1,7 fois moins d'erreurs (p = 0.034, test de Fisher exact, 16 erreurs contre 27), par rapport à l'interface textuelle similaire à celles existantes (Lamy *et al.*, 2008c). Le temps de réponse avec l'interface textuelle augmente fortement avec la longueur du RCP, et lorsque l'on demande au médecin de rechercher un item qui est absent du RCP. Avec l'interface graphique, la longueur du RCP a une influence plus faible sur le temps de réponse, et la recherche d'un item présent ou absent n'a aucune influence, ce qui était attendu car "Monsieur VCM" représente de manière explicite l'absence d'item concernant une localisation anatomo-fonctionnelle ou une étiologie.

4 Discussion et conclusion

4.1 Analyse d'un corpus de texte

L'analyse d'un corpus de RCP a permis de sélectionner les deux axes principaux de l'interface graphique. Le corpus comprenait 258 RCP sur un total d'environ 5000, cependant ces RCP sont représentatifs de l'ensemble et, faisant partie de la liste des médicaments essentiels de l'OMS, sont vraisemblablement parmi les plus utilisés. Nous avons réalisé une analyse assez simple ; des méthodes plus sophistiquées auraient pu être mise en oeuvre mais n'était pas nécessaires. Une approche similaire a déjà été utilisée pour analyser la section indication du RCP (Duclos & Venot, 2000). Les étapes manuelles ont été restreintes à l'ensemble des mots présents dans les candidats termes (sans les doublons), et non à l'ensemble de tous les candidats termes. Nous n'avons fait appel qu'à un seul expert, car la variabilité inter-expert est vraisemblablement limitée vu la faible granularité de l'étude ; il pourrait être intéressant de vérifier ce point. Les listes de mots marqués par l'expert pourrait resservir à d'autres études de ce type. Enfin, l'utilisation d'un filtrage aurait pu permettre de retirer les candidats termes non interprétables avant leur étude.

Les résultats de l'analyse montrent l'importance de la dimension anatomo-fonctionnelle. Ce résultat n'est pas surprenant car on retrouve fréquemment cette dimension, par exemple dans les classifications médicales. En revanche, il est intéressant de constater que l'utilisation conjointe des deux attributs localisation anatomo-fonctionnelle et étiologie permet de couvrir la quasi-totalité des contre-indications et des effets indésirables. Pour les médicaments, la plupart sont associés à une pathologie traitée qui peut être caractérisée par au moins un de ces deux attributs, mais il reste cependant des classes chimiques ou pharmacologiques qui ne correspondent pas directement à une pathologie traitée, ainsi que certains médicaments qui peuvent traiter plusieurs pathologies.

4.2 Interface graphique "Monsieur VCM"

"Monsieur VCM" répond aux besoins définis en introduction, en offrant une vue d'ensemble graphique des contre-indications, effets indésirables ou interactions médicamenteuses d'un RCP, et en indiquant de manière explicite, par un pic-

togramme grisé, l'absence d'élément d'information concernant une localisation anatomo-fonctionnelle ou une étiologie. Les mêmes cases étant toujours placées au même endroit, "Monsieur VCM" permet de trouver rapidement les informations correspondant à un localisation ou une étiologie donné.

L'une des particularités de cette interface est de combiner ensemble les icônes qui se rapportent à la même localisation anatomo-fonctionnelle ou à la même étiologie, en créant une icône plus générale. Ce mécanisme fait varier le niveau d'abstraction des icônes en fonction de la quantité d'information disponible, de sorte à avoir un "Monsieur VCM" dont la taille reste constante : un RCP plus long se traduira par un "Monsieur VCM" avec des icônes plus générales mais pas plus volumineux. Une autre particularité est de représenter de manière explicite l'absence d'éléments concernant une localisation ou une étiologie donnée.

Plusieurs auteurs ont proposé des interfaces employant des schémas anatomiques (Kirby *et al.*, 1996; McCullagh *et al.*, 2003; Sundvall *et al.*, 2007). Comparés à ces interfaces, "Monsieur VCM" a l'avantage d'utiliser un langage iconique, et de ne pas être purement anatomique grâce à l'ajout d'un axe étiologique. Par ailleurs, les interfaces précédemment publiées était destinées à la visualisation de données médicales et non de connaissances, et proposaient des schémas anatomiques plus réalistes. Pour la visualisation des connaissances, un schéma moins réaliste permet un plus haut niveau d'abstraction ; par exemple, "Monsieur VCM" ne distingue pas le rein gauche et le rein droit car il n'existe pas de contre-indication spécifique à l'un des reins. En revanche, pour visualiser les données médicales d'un patient, il peut être utile de distinguer les deux. Enfin, la généralisation de l'interface graphique iconique à d'autres cultures que la culture occidentale reste à évaluer.

4.3 Conclusion

Nous avons présenté une méthode pour concevoir une interface graphique pour l'accès aux contre-indications, effets indésirables et interactions médicamenteuses. La méthode repose sur l'analyse d'un corpus de texte afin de dégager les principaux axes permettant d'organiser les pathologies et les médicaments, puis sur l'utilisation conjointe de techniques de visualisation et d'un langage graphique. L'évaluation de cette interface auprès d'un groupe de médecins a donné de bons résultats.

Nous envisageons à présent la conception d'interfaces pour d'autres types de connaissances médicales, comme celles contenues dans les guides de bonnes pratiques cliniques. Une autre perspective est la construction d'une ontologie pour les icônes VCM et les "Messieurs VCM", afin d'automatiser certains processus comme la correspondance entre les icônes et les classifications médicales existantes. Par ailleurs, les problèmes de volume de connaissances ne sont pas spécifiques au domaine médical, et nous pensons qu'une approche similaire pourrait être appliquée dans d'autres domaines, par exemple à des textes de lois.

References

Andrews K. (2002). *Information visualisation: tutorial notes*. Graz University of Technology.

Bouillon P. & Vandooren F. (1998). *Traitement automatique des langues naturelles*. Paris, Bruxelles: De Boeck Université.

Bourigault D. (1995). LEXTER, a terminology extraction software for knowledge acquisition from texts. In *9th knowledge acquisition for knowledge based system workshop (KAW '95)*, Banff, Canada.

Chang S. (1986). *Visual Languages*, chapter Iconic Visual Languages, p. 1–7. Plenum Press, New York.

Duclos C. & Venot A. (2000). Structured representation of drug indications: lexical and semantic analysis and object-oriented modeling. *Methods Inf Med*, **39**, 83–87.

Furnas G. (1986). Generalized Fisheye views. In *Proceedings of the Human Factors in Computing Systems CHI '86 conference*, p. 16–23.

Hascoët M. & Beaudouin-Lafon M. (2001). Visualisation interactive d'information. *Information, interaction, intelligence (I3)*, **1**(1), 77–108.

Kirby J., Cope N., Souza A., Fowler H. & Gain R. (1996). The PEN&PAD data entry system: from prototype to practical system. In J. Brender, J. Christensen, J.-R. Scherrer & P. McNair, Eds., *Medical Informatics Europe (MIE-96)*, p. 430–434, Copenhagen: IOS Press.

Lamy J.-B., Duclos C., Bar-Hen A., Ouvrard P. & Venot A. (2008a). An iconic language for the graphical representation of medical concepts. *BMC Medical Informatics and Decision Making*, **8**(16).

Lamy J.-B., Duclos C., Rialle V. & Venot A. (2008b). Quelle méthodologie tenant compte des sciences cognitives pour la conception des langages graphiques ? *numéro spécial de la revue d'intelligence artificielle (RIA)*, **22**(3-4), 265–280.

Lamy J.-B., Venot A., Bar-Hen A., Ouvrard P. & Duclos C. (2008c). Design of a graphical and interactive interface for facilitating access to drug contraindications, cautions for use, interactions and adverse effects. *BMC Medical Informatics and Decision Making*, **8**(21).

McCullagh P., McGuigan J., Fegan M. & Lowe-Strong A. (2003). Structure data entry using graphical input: recording symptoms for multiple sclerosis. *Stud Health Technol Inform*, **95**, 673–8.

Meunier J.-G. (1998). The categorial structure of iconic languages. *Theory & Psychology*, **8**(6), 805–825.

Moore T., Cohen M. & Furberg C. (2007). Serious Adverse Drug Events Reported to the Food and Drug Administration, 1998-2005. *Arch Intern Med*, **167**, 1752–1759.

Paivio A. (1990). *Mental representations: a dual coding approach*. Oxford University Press, Incorporated.

Sundvall E., Nyström N., Forss M., Chen R., Petersson H. & Åhlfeldt H. (2007). Graphical overview and navigation of Electronic Health Records in a prototyping environment using Google Earth and openEHR Archetypes. In K. Kuhn, J. Warren & T.-Y. Leong, Eds., *MEDINFO 2007*: IOS Press.

COBRA : Une plate-forme de RàPC basée sur des ontologies

Amjad Abou Assali[1], Dominique Lenne[1], Bruno Debray[2]
et Sébastien Bouchet[2]

[1] Université de Technologie de Compiègne, CNRS
HEUDIASYC
{aabouass,dominique.lenne}@utc.fr

[2] INERIS[‡]
{bruno.debray,sebastien.bouchet}@ineris.fr

Résumé :
Cet article présente un projet en cours qui a pour objectif de développer une plate-forme de RàPC pour le diagnostic basée sur des ontologies, appelée COBRA. Cette plate-forme est constituée de deux parties principales : les modèles de connaissances décrits par des ontologies, et les processus de raisonnement. Nous travaillons actuellement sur la défaillance des barrières de sécurité installées sur des sites industriels. Cependant, notre objectif est de rendre la plate-forme générique et indépendante du domaine d'application. Nous affirmons que, pour mieux exploiter les avantages des ontologies dans les systèmes de RàPC, il est important de pouvoir utiliser n'importe quel concept dans la description des cas. Ainsi, COBRA permet de définir les attributs de chaque cas dynamiquement au moment de l'exécution, ce qui conduit à une base de cas hétérogène. Dans cet article, nous présentons l'architecture de la plate-forme, les modèles de connaissances, les processus principaux, ainsi que les problèmes rencontrés en travaillant avec des cas hétérogènes.

Mots-clés : Raisonnement à partir de cas, Ontologie, Base de cas hétérogène.

1 Introduction

La gestion des risques est devenue une préoccupation importante sur la plupart des sites industriels. Pour réduire les risques potentiels, des barrières de sécurité sont proposées par des experts. Toutefois, ces barrières peuvent échouer à assurer la fonction de sécurité pour laquelle elles ont été installées, et des accidents peuvent se produire. Dans ce contexte, un expert industriel intervient pour diagnostiquer le dysfonctionnement des barrières. Dans un premier temps, il essaie de se remémorer des expériences de dysfonctionnement observées dans des situations similaires. L'hypothèse que l'expert

[‡]Institut National de l'Environnement industriel et des RISques.

fait est que “si une barrière n’a pas bien fonctionné dans une autre situation similaire, il est fortement probable qu’elle ne fonctionne pas, dans la situation actuelle, *pour des raisons similaires*”. Quand l’expert avance dans son diagnostic, il cherche parfois à obtenir plus d’informations sur la situation actuelle pour pouvoir trouver la bonne cause de défaillance. Pour simuler l’activité de l’expert, nous utilisons une approche de RàPC (Riesbeck & Schank, 1989) conversationnelle.

Le RàPC est une approche de résolution de problèmes. Il a pour objectif de résoudre un nouveau problème, appelé *problème cible*, à l’aide d’un ensemble de problèmes déjà résolus, appelés *problèmes sources*. Les approches *knowledge-intensive* du RàPC (KI-CBR) sont celles pour lesquelles les connaissances du domaine jouent un rôle fondamental (et pas uniquement la base de cas). Dans notre travail, nous suivons une approche knowledge-intensive et conversationnelle car les cas sont enrichis au fur et à mesure que l’expert avance dans son analyse.

Nous travaillons actuellement sur la défaillance des barrières de sécurité installées sur des sites industriels, en particulier les capteurs de gaz. Ces capteurs déclenchent une alarme lorsqu’il y a une fuite de certains gaz quelque part dans le site. Dans ce contexte, un cas représente un diagnostic de la défaillance d’un capteur de gaz dans un environnement donné.

Nous présentons, dans cet article, COBRA (Conversational Ontology-based CBR for Risk Analysis), une plate-forme de KI-CBR indépendante du domaine d’application. Cette plate-forme est basée sur des modèles de connaissances représentés par des ontologies pour décrire le domaine et les cas. Nous affirmons qu’il est important de ne pas prédéfinir la structure (les attributs) des cas dans les systèmes de KI-CBR, mais de permettre à l’utilisateur de décrire ses cas par n’importe quel concept de l’ontologie de domaine. Par conséquent, nous obtenons une base de cas hétérogène où les cas peuvent être décrits par différents attributs. En outre, nous présentons les processus d’authoring et de remémoration des cas, ainsi que les mesures de similarité utilisées pour pallier les problèmes liés à l’hétérogénéité des cas.

2 État de l’art

L’intégration des connaissances génériques du domaine d’application dans les systèmes de KI-CBR a été un aspect important dans plusieurs projets. Dans l’architecture de CREEK (Aamodt, 1994), nous trouvons un couplage assez fort entre les connaissances des cas et celles du domaine. Ainsi, les cas sont immergés dans un modèle générique du domaine représenté par un réseau sémantique. Fuchs & Mille (2005) ont proposé une modélisation du RàPC au niveau connaissance. Ils ont distingué quatre modèles de connaissance : 1) le modèle conceptuel du domaine décrivant les concepts utilisés pour décrire l’ontologie du domaine indépendamment du raisonnement ; 2) le modèle de cas qui sépare le cas en problème, solution, et trace de raisonnement ; 3) les modèles de tâches de raisonnement qui comprennent un modèle de spécification et un autre de décomposition de tâches ; 4) et les modèles supports du raisonnement. D’Aquin *et al.* (2006) ont travaillé sur l’intégration du RàPC dans le Web sémantique. Pour cela, ils ont proposé une extension de OWL (Ontology Web Language) permettant de représenter les connaissances d’adaptation du RàPC. L’expression des connaissances

du domaine et des cas en OWL leur a permis de rajouter au système de RàPC les capacités de raisonnement propres à OWL en exploitant, par exemple, la subsumption et l'instanciation. Dìaz-Agudo & Gonzàlez-Calero (2000) ont proposé une architecture indépendante du domaine qui aide à l'intégration d'ontologies dans les applications de RàPC. Leur approche consiste à construire des systèmes intégrés qui combinent des connaissances spécifiques aux cas avec des modèles génériques des connaissances du domaine. Ils ont présenté CBROnto (Dìaz-Agudo & Gonzàlez-Calero, 2002), une ontologie de tâche/méthode qui fournit le vocabulaire nécessaire pour décrire les éléments impliqués dans les processus de RàPC, et qui permet également d'intégrer différentes ontologies de domaine. CBROnto a été réutilisée plus tard par jCOLIBRI, un framework orienté-objet (en JAVA) assez puissant pour la construction de systèmes de RàPC (Recio-Garcìa *et al.*, 2006; Dìaz-Agudo *et al.*, 2007). jCOLIBRI sépare la gestion des bases de cas en deux aspects : la persistance et l'organisation en mémoire, ce qui permet d'avoir différents supports de stockage de cas (fichiers text/XML, ontologie, *etc.*) accessibles via des connecteurs spécifiques. Toutefois, jCOLIBRI ne permet pas le traitement de bases de cas dynamiques et hétérogènes, ce qui nous a amené à développer une couche supplémentaire pour pallier ce manque.

3 Architecture de COBRA

Plusieurs architectures des systèmes de RàPC ont été proposées dans la littérature. Ces architectures partagent plus ou moins les mêmes composantes. Aamodt & Plaza (1994) ont présenté le cycle fameux de RàPC constitué des processus : REMÉMORER, ADAPTER, RÉVISER, et MÉMORISER. Un cinquième processus, ÉLABORER, a été distingué plus tard (voir Renaud *et al.* (2007)). En 2002, Lamontagne & Lapalme ont présenté une vue globale où l'on trouve les processus hors-ligne/en-ligne, et les connaissances ("knowledge containers") permettant de préserver et exploiter les expériences passées. Comme le montre la figure 1, COBRA est basé sur ces architectures et est composé de deux parties principales :

- *Processus* : cette partie est constituée d'un processus hors-ligne, *authoring des cas*, et de six processus de raisonnement : ÉLABORER, REMÉMORER, DIAGNOSTIQUER, ENRICHIR, VALIDER et MÉMORISER. Dans les systèmes conversationnels de RàPC pour le diagnostic, il est important de distinguer deux processus fondamentaux, *diagnostiquer* et *enrichir*. Dans la phase de "diagnostic", le système essaie d'identifier les causes de défaillance à partir des cas similaires. Si aucune cause n'est trouvée, ou si le diagnostic proposé par le système n'a pas été validé par l'utilisateur, le système demande à l'utilisateur d'enrichir la description du cas cible pour chercher de nouveaux cas similaires. Ce cycle se reproduit jusqu'à ce qu'un bon diagnostic soit proposé par le système, ou qu'aucune solution ne puisse plus être trouvée.
- *Connaissances* : dans les systèmes de RàPC, on distingue quatre catégories de connaissances ("knowledge containers") : vocabulaire, base de cas, mesures de similarité, et connaissances d'adaptation. Dans notre système, nous retrouvons les trois premières catégories, mais nous n'avons pas de règles d'adaptation pour le moment.

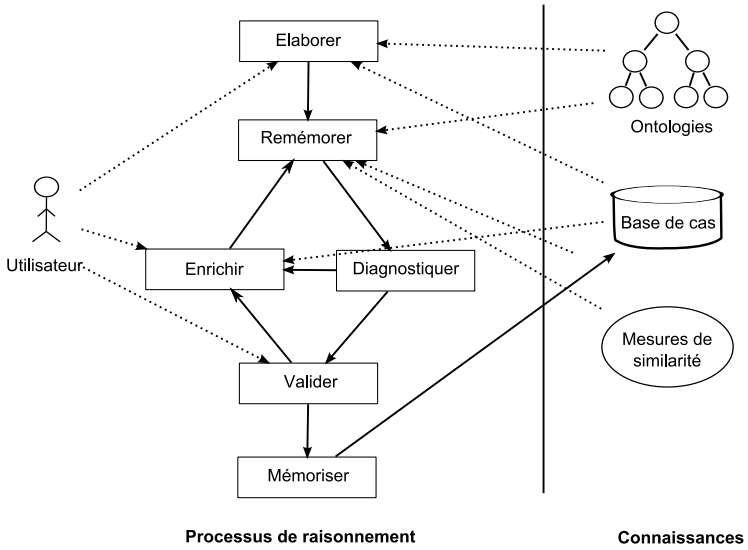


FIG. 1 – Architecture de COBRA.

Pour décrire le vocabulaire et la base de cas, le système se base sur deux modèles de connaissances : les modèles de domaine et de cas.

3.1 Modèle de domaine

Ce modèle représente les connaissances du domaine sous forme d'une ontologie. Dans les systèmes de KI-CBR, les ontologies jouent un rôle important (Recio-Garcìa *et al.*, 2006) comme vocabulaire pour décrire les cas, comme structure de connaissances où les cas sont localisés, et comme source de connaissance permettant le raisonnement sémantique dans les méthodes de calcul de similarité.

Dans ce travail, nous utilisons l'ontologie noyau développée dans (Abou Assali *et al.*, 2008), qui contient des concepts génériques sur la sécurité industrielle tels que : Événement, Phénomène dangereux, Explosion, Effet, *etc.* Une autre ontologie de domaine a été élaborée dont les concepts sont des spécialisations de l'ontologie noyau. Elle décrit le domaine des barrières de sécurité, le domaine d'application actuel (voir FIG. 2). Les ontologies sont décrites en OWL Lite, et elles ont été développées par plusieurs experts de l'INERIS avec l'aide d'un expert en ontologie.

3.2 Modèle de cas

Un cas dans notre système est un cas de diagnostic. Il contient trois parties principales : la partie *description*, qui décrit le contexte dans lequel a été réalisé le diagnostic, la partie *mode de défaillance*, et la partie *causes*. Prenons, par exemple, la défaillance de capteurs de gaz. La partie description peut contenir : le gaz à mesurer, la technologie du capteur utilisé, le seuil de la concentration d'alarme (*i.e.* à quelle concentration

du gaz le capteur doit déclencher une alarme), *etc.* Le mode de défaillance peut être : une fausse alarme, une absence de détection de gaz, *etc.* Enfin, la cause de défaillance peut être : la technologie du capteur est inadaptée pour le gaz à mesurer, un mauvais calibrage du capteur, *etc.* Un cas est généralement décrit par un couple (*problème*, *solution*). Selon notre modèle, les parties *description* et *mode de défaillance* correspondent à la partie *problème*, et la partie *causes* correspond à la partie *solution*.

Pour améliorer la communication entre la base de cas et le modèle de domaine, notre modèle de cas est représenté à l'aide d'une ontologie qui intègre le modèle de domaine. Nous nous inspirons en cela de l'approche utilisée dans jCOLIBRI. Cette ontologie contient les concepts racines suivants (FIG. 2) :

– CBR-CASE qui subsume les concepts représentant les différents types de cas qui peuvent exister dans le système.
– CBR-DESCRIPTION qui subsume les concepts représentant les parties d'un cas (mode de défaillance, cause, *etc.*).
– CBR-INDEX qui permet d'intégrer les concepts du modèle de domaine.

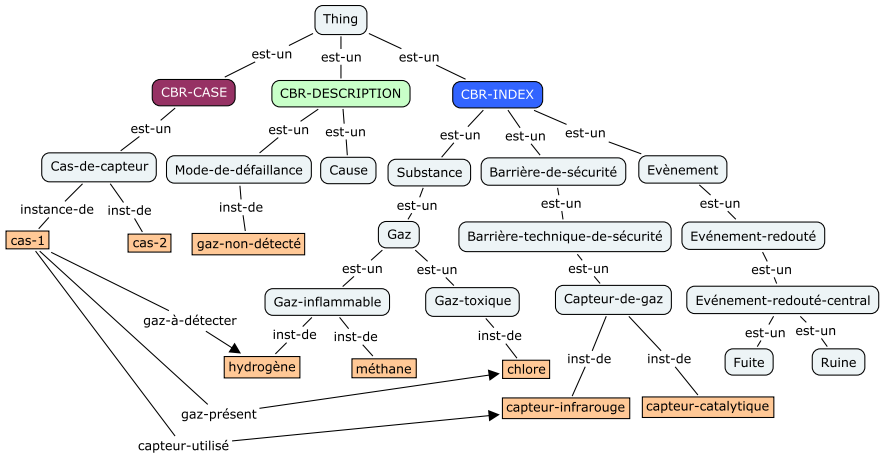


FIG. 2 – Modèle de cas.

Les cas sont alors représentés par des instances de l'ontologie, et ils ont donc deux types d'attributs :

– des attributs simples correspondant à des propriétés *data-type* de l'ontologie qui prennent des valeurs simples, *i.e.* string, int, float, *etc.*
– des attributs complexes correspondant à des instances de l'ontologie.

4 Processus de RàPC

Nous décrivons dans la suite les premiers processus de notre cycle de RàPC : authoring des cas, élaboration d'un cas cible (d'une requête), et remémoration de cas.

4.1 Authoring des cas

Pour initialiser les bases de cas, différents moyens peuvent être envisagés (Yang *et al.*, 2008). Dans notre travail, des experts du domaine ont été sollicités, dans un premier temps, pour construire des cas de diagnostic de la défaillance de capteurs de gaz à partir de leur expérience. Puis, des cas ont été rajoutés à partir des documentations existantes. Les cas sont décrits par des concepts du modèle de domaine. Toutefois, nous nous sommes rendu compte que nos cas ne partagent pas toujours la même structure, autrement dit, les mêmes attributs. Pour cela, deux objectifs ont émergé :

- Tout d'abord, les experts ne doivent pas être limités à un certain nombre d'attributs pour décrire leur expérience. Ils doivent pouvoir utiliser tout concept du modèle de domaine dont ils ont besoin. Cela conduit donc à une base de cas hétérogène, ce qui complique la phase de remémoration de cas.
- Comme les cas peuvent être décrits par différents attributs, il est utile d'aider l'expert durant l'authoring de son expérience. Pour ce faire, le système lui montre une liste des concepts les plus utilisés dans les cas similaires ordonnancés suivant leur importance. La base de cas est alors exploitée non seulement dans la phase de remémoration, mais aussi dans la phase d'authoring des cas.

Le processus d'élaboration de cas cibles est similaire. Par contre, des poids peuvent être associés aux attributs d'un cas cible pour être pris en compte dans la phase de remémoration (voir la section suivante).

4.2 Remémoration de cas

Dans cette phase, des mesures de similarité sont utilisées pour récupérer les cas similaires à un cas cible. En général, avec les structures orientées-objet des cas, les mesures de similarité suivent le principe "local-global" (Bergmann & Stahl, 1998; Richter, 2008) qui dit : "le but est de déterminer la similarité entre deux objets, un objet représentant le cas source (ou une partie du cas) et un autre objet représentant le cas cible (ou une partie du cas). Cette similarité est appelée la similarité globale, et est calculée d'une manière récursive ; *i.e.* pour chaque attribut simple, une mesure de similarité *locale* détermine la similarité entre les deux valeurs d'attribut. En revanche, pour chaque attribut complexe, une mesure de similarité *globale* est utilisée. Enfin, les valeurs des similarités locales et globales sont agrégées, d'une manière récursive, pour donner la similarité globale des deux objets comparés".

D'un point de vue ontologique, le calcul de similarité entre deux concepts de l'ontologie peut être divisé en deux composantes (Bergmann & Stahl, 1998; Recio-Garcìa *et al.*, 2006) : *une similarité basée-concept* (ou similarité intra-classe) qui dépend de l'emplacement des concepts dans l'ontologie, et *une similarité basée-slot* (ou similarité inter-classe) qui dépend des valeurs des attributs communs des objets comparés.

4.2.1 Mesures de similarité

Les attributs d'un cas cible n'ont pas toujours la même importance dans le calcul de similarité. Ainsi, il est important de permettre à l'utilisateur d'associer à chaque attribut

un certain poids. Dans notre travail, les poids peuvent être attribués à deux niveaux différents du cas cible :

- Les attributs simples peuvent avoir l'un de trois modes de calcul de similarité :
  - IGNORE : l'attribut n'a pas d'importance.
  - EXACT : qui permet de vérifier l'égalité stricte des valeurs de l'attribut.
  - NUMÉRIQUE : qui est applicable aux attributs numériques seulement. Plus les deux valeurs sont proches l'une de l'autre, plus elles sont similaires.
- Les attributs complexes ont un poids dans l'intervalle [0, 1]. En outre, chaque attribut simple d'un attribut complexe peut avoir l'un des trois modes de calcul (IGNORE, EXACT, NUMÉRIQUE).

Nous allons expliquer, dans la suite, les mesures de similarité utilisées dans COBRA.

Soit $Q = \{q_i : 1 \leq i \leq n, n \in \mathbb{N}^*\}$ un cas cible pour lequel on cherche des cas similaires, où q_i est un attribut simple ou bien complexe, et soit $\Omega = \{C_j : 1 \leq j \leq k, k \in \mathbb{N}^*\}$ la base de cas, où $C_j = \{c_{jl} : 1 \leq l \leq m_j, m_j \in \mathbb{N}^*\}$. La similarité basée-concept, sim_{cpt}, est définie comme suit :

Pour chaque attribut complexe, $q \in Q$ et $c \in C$,

$$sim_{cpt}(q, c) = w_q * \frac{2 * prof(LCS(q, c))}{prof(q) + prof(c)} \tag{1}$$

où w_q est le poids associé à q, $prof$ est la profondeur d'un concept (ou d'une instance) dans l'ontologie, et LCS est le plus petit subsumant commun (Least Common Subsumer) de deux instances. Dans un cas particulier, quand q et c représentent la même instance, nous avons : $prof(LCS(q, c)) = prof(q)$.

La similarité basée-slot, sim_{slt}, est définie comme suit :

$$sim_{slt}(q, c) = \frac{\sum_{s \in CS} sim(q.s, c.s)}{|CS|} \tag{2}$$

où CS est l'ensemble des attributs simples en commun entre q et c (Common Slots), $|CS|$ est sa cardinalité, $q.s$ (ou $c.s$) représente l'attribut simple s de q (ou de c), et $sim(q.s, c.s)$ est la similarité entre ces deux attributs. Pour le moment, nous considérons seulement les deux premiers modes (IGNORE, EXACT), et donc $sim(q.s, c.s)$ est définie comme suit :

$$sim(q.s, c.s) = \begin{cases} 1 & \text{si} \quad (w_{q.s} = exact) \wedge (v_{q.s} = v_{c.s}) \\ 0 & \text{sinon} \end{cases}$$

où $w_{q.s}$ est le mode associé à l'attribut $q.s$, et $v_{q.s}$ est la valeur de cet attribut dans q.

La mesure globale de similarité entre les deux attributs complexes, q et c, est définie par la formule suivante (Zhang *et al.*, 2006) :

$$sim(q, c) = (1 - \alpha) * sim_{cpt}(q, c) + \alpha * sim_{slt}(q, c) \tag{3}$$

où α est un paramètre d'expérience (actuellement, $\alpha = 0.4$).

Pour calculer la similarité basée-concept, chaque attribut complexe du cas cible est comparé à son attribut correspondant d'un autre cas source. Dans les bases de cas *homogènes*, tous les cas partagent la même structure prédéfinie, et donc la correspondance entre les attributs complexes des cas est déjà définie. En revanche, dans les bases de cas *hétérogènes*, cette correspondance n'est pas préalablement définie. En conséquence, avant de calculer la similarité basée-concept, il faut déterminer les attributs complexes correspondants.

4.2.2 Déterminer les attributs correspondants

Pour chaque attribut complexe $q' \in Q$, soit $c' \in C_j$ l'attribut complexe correspondant dans le cas $C_j \in \Omega$. Donc,

$$sim_{cpt}(q', c') = \max_{1 \leq l \leq m_j} (sim_{cpt}(q', c_{jl})) \tag{4}$$

Nous constatons que la prise en compte de la similarité maximale par rapport à un seul cas n'est pas suffisante en tant que telle. Car il se peut que c' soit l'attribut le plus similaire à q' dans le cas C_j alors qu'en réalité q' n'a aucun attribut correspondant dans C_j. Pour cela, il faut également comparer cette similarité avec la similarité maximale que l'on peut obtenir sur l'ensemble des cas, ce qui conduit à la satisfaction de la condition suivante :

$$\frac{sim_{cpt}(q', c')}{\max_{1 \leq j \leq k, 1 \leq l \leq m_j}(sim_{cpt}(q', c_{jl}))} \geq \beta \tag{5}$$

où β est un certain seuil (actuellement, $\beta = 0.7$).

4.2.3 Exemple

Prenons, par exemple, la partie suivante d'une description d'un cas : "Sur un site industriel, un capteur de gaz *infrarouge* a été installé pour détecter le *méthane*. D'autres gaz étaient présents sur le site dont *l'ammoniac*. Le capteur n'a pas bien fonctionné et une explosion s'est produite". Supposons maintenant une requête cherchant les cas où un capteur de gaz (peu importe sa technologie) a été utilisé pour détecter *l'hydrogène*. Nous avons alors :
$C =$ { capteur-infrarouge, méthane, ammoniac }
$Q =$ { capteur-de-gaz, hydrogène }
Supposons que $w_{hydrogene} = 1$, afin d'identifier l'attribut correspondant à l'hydrogène (de la requête), nous trouvons (voir FIG. 3, Formules (1), (4)) :

$sim_{cpt}(\text{hydrogène}_Q, \text{capteur-infrarouge}_C) = 0$
$sim_{cpt}(\text{hydrogène}_Q, \text{méthane}_C) = (2 * 3)/8 = 0.75$
$sim_{cpt}(\text{hydrogène}_Q, \text{ammoniac}_C) = 0.5$

Donc, le *méthane* est l'attribut qui correspond mieux à *l'hydrogène* (En ne considérant qu'un seul cas).

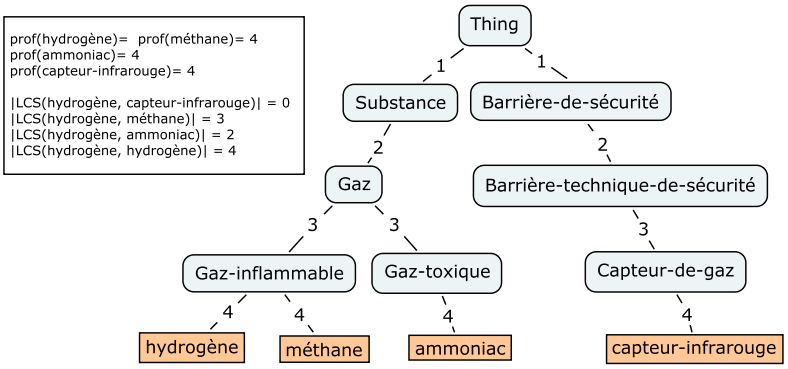


FIG. 3 – Partie du modèle de domaine.

4.2.4 Notion de rôle

Bien que la solution proposée pour déterminer la correspondance entre les attributs complexes soit efficace, elle donne parfois des résultats ambigus quand plusieurs attributs candidats sont proposés. Modifions l'exemple décrit précédemment : "l'un des gaz présents sur le site, en plus de l'ammoniac, était *l'hydrogène*" ; *i.e.* $C = \{$ capteur-infrarouge, méthane, ammoniac, hydrogène $\}$.

En suivant la méthode précédente, nous trouvons que l'hydrogène du cas C est l'attribut correspondant à l'hydrogène de Q. Toutefois, ce n'est pas le résultat souhaité puisque nous cherchons les cas où l'hydrogène était le gaz à détecter et non pas un gaz présent sur le site.

Pour lever cette ambiguïté, nous proposons de rajouter la notion de rôle ; *c'est-à-dire*, pour chaque attribut complexe pouvant jouer différents rôles dans un cas, son rôle doit être précisé dans chaque cas. En effet, le rôle représente dans l'ontologie une relation (de type objet) entre le cas et son attribut (dans cet exemple, $C \rightarrow$ *gaz-à-détecter* $\rightarrow$ méthane).

Pour trouver la correspondance entre les attributs complexes, nous combinons les deux approches : tout d'abord, les attributs ayant le même rôle sont identifiés ; ensuite, on applique la méthode proposée dans la section 4.2.2 aux autres attributs.

5 Plate-forme COBRA

COBRA est une plate-forme de RàPC indépendante du domaine. Elle permet la construction de systèmes de RàPC dont les connaissances sont décrites par des ontologies. La plate-forme est développée en JAVA en tant qu'application basée sur Eclipse[1], grâce au framework RCP[2] (Rich Client Platform). Elle profite ainsi de beaucoup de fonctionnalités offertes par Eclipse.

[1] http ://www.eclipse.org/

[2] http ://wiki.eclipse.org/index.php/Rich_Client_Platform

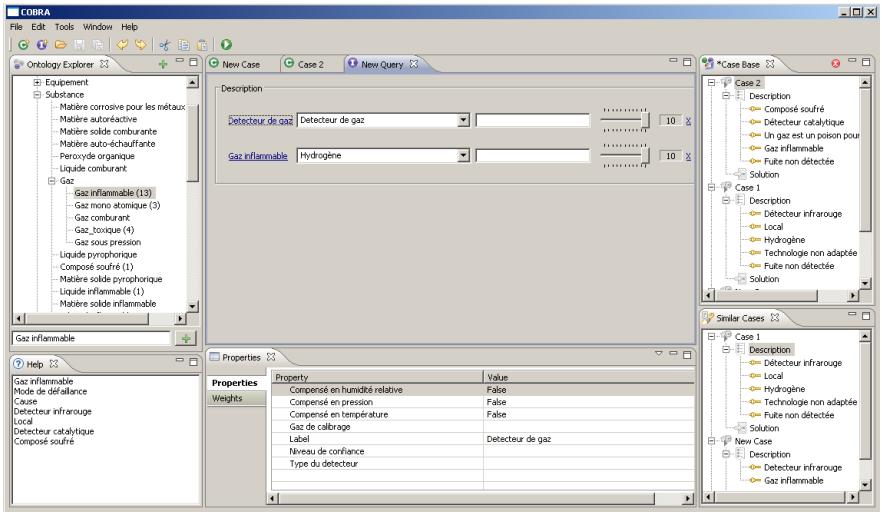


FIG. 4 – La plate-forme COBRA.

La plate-forme contient les onglets (viewers) principaux suivants (FIG. 4) : l'explorateur d'ontologie (en haut à gauche) qui montre les concepts de l'ontologie de domaine, la base de cas (en haut à droite), l'onglet des cas similaires par rapport à un cas cible, l'onglet d'aide qui propose à l'utilisateur les concepts candidats et les plus utilisés dans la base de cas, et l'onglet des propriétés permettant de modifier les valeurs de l'instance sélectionnée (d'un cas ou d'un attribut de cas).

L'authoring des cas se fait en créant un nouveau cas, puis en renseignant ses attributs simples et/ou en y ajoutant des attributs complexes (instances de l'ontologie). L'utilisateur peut choisir parmi les instances présentes, ou il peut créer et ajouter ses propres instances, ce qui permet d'enrichir la base de connaissances. Ensuite, l'utilisateur peut associer (ou créer) des rôles aux attributs complexes si nécessaire.

Comme le montre la figure 5, la plate-forme est basée sur JCOLIBRI, et nous avons étendu cette API par une couche qui permet le traitement de bases de cas dynamiques et hétérogènes. Cette couche contient notre propre connecteur d'ontologie ainsi que le module de calcul de similarité. Grâce à cette architecture, le développement d'un nouveau système de RàPC se fait en fournissant l'ontologie de domaine, et en paramétrant des fichiers XML de configuration.

6 Conclusion et perspectives

Nous avons présenté dans cet article notre plate-forme COBRA, une plate-forme de KI-CBR indépendante du domaine. Nous travaillons actuellement sur le domaine des barrières de sécurité, en particulier les capteurs de gaz, installées sur des sites industriels. Dans ce contexte, COBRA sera utilisé pour répondre à deux objectifs :

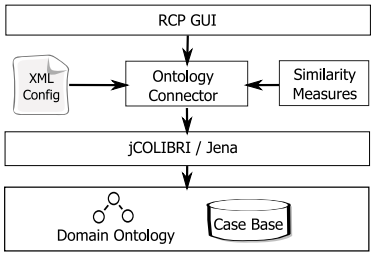


FIG. 5 – Les composantes de COBRA.

– Capitaliser les connaissances autour de la défaillance des capteurs de gaz.
– Fournir une aide aux experts et aux ingénieurs de sécurité pour pouvoir diagnostiquer les causes de défaillance des capteurs dans des conditions industrielles.

COBRA permet aux cas d'avoir différentes structures, et donc de travailler avec des bases de cas hétérogènes. Pour développer COBRA, nous nous sommes appuyés sur l'API jCOLIBRI. Cette API ne permet pas le traitement de bases de cas dynamiques et hétérogènes, car les attributs de cas doivent être définis auparavant (dans le code source) avant de lancer la plate-forme. Nous avons donc gardé les aspects intéressants de jCOLIBRI, et ajouté une couche supplémentaire afin de pallier ce manque.

L'hétérogénéité des cas complique la détermination de la correspondance entre les attributs complexes (d'un cas cible et d'autres cas sources). Dans cet article, nous avons proposé deux approches complémentaires pour résoudre ce problème basées sur les calculs de similarité et sur la définition des rôles des attributs dans leurs cas.

Actuellement, nous sommes en train de développer les autres processus de RàPC et d'étudier le lien qu'il peuvent avoir avec l'ontologie de domaine. Parallèlement, nous rajoutons de nouveaux cas (sur la défaillance de capteurs de gaz) à la base de cas. Nous allons également évaluer ce travail auprès d'experts de l'INERIS à deux niveaux :

– Le premier niveau concerne l'architecture proposée de RàPC, par exemple : à quel point la structure des cas et les processus de raisonnement sont-ils proches de l'activité réelle de l'expert ? Quels sont les concepts à rajouter à l'ontologie de domaine pour pouvoir décrire les nouveaux cas ? L'expert trouve-t-il les propositions d'aide intéressantes ?
– Le deuxième niveau concerne les solutions données par le système. Après l'ajout d'un certain nombre de cas, les experts doivent vérifier la qualité du diagnostic proposé par le système par rapport à certains cas cibles.

Références

AAMODT A. (1994). Explanation-Driven Case-Based Reasoning. *Lecture Notes In Computer Science*, p. 274–274.

AAMODT A. & PLAZA E. (1994). Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, **7**(1), 39–59.

ABOU ASSALI A., LENNE D. & DEBRAY B. (2008). Ontology development for industrial risk analysis. In *IEEE International Conference on Information & Communication Technologies : from Theory to Applications (ICTTA'08)*, Damascus, Syria.

BERGMANN R. & STAHL A. (1998). Similarity measures for object-oriented case representations. In *Proceedings of the European Workshop on Case-Based Reasoning, EWCBR'98*.

D'AQUIN M., LIEBER J. & NAPOLI A. (2006). *Artificial Intelligence : Methodology, Systems, and Applications*, volume Volume 4183/2006 of *Lecture Notes in Computer Science*, chapter Case-Based Reasoning Within Semantic Web Technologies. Springer Berlin / Heidelberg.

DÌAZ-AGUDO B. & GONZÀLEZ-CALERO P. (2000). An architecture for knowledge intensive CBR systems. *Advances in Case-Based Reasoning–(EWCBR'00). Springer-Verlag, Berlin Heidelberg New York*.

DÌAZ-AGUDO B. & GONZÀLEZ-CALERO P. (2002). CBROnto : a task/method ontology for CBR. *Procs. of the 15th International FLAIRS*, **2**, 101–106.

DÌAZ-AGUDO B., GONZÀLEZ-CALERO P., RECIO-GARCÌA J. & SÀNCHEZ-RUIZ-GRANADOS A. (2007). Building CBR systems with jcolibri. *Science of Computer Programming*, **69**(1-3), 68–75. Special issue on Experimental Software and Toolkits.

FUCHS B. & MILLE A. (2005). Une modélisation au niveau connaissance du raisonnement à partir de cas. In L'HARMATTAN, Ed., *Ingénierie des connaissances*.

LAMONTAGNE L. & LAPALME G. (2002). Raisonnement à base de cas textuels : Etat de l'art et perspectives. *Revue d'intelligence artificielle*, **16**(3), 339–366.

RECIO-GARCÌA J., DÌAZ-AGUDO B., GONZÀLEZ-CALERO P. & SANCHEZ A. (2006). Ontology based CBR with jCOLIBRI. *Applications and Innovations in Intelligent Systems*, **14**, 149–162.

RENAUD J., MORELLO B., FUCHS B. & LIEBER J. (2007). Raisonnement à Partir de Cas 1 : conception et configuration de produits. *Hermes-Lavoisier, February*, **1**.

RICHTER M. (2008). *Case-Based Reasoning on Images and Signals*, volume 73/2008 of *Studies in Computational Intelligence*, chapter Similarity, p. 25–90. Springer Berlin / Heidelberg.

RIESBECK C. & SCHANK R. (1989). *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates.

YANG C., FARLEY B. & ORCHARD B. (2008). Automated Case Creation and Management for Diagnostic CBR Systems. *Applied Intelligence : The International Journal of Artificial Intelligence, Neural Networks and Complex Problem-Solving Technologies*, **28**(1), 17–28.

ZHANG K., TANG J., HONG M., LI J. & WEI W. (2006). Weighted Ontology-Based Search Exploiting Semantic Similarity. *Lecture Notes In Computer Science*, **3841**, 498.

Démarches sémantiques de recherche d'information sur le Web

Olivier Corby[1], Catherine Faron-Zucker[2] et Isabelle Mirbel[1,2]

[1]EDELWEISS, INRIA Sophia Antipolis, Sophia Antipolis - France
[2]KEWI, I3S, Université de Nice Sophia - CNRS, Sophia Antipolis - France
olivier.corby@sophia.inria.fr,
catherine.faron-zucker@unice.fr, isabelle.mirbel@unice.fr

Type de communication : Recherche
Thèmes : Web sémantique

Résumé : L'examen de différents projets de recherche visant à supporter les activités des membres d'une communauté à l'aide d'une mémoire collective met en évidence l'intérêt de capitaliser les requêtes formulées à la mémoire et plus généralement les savoir-faire experts d'une communauté en matière de recherche d'information. L'enjeu est de donner les moyens aux membres d'une communauté de réutiliser et partager ces savoir-faire pour retrouver des informations précises et complètes par composition des résultats de requêtes sur différentes sources d'information. Dans cet article, nous proposons un modèle fondé sur les standards du web sémantique pour capitaliser, réutiliser et partager des séquences complexes de requêtes que nous appelons démarches de recherche. Notre modèle est le résultat d'une adaptation de la représentation intentionnelle de processus : nous explicitons les sous-buts qui gouvernent l'organisation d'une démarche de recherche et l'ordre selon lequel ces sous-buts doivent être satisfaits. Les démarches de recherche sont représentées en RDF et opérationalisées par des requêtes représentées en SPARQL. L'instanciation d'une démarche repose sur la mise en oeuvre d'un mécanisme de chaînage arrière sur ces règles.

Mots-clés : Web sémantique, SPARQL, RDF(S), requêtes, démarches de recherche

1 Introduction

Les mémoires collectives permettent de supporter les activités de capitalisation, gestion et diffusion des connaissances au sein d'une communauté. Les ressources de la communauté y sont indexées par des annotations sémantiques qui explicitent et formalisent leur contenu informatif. Les membres d'une communauté exploitent alors leur mémoire communautaire en exprimant des requêtes leur permettant de rechercher des ressources pertinentes pour mener à bien leurs activités. La recherche des ressources de la mémoire qui répondent à une requête repose sur la manipulation formelle des annotations des ressources et peut être guidée par des ontologies du domaine.

L'examen de projets de recherche visant à assister les activités de communauté par une mémoire collective auxquels nous avons participé tels que le projet européen SevenPro (Cherfi *et al.*, 2008), les projets ANR e-WOK HUB (eWOKHUB Consortium, 2008) et Immunosearch (Kefi-Khelif *et al.*, 2008) ou le projet C3R (Yurchyshyna *et al.*, 2008) met en évidence le besoin de capitaliser les requêtes formulées par les utilisateurs dans une base idoine afin de permettre à leurs auteurs de les réutiliser, voire de les échanger avec les autres membres de la communauté. Plus généralement, la capitalisation des démarches de recherche d'information devient un véritable enjeu dans de nombreux domaines. En effet, des stratégies précises sont mises en oeuvre par les experts d'un domaine pour retrouver l'information nécessaire à leurs activités (Bhavnani *et al.*, 2003) et elles sont souvent difficiles à acquérir par des utilisateurs néophytes. Ces stratégies, de plus en plus critiques de par la spécialisation et la multiplication des bases de connaissances, sont rarement capitalisées et ne sont exploitées ni dans les outils de recherche comme Google ni dans les portails de domaine (Bhavnani *et al.*, 2003).

(Bhavnani *et al.*, 2003) proposent une approche pour expliciter les procédures critiques de recherche d'information dans le domaine médical à l'aide de ce qu'ils appellent des portails de stratégies de recherche (≪strategy hubs≫). En partant d'un ensemble de questions types du domaine, ils définissent un ensemble de patrons représentant des procédures de recherche. Une procédure de recherche est représentée par un ensemble ordonné de sous-buts et pour chaque procédure de recherche, des liens vers les sources d'information pertinentes sont établis.

Les auteurs de (Buffereau *et al.*, 2003) proposent un environnement de navigation parmi des ressources Web qui distingue 3 niveaux de connaissances : (i) un niveau support de la connaissance qui rassemble les ressources du Web dans le domaine d'application, (ii) un niveau représentation de la connaissance qui rassemble les méta-données relatives aux ressources du niveau précédent et (iii) un niveau transmission de la connaissance qui propose des parcours (appelés ≪e-parcours≫) pour exploiter les ressources du Web au travers des méta-données qui leur sont associées. Les *e-parcours* sont composés d'étapes caractérisées par une intention ou un titre, un sujet et une illustration. Les illustrations sont les ressources Web relatives à l'étape du *e-parcours*.

Des exemples de procédures de recherche permettant une navigation dynamique existent également dans les systèmes d'apprentissage en ligne fondés sur les modèles du Web sémantique. Dans (Yessad *et al.*, 2008; Dehors & Faron-Zucker, 2006), nous modélisons l'approche pédagogique adoptée par un assemblage de requêtes paramétrées et les ressources dont les annotations répondent à ces requêtes composent dynamiquement des documents pédagogiques présentés à l'apprenant qui navigue dans le système. Dans nos travaux appliqués au domaine du bâtiment, nous avons explicité auprès des experts du domaine des procédures de vérification de conformité que nous représentons par un ordonnanceur des requêtes de conformité (Yurchyshyna *et al.*, 2008) auxquelles est soumise l'annotation de la maquette numérique d'un projet de construction à contrôler. L'ordonnancement dépend des classes auxquelles appartiennent les requêtes à mettre en oeuvre (types de textes réglementaires, types de bâtiments, parties de la maquette visée), c'est-à-dire que l'ordonnancement des classes de requêtes est fixe.

L'approche que nous proposons ici consiste en un modèle pour capitaliser, réutiliser et partager des requêtes de recherche d'information et pour les organiser en des démarches

de recherche formalisées également réutilisables et partageables. Comme dans les travaux cités plus haut, exceptés ceux de (Bhavnani *et al.*, 2003), notre approche repose sur les techniques et modèles du Web sémantique. Nous tirons parti des capacités d'inférence des ontologies capitalisant la connaissance d'un domaine d'application et nous utilisons le language SPARQL qui permet de représenter des requêtes plus riches que les patrons proposés dans l'approche de (Bhavnani *et al.*, 2003). Comme dans (Buffereau *et al.*, 2003), nous proposons un modèle pour formaliser les buts et les sous-buts, qui sont exprimés en language naturel dans (Bhavnani *et al.*, 2003). Enfin, notre proposition diffère de celle de (Buffereau *et al.*, 2003) dans le sens où elle permet la réutilisation et le partage d'étapes (i.e. de sous-buts) entre les procédures de recherche, alors que l'approche présentée dans (Buffereau *et al.*, 2003) positionne le partage uniquement au niveau des ressources Web. Précisément, nous nous intéressons à la construction et l'exploitation d'une base de requêtes rendant opérationnelles les étapes d'un scénario de recherche d'information que nous appelons démarche de recherche d'information. Cette base de connaissances peut être vue comme une mémoire épisodique dans laquelle les démarches de recherche sont construites dynamiquement en fonction du contexte.

La suite de l'article est organisée de la façon suivante. Dans la section 2 nous définissons la notion de démarche de recherche d'information et nous expliquons notre choix d'une modélisation intentionnelle de telles démarches. Dans la section 3, nous détaillons comment nous avons adapté un modèle intentionnel de représentation des processus au domaine du Web sémantique, au travers de la proposition d'une ontologie idoine permettant l'annotation de démarche de recherche d'information. Dans la section 4, nous présentons la façon dont les démarches de recherche sont mises en oeuvre à l'aide de règles.

2 Des démarches intentionnelles pour une compréhension globale d'une procédure de recherche

Nous définissons la notion de démarche de recherche d'information comme une séquence de recherches atomiques qui doivent être effectuées par un expert du domaine pour mener à bien une tâche ou un processus métier. Une démarche de recherche peut être vue comme un type particulier de processus métier constitué exclusivement d'activités de recherche d'information. Différents modèles de représentation des processus métiers ont été proposés dans la littérature (Nurcan & Edme, 2005).

Notre travail porte sur des moyens de capitaliser et d'expliciter des démarches de recherche permettant d'avoir une compréhension globale d'un sujet à partir de sources d'information dispersées dans différentes bases de connaissances et de données.

Pour favoriser le transfert de connaissances en matière de recherche d'information des experts vers les néophytes au sein d'une communauté, nous cherchons à nous appuyer sur un modèle de représentation des démarches ayant les caractéristiques suivantes :

– Une représentation modulaire des démarches de recherche d'information afin de favoriser leur partage et leur réutilisation,
– Une représentation des intentions (c'est-à-dire du pourquoi) de la recherche afin

de favoriser le transfert de connaissances des experts vers les néophytes et ainsi permettre à ces dernier de comprendre de façon globale un sujet sur lequel aucune source de données ne contient l'ensemble des informations pertinentes,
– Une représentation de la connaissance sur les démarches à plusieurs niveaux d'abstraction afin de prendre en considération les différents niveaux d'expertise des membres de la communauté.

Pour cela, l'approche originale que nous proposons de mettre en oeuvre est fondée sur l'adaptation au Web sémantique de la représentation intentionnelle d'un processus proposé dans (Rolland *et al.*, 1999; Rolland, 2007).

2.1 Le modèle de carte

D'après (Rolland *et al.*, 1999; Rolland, 2007), une carte est un modèle de processus dans lequel un ordonnancement non déterministe d'intentions et de stratégies de réalisation de ces intentions est représenté. Dans notre cas, nous nous concentrons sur des intentions et des stratégies de recherche. Une carte est un graphe nommé orienté ayant des intentions pour noeuds et des stratégies pour arcs entre les intentions. Une intention de recherche représente un but qui peut être atteint en suivant une stratégie de recherche. Une intention exprime ce qui est voulu (un état, un résultat) indépendamment de par qui, quand et où l'intention est réalisée. Deux intentions particulières sont distinguées : l'intention de début de démarche (*début*) et l'intention de fin de la démarche (*fin*). Une carte consiste donc en un ensemble de sections, chacune d'elle représentée par un triplet (intention source, stratégie, intention cible). Une stratégie représente la façon dont l'intention cible est réalisable à partir d'une intention source. Une carte contient un ensemble fini de chemins de l'intention *début* à l'intention *fin*, chacun décrivant une façon de satisfaire le but de la démarche de recherche d'information décrite.

La figure 1 montre un exemple de démarche pour rechercher des informations sur les bases de données relationnelles, par exemple dans le cadre de la conception d'un cours sur ce sujet dans une communauté d'enseignants. Pour mener à bien l'intention globale de la démarche (rechercher des ressources sur les bases de données relationnelles) il est recommandé de décomposer la recherche suivant les intentions présentées dans la figure 1. On peut remarquer que l'intention de recherche de ressources sur l'historique des bases de données relationnelles est facultative (il existe deux chemins de l'intention *début* à l'intention *fin*, l'un incluant cette intention, l'autre non) et que deux stratégies sont proposées pour réaliser l'intention de chercher des ressources sur le pilotage d'une base de données relationnelle depuis un langage de programmation.

Toujours d'après (Rolland *et al.*, 1999), à chaque section de la carte correspond une directive de réalisation d'intention (DRI) qui fournit des moyens opérationnels ou intentionnels de réaliser l'intention cible. Dans notre approche, une DRI opérationnelle correspond à l'exécution d'une requête SPARQL sur la mémoire de la communauté et une DRI intentionnelle est représentée par une carte définissant de façon plus détaillée (i.e. décomposant l'intention cible en sous-buts) la stratégie permettant de réaliser l'intention cible. La figure 2 montre un exemple de DRI intentionnelle pour la section mise en évidence dans la figure 1 et un exemple de DRI opérationnelle pour la section ayant l'intention *début* comme intention source dans la DRI intentionnelle de la figure 2.

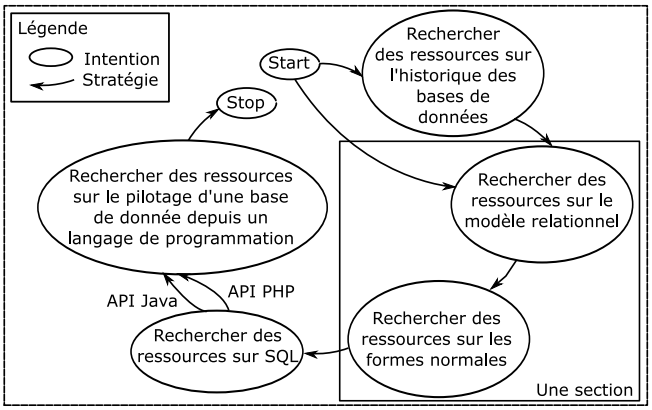


FIGURE 1 – Exemple de démarche

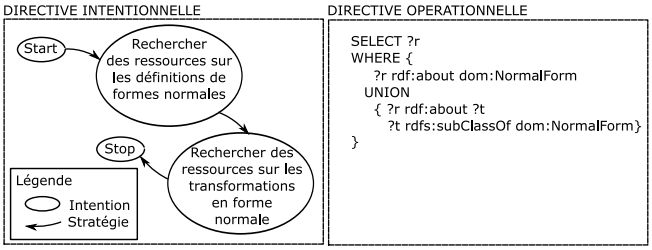


FIGURE 2 – Exemple de DRI

2.2 Intentions et stratégies

Pour permettre la réutilisation et le partage de requêtes de recherche d'information (i.e. DRI opérationnelles) et de démarches de recherche d'information (i.e. DRI intentionnelles), nous formalisons les notions de stratégie et d'intention. Pour cela, nous nous appuyons sur les travaux présentés dans (Prat, 1999), qui se sont déjà montrés pertinents pour formaliser des buts (Ralyte, 2001; Guzelian, 2007; Rolland, 2007). D'après (Prat, 1999), une intention est caractérisée par un verbe et des paramètres qui jouent un rôle particulier vis-à-vis du verbe. Parmi ces paramètres, il y a notamment l'objet sur lequel porte l'action décrite par le verbe. Dans (Ralyte, 2001; Rolland, 2007) d'autres paramètres sont proposés tels que le bénéficiaire du résultat de la réalisation de l'intention, l'objet à partir duquel l'action est réalisée ou une propriété qui devra être préservée durant la réalisation de l'intention. Ces paramètres permettent de formaliser la notion de stratégie. Dans l'exemple de la figure 1, les intentions sont spécifiées à l'aide du verbe *Search* et des objets du domaine, par exemple *Normal Form* et *Relational Model* ; les stratégies à l'aide des paramètres *Java API* et *PHP API*.

3 Démarches sémantiques de recherche d'information

Nous considérons les démarches comme des ressources de la communauté et, à ce titre, nous les annotons afin de les indexer dans la mémoire collective de la communauté. Au-delà de la proposition d'un moyen alternatif d'organiser et d'accéder aux ressources d'une mémoire communautaire à travers la spécification de démarches décrivant des stratégies d'accès aux ressources pertinentes, l'annotation des démarches est un moyen de capitaliser la connaissance intrinsèque aux démarches elles-mêmes.

3.1 Une ontologie pour les démarches sémantiques de recherche d'information

Nous avons construit une ontologie RDFS pour annoter les démarches de recherche d'information qui rassemble les concepts et les relations du modèle de carte et ceux qui participent à la définition des notions d'intention et de stratégie telles que définies plus haut.

On retrouve comme classes principales de cette ontologie *Section*, *Intention*, *IntentionAchievementGuideline* qui représente une DRI, et *Resource* qui représente les ressources faisant l'objet du processus de recherche. Les intentions de début et de fin de démarche apparaissent comme sous-classes de la classe *Intention*. Une section est constituée d'une intention source, d'une intention cible et d'une stratégie (propriétés *hasTarget*, *hasSource* et *hasStrategy*). Des ressources Web sont dynamiquement associées aux sections dont elles permettent de satisfaire les intentions à l'aide de la propriété *hasResource*. Chaque section est rendue opérationnelle par une DRI (propriété *operationalizedBy*). Une intention est spécifiée à l'aide d'un verbe et d'un objet (classes *Verb* et *Object* et des propriétés *hasVerb* et *hasObject*). La classe *IntentionAchievementGuideline* est spécialisée en deux sous-classes *Map* et *GenericQuery* qui traduisent respectivement les notions de DRI intentionnelle et de DRI opérationnelle introduites précédemment.

Nous n'exploitons pour l'instant qu'un seul verbe, correspondant à la classe *Search* qui est instance de la (méta-) classe *Verb*, et nous considérons les concepts du domaine d'application de la recherche d'information comme des instances de la (méta-) classe *Object*. Les sous-classes de la classe *Parameter* sont instanciées en différentes classes qui modélisent une forme de contexte dans les processus de recherche d'information. Nous distinguons les informations contextuelles qui dépendent du domaine des informations recherchées de celles qui en sont indépendantes. Par exemple, les classes *Neophyte* et *Expert*, instances de la (méta-) classe *Beneficiary* sont donc toutes deux indépendantes du domaine : elles indiquent à quel type de membre le résultat de l'exécution de la DRI associée à une section de démarche est dédié. De même, les classes *DetailedDescription* et *ShortDescription*, instances de la classe *Quality*, sont indépendantes du domaine : elles indiquent si le résultat de l'exécution de la DRI associée à une section doit être constitué de ressources décrivant de façon détaillée ou non les informations résultat de la recherche. Au contraire, les classes *Java API* et *PHP API* instances de la classe *Manner* sont des exemples de concepts dépendants du domaine.

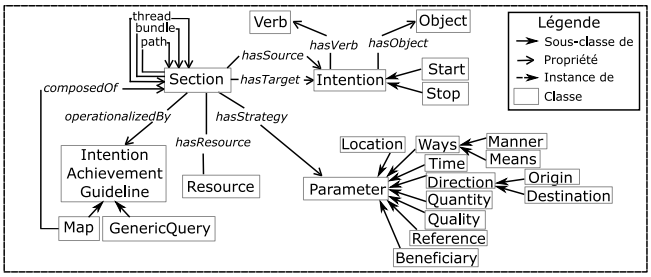


FIGURE 3 – Une ontologie pour les démarches de recherche d'information

Nous travaillons actuellement à la construction d'une ontologie indépendante du domaine d'application pour enrichir les extensions des concepts *Verb* et *Parameter*.

La figure 3 montre l'ontologie que nous avons construite pour annoter les démarches de recherche d'information.

3.2 Représentation des démarches de recherche d'information

Nous représentons les démarches de recherche d'information par des annotations RDF qui reposent sur l'ontologie décrite ci-dessus. Nous exploitons les capacités de raisonnement qu'offre le modèle RDF pour organiser les démarches de recherche d'information et pour les retrouver afin de les réutiliser.

Sur la carte décrite dans la figure 2, la section ayant pour source l'intention intitulée ≪Rechercher des ressources sur les définitions de formes normales≫ et pour cible celle intitulée ≪Rechercher des ressources sur les transformations en forme normale≫ est représentée en RDF de la façon suivante où `map` correspond au namespace de l'ontologie pour les démarches de recherche d'information et `dom` celui de l'ontologie de domaine :

```
<rdf:RDF xmlns:rdf="..." xmlns:map="..."xmlns:dom="...">
 <map:Section>
   <map:hasSource>
     <map:Intention rdf:nodeID="ii">
       <map:hasVerb rdf:resource="&dom;Search"/>
       <map:hasObject rdf:resource="&dom;NFDefinition"/>
     </map:Intention>
   </map:hasSource>
   <map:hasTarget>
     <map:Intention rdf:nodeID="ij">
       <map:hasVerb rdf:resource="&dom;Search"/>
       <map:hasObject rdf:resource="&dom;NFRule"/>
     </map:Intention>
   </map:hasTarget>
 </map:Section>
</rdf:RDF>
```

3.3 Partage et réutilisation de démarches

3.3.1 Des fragments de démarches

Une DRI et la section à laquelle elle est associée forment ce que nous appelons un fragment de démarche de recherche d'information. La définition de la section constitue la signature du fragment. Elle précise l'intention source, l'intention cible et la stratégie du fragment (seule la spécification de l'intention cible est obligatoire). Les intentions sont spécifiées à l'aide d'instances de la classe *Object* et d'instances de la classe *Verb*. La stratégie est spécifiée à l'aide d'instances de la classe *Parameter*.

La DRI constitue le corps du fragment de démarche. Elle est soit intentionnelle soit opérationnelle. Une DRI intentionnelle est une carte permettant de satisfaire l'intention cible considérée dans la signature du fragment, en partant éventuellement de l'intention source si elle est présente dans la signature du fragment et en suivant une stratégie éventuellement décrite dans la signature du fragment. Une DRI opérationnelle est une requête permettant de retrouver les ressources pertinentes dans la mémoire de la communauté. Les fragments correspondants aux DRIs présentées dans la figure 2 sont représentés dans la figure 4.

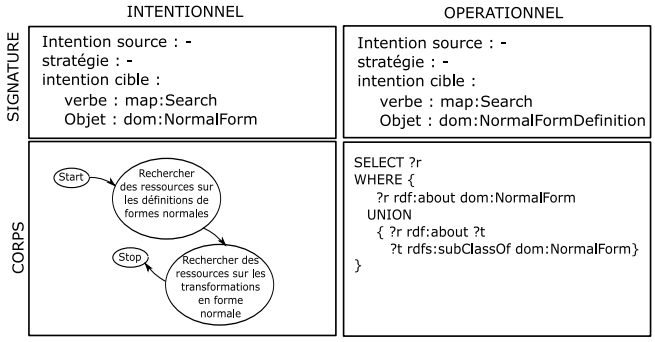


FIGURE 4 – Exemples de fragments et de règles

Lorsqu'un membre expert de la communauté souhaite conserver dans la mémoire de la communauté une recherche d'information, il définit sa démarche sous forme d'un ensemble de sections reliées les unes aux autres et chacune opérationnalisée à l'aide d'au moins un fragment, nouvellement créé pour l'occasion ou réutilisé parmi les fragments déjà présents dans la mémoire de la communauté. Les DRI intentionnelles des nouveaux fragments devront à leur tour être rendues opérationnelles jusqu'à proposer des DRI opérationnelles pour tous les sous-buts de la démarche.

Un fragment de démarche peut être réutilisé d'une démarche à l'autre si ces démarches partagent une même section ou des éléments de section. En effet, deux démarches incluant la réalisation d'une même intention peuvent être partiellement opérationnalisées par un même fragment dont la signature ne porte que sur la réalisation de l'intention cible (et ne contraint ni la situation de départ ni la stratégie de réalisation de l'intention).

La façon dont nous concevons la réutilisation de démarches de recherche d'infor-

mation est fondée sur la connexion dynamique de différents fragments pour rendre opérationnelle une démarche dans son ensemble en combinant les fragments de démarche d'après leur signature. Par exemple, un fragment ayant une intention cible *I* peut se combiner avec tout fragment ayant pour intention source *I*. Plusieurs fragments ayant *IS* pour intention source, *S* pour stratégie et *IC* pour intention cible peuvent être associés à une même section de démarche de recherche d'information et ainsi constituer autant de moyens différents de rendre opérationnelle la section de démarche considérée, c'est-à-dire de trouver un ensemble de ressources pertinentes dans la mémoire sémantique de la communauté.

3.3.2 Des fragments aux règles

Dans notre approche, nous implémentons chaque fragment de démarche sous forme d'une règle dont la conclusion correspond à la signature du fragment (une section de carte) et dont la prémisse correspond au corps du fragment (c'est-à-dire soit une requête SPARQL dans le cas d'une DRI opérationnelle soit une carte dans le cas d'une DRI intentionnelle). Nous distinguons les règles concrètes dont la prémisse est constituée d'une requête SPARQL des règles abstraites dont la prémisse est constituée d'une carte. Le langage SPARQL constitue un cadre unifié pour représenter règles concrètes et règles abstraites en reposant sur la forme de requête CONSTRUCT-WHERE. La clause WHERE d'une telle requête peut être vue comme la prémisse d'une règle et la clause CONSTRUCT comme sa conclusion. Une telle requête construit des graphes RDF en substituant aux variables de sa clause CONSTRUCT les valeurs qui satisfont sa clause WHERE (retrouvées en recherchant les appariements possibles de sa clause WHERE avec les données RDF interrogées). Ainsi, nous représentons un fragment de démarche par une requête SPARQL dont la clause CONSTRUCT est un patron de graphe permettant de construire la représentation RDF de la section de la carte considérée et dont la clause WHERE est un patron de graphe qui représente soit une carte (règle abstraite) soit un critère de recherche de ressources pertinentes (règle concrète). Dans le cas d'une règle abstraite, la clause WHERE est un patron de graphe représentant une sous-carte permettant la réalisation de l'intention cible du fragment. Dans le cas d'une règle concrète, la clause WHERE est un graphe qui permet de retrouver les ressources pertinentes, c'est-à-dire celles avec les annotations RDF desquelles il existe des appariements avec ce graphe requête. Les règles correspondant aux fragments présentés dans la figure 4 sont représentées dans la figure 5.

Une section de carte peut être implémentée par différentes requêtes, concrètes ou abstraites, afin de proposer différentes possibilités de recherche à différents niveaux d'abstraction. Ces requêtes partagent la même clause CONSTRUCT (fragments de même signature) et diffèrent par le contenu de leur clause WHERE.

La mise en oeuvre d'une démarche de recherche d'information, i.e. la combinaison de fragments de démarche en une démarche globale, est réalisée en appliquant les règles qui implémentent les fragments de démarche en chaînage arrière. Nous nous appuyons sur le moteur sémantique CORESE [1](Corby *et al.*, 2006) à la fois pour le chaînage arrière sur la base de requêtes SPARQL représentant des règles et pour trouver les réponses à

1. http ://www-sop.inria.fr/edelweiss/software/corese/

REGLE ABSTRAITE

```
CONSTRUCT {
  _:s map:hasTarget _:i
  _:i map:hasObject d:NormalForm
  _:s map:operationalizedBy ?g
}
WHERE {
  GRAPH ?g {
    ?s1 map:hasSource ?i0
    ?i0 rdf:type map:Start
    ?s1 map:hasTarget ?i1
    ?i1 map:hasObject d:NormalFormDef
    ?s2 map:hasSource ?i1
    ?s2 map:hasTarget ?i2
    ?i2 map:hasObject d:NormalizationRule
    ?s3 map:hasSource ?i2
    ?s3 map:hasTarget ?i4
    ?i4 rdf:type map:Stop }
}
```

REGLE CONCRETE

```
CONSTRUCT {
  _:s map:hasTarget _:i
  _:i map:hasObject d:NormalFormDefinition
  _:s map:hasResource ?r
}
WHERE {
    ?r rdf:about dom:NormalForm
 UNION
  { ?r rdf:about ?t
    ?t rdfs:subClassOf dom:NormalForm }
}
```

FIGURE 5 – Exemples de fragments et de règles

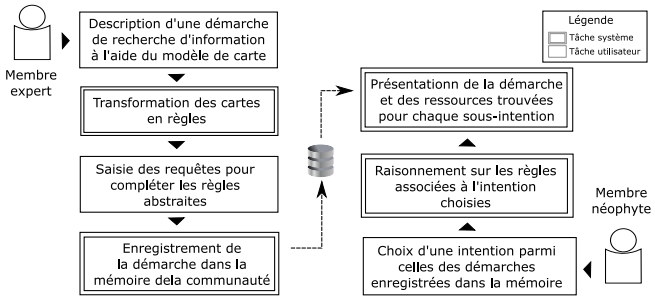


FIGURE 6 – Processus de définition et de mise en oeuvre de démarches

une requête dans la base d'annotations RDF des ressources de la mémoire sémantique de la communauté. Soulignons que seules les requêtes sont capitalisées dans la mémoire de la communauté et non les cartes. Ces dernières sont créées dynamiquement tout au long du processus de chaînage arrière, comme des sous-buts temporaires, jusqu'à ce que toutes les resources dont les annotations correspondent à des sous-buts et donc à l'intention globale de la démarche de recherche d'information soient retrouvées. Un membre de la communauté qui souhaite trouver une démarche de recherche d'information sur un sujet donné doit en fait rechercher la règle qui correspond à l'intention de recherche qu'il souhaite satisfaire (intention globale) et bénéficie de toutes les règles et de toutes les annotations de ressources présentes dans la mémoire sémantique de la communauté au moment de la mise en oeuvre de la démarche. La mémoire de la communauté évolue au cours du temps et ainsi l'ensemble des ressources retournées peut également varier dans le temps. En d'autres termes, l'instantiation de la démarche de recherche d'information dans son ensemble est réalisée au moment de son exécution et dépend donc des ressources disponibles dans la mémoire collective à ce moment-là. Le processus de définition d'une démarche par un expert et le processus de mise en oeuvre d'une démarche par un néophytes sont schématisés dans la figure 6.

4 Conclusion

Dans cet article nous avons présenté une approche, fondée sur les techniques et modèles du Web sémantique, qui a pour but de capitaliser, partager et réutiliser des requêtes et des démarches de recherche d'information. En proposant une modélisation intentionnelle et sémantique des démarches de recherche d'information, notre but est de capitaliser de la connaissance sur les pratiques en recherche d'information au sein d'une communauté et cela au travers de séquences structurées de tâches de recherche. Pour cela, en partant d'un formalisme de représentation des processus guidé par les intentions, nous avons proposé une ontologie permettant d'annoter des démarches de recherche d'information et nous avons opérationnalisé les directives correspondant aux fragments de ces démarches à l'aide de règles implémentées par des requêtes SPARQL. Les démarches de recherche d'information sont mises en oeuvre à l'aide d'un mécanisme de chaînage arrière appliqué sur la base de règles et sur les annotations RDF des ressources de la communauté.

Cinq perspectives se dessinent à l'issue de ce travail :

- La proposition d'une ontologie de concepts indépendants du domaine d'application pour enrichir l'ensemble des classes instances de *Verb* et *Parameter* dans l'ontologie des démarches ;
- La proposition d'un modèle de requête générique permettant la représentation de requêtes récurrentes dans les fragments de démarche ;
- La prise en compte des profils des membres de la communauté et du contexte de mise en oeuvre de la démarche afin d'affiner le processus de sélection des fragments de démarche lors de l'opérationnalisation d'une démarche de recherche d'information ;
- La proposition d'une ontologie et d'un modèle pour annoter les démarches et ainsi faciliter leur partage au sein de la communauté, notamment en fonction des profils des membres et de leurs contextes de travail.
- Le développement d'un outil pour mettre en oeuvre cette approche et la proposition d'interfaces idoines pour aider les membres de la communauté à naviguer dans la mémoire sémantique dédiée aux stratégies de recherche.

Ce travail est le point de départ du projet DESIR (COLOR INRIA) qui démarre cette année en partenariat avec deux équipes de l'INRA et qui vise à l'explicitation et la capitalisation des processus de recherche d'information d'agronomes et généticiens sur différentes bases de données hétérogènes.

Références

BHAVNANI S., BICHAKJIAN C., JOHNSON T., LITTLE R., PECK F., SCHWARTZ J. & STRECHER V. (2003). Strategy hubs : Next-generation domain protals with search procedures. In G. COCKTON & P. KORHONEN, Eds., *ACM Conference on Human Factors in Computing Systems*, p. 393–400, Florida, USA : ACM.

BUFFEREAU B., DUCHET P. & PICOUET P. (2003). Generating guided tours to facilitate learning from a set of indexed resources. In *IEEE International Conference on*

Advanced Learning Technologies (ICALT), p. 492, Athens, Greece : IEEE Computer Society.

CHERFI H., CORBY O., FARON-ZUCKER C. & KHELIF K. (2008). Semantic annotation of texts with RDF graph contexts. In P. W. EKLUND & O. HAEMMERLÉ, Eds., *International Conference on Conceptual Structures (ICCS'2008)*, p. 75–82, Krakow, Poland : CEUR-WS.org.

CORBY O., DIENG-KUNTZ R., FARON-ZUCKER C. & GANDON F. (2006). Searching the semantic web : Approximate query processing based on ontologies. *IEEE Intelligent Systems Journal*, **21**(1).

DEHORS S. & FARON-ZUCKER C. (2006). Qbls : A semantic web based learning system. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*, Orlando, FL, USA.

EWOKHUB CONSORTIUM (2008). Semantic hubs for geographical projects. In CEUR-WS.ORG, Ed., *Semantic Metadata Management and Applications (SeMMA), workshop at ESWC*, p. 3–17, Teneriffe, Spain : Khalid Belhajjame and Mathieu d'Aquin and Peter Haase and Paolo Missier.

GUZELIAN G. (2007). *Modelisation et specification de composants réutilisables pour la conception de systèmes d'information*. PhD thesis, Université Aix Marseille.

KEFI-KHELIF L., DEMARCHEZ M. & COLLARD M. (2008). A knowledge base approach for genomics data analysis. In *International Conference on Semantic Systems, I-Semantics 2008*, Graz, Austria.

NURCAN S. & EDME M. (2005). Intention-driven modeling for flexible workflow applications. *Software Process : Improvement and Practice*, **10**(4), 363–377.

PRAT N. (1999). *Réutilisation de la trace par apprentissage dans un environnement pour l'ingénierie des processus*. PhD thesis, Université Paris I - Sorbonne.

RALYTE J. (2001). *Ingénierie des méthodes à base de composants*. PhD thesis, Université Paris I - Sorbonne.

ROLLAND C. (2007). *Conceptual Modelling in Information Systems Engineering*, chapter Capturing System Intentionality with Maps. Springer-Verlag.

ROLLAND C., PRAKASH N. & BENJAMEN A. (1999). A multi-model view of process modelling. *Requirements Engineering*, **4**(4), 169–187.

YESSAD A., FARON-ZUCKER C., DIENG-KUNTZ R. & LASKRI M. (2008). Ontology-driven adaptive course generation for web-based education. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*, Vienna, Austria.

YURCHYSHYNA A., FARON-ZUCKER C., MIRBEL I., SALL B., LE THANH N. & ZARLI A. (2008). Une approche ontologique pour formaliser la connaissance experte dans le modèle du contrôle de conformité en construction. In Y. T. YANNICK PRIÉ, Ed., *19ième journées francophones d'ingénierie des connaissances*, Nancy, France : Capaudes Editions.

Modélisation systématique de recommandations de pratique clinique : une étude théorique et pratique sur la prise en charge de l'hypertension artérielle

Brigitte Séroussi[1], Jacques Bouaud[2], Denké L. Denké[1], Jacques Julien[3], Hector Falcoff[4]

[1] Université Paris 6, UFR de Médecine, Paris, France; AP-HP, Hôpital Tenon, Département de Santé Publique, Paris, France; Université Paris 13, UFR SMBH, LIM&BIO, Bobigny, France; APREC, Paris, France.
[2] AP-HP, DSI, STIM, Paris, France; INSERM, UMR_S 872, eq. 20, Paris, France.
[3] AP-HP, HEGP, Service de Médecine Vasculaire et Hypertension Artérielle, Paris, France.
[4] SFTG, Paris, France; Université Paris 5, Faculté de Médecine, Département de Médecine Générale, Paris, France.
brigitte.seroussi@tnn.aphp.fr, jacques.bouaud@sap.aphp.fr

Résumé : Les recommandations de pratique clinique (RPC) cherchent à promouvoir les résultats d'une « médecine fondée sur les faits » pour améliorer la qualité des soins. L'implémentation des RPC dans des systèmes d'aide à la décision médicale (SADM) permet d'améliorer la mise en oeuvre, en pratique, des recommandations par les médecins. Nous avons développé un SADM, ASTI-MG, sur la base des RPC pour la prise en charge de l'hypertension artérielle. Nous avons utilisé la modélisation pour caractériser formellement le statut en terme de niveau de preuve des stratégies thérapeutiques associées à chacun des profils patient. Trois statuts ont été identifiés : les recommandations avec niveau de preuve, celles fondées sur le consensus de professionels, et les stratégies établies après avis d'expert hors RPC. Nous avons étudié la distribution du statut des stratégies thérapeutiques associées aux profils « théoriques » de la base de connaissances d'ASTI-MG et nous l'avons comparée à celle réalisée par un échantillon de 435 profils « pratiques » de patients réels. Les recommandations avec niveau de preuve représentent 0,5 % des possibles pour 8,3 % des pratiques. Presque la moitié des patients (44,8 %) sont dans des situations non couvertes par les RPC. Ceci pose la question de l'aide que peuvent fournir les RPC et celle de l'acceptabilité des propositions des SADM.

Mots-clés : modélisation des connaissances, recommandations de pratique clinique, aide à la décision, *evidence-based medicine*, hypertension artérielle, étude de pratiques

1 Introduction

Afin d'optimiser la qualité des soins, les médecins doivent mettre en œuvre les principes de la « médecine fondée sur les faits » ou *evidence-based medicine* (EBM) en reprenant le terme anglo-saxon couramment utilisé. Ces principes se définissent comme *« l'utilisation explicite, judicieuse et consciencieuse des dernières données issues de la recherche disponibles au moment de la prise de décisions concernant les soins à prodiguer à un patient donné ou à des populations »* (Sackett *et al.*, 1996).

Les « recommandations de pratique clinique » (RPC) ou « guides de bonnes pratiques » (GBP) sont élaborées dans un objectif de synthèse des productions de l'EBM sur une problématique médicale donnée. Ainsi les RPC sont des sources d'information dans lesquelles les résultats de la recherche clinique sont gradés selon leur validité par des experts méthodologistes puis modulés par des cliniciens experts en fonction de leur pertinence pour la pratique médicale. Produits sous l'égide de sociétés savantes ou d'agences nationales de santé, comme en France la Haute Autorité de Santé, les RPC sont des documents structurés répertoriant un ensemble de situations particulières pour lesquelles certains plans d'actions, dans le registre de la prévention, du diagnostic ou de la thérapeutique, sont recommandés avec un grade donné traduisant la force de la recommandation. Afin d'assurer la validité et la fiabilité nécessaires à leur mise en application, les RPC doivent être fondées sur les faits et le statut, en terme de « niveau de preuve », de chaque recommandation doit être explicite. Malheureusement, une faible proportion des recommandations repose sur les résultats d'essais cliniques randomisés (grade A), d'essais cliniques de faible puissance ou de cohortes (grade B), ou enfin de séries de cas témoins (grade C). En l'absence d'étude, la majeure partie des propositions thérapeutiques rapportées dans les RPC repose sur des accords professionnels (grade D) considérés comme consensuels par le groupe d'experts développant les RPC.

De nombreuses études ont montré que la seule diffusion des RPC sous la forme de documents avait peu d'impact sur les pratiques médicales (Matillon & Durieux, 2000). En revanche, l'utilisation de systèmes d'aide à la décision médicale (SADM) serait susceptible d'augmenter l'observance des recommandations par les praticiens (Shiffman *et al.*, 1999; Garg *et al.*, 2005; Nies *et al.*, 2006), même si les conditions de succès des SADM ne sont pas élucidées en pratique. ASTI (Séroussi *et al.*, 2001) est un SADM permettant la diffusion des RPC, développé à l'intention des médecins généralistes. Le mode guidé d'ASTI (ASTI-MG), développé selon les principes de l'aide à la décision documentaire (Bouaud & Séroussi, 2005) est un système utilisé « à la demande » sur l'initiative du médecin qui recherche une solution thérapeutique à la prise en charge d'un patient donné. La base de connaissances (BC) du mode guidé a été développée de façon à fournir des propositions thérapeutiques dans *toutes* les situations cliniques. Nous avons utilisé le formalisme de l'arbre de décision pour représenter l'ensemble des profils patient pris en charge, chaque profil étant caractérisé par un ensemble de variables discrètes, les propositions thérapeutiques figurant au niveau des feuilles de l'arbre. Lors de la modélisation des connaissances, afin de garantir la complétude et la cohérence de la BC, le processus de formalisation du texte des RPC (Shiffman *et al.*, 2004) a souvent nécessité de faire des hypothèses d'interprétation (Georg *et al.*, 2003) et de combiner entre elles différentes sources de connaissances. Ainsi, pour de nombreux profils pa-

tient pris en compte par ASTI-MG, il existe une recommandation thérapeutique avec niveau de preuve uniquement pour des sous profils élémentaires correspondant à un petit sous-ensemble de critères. Par exemple, il existe des recommandations avec niveau de preuve pour un patient avec HTA et diabète, ou HTA et insuffisance rénale, mais pas pour un patient avec HTA, diabète et insuffisance rénale. En l'absence de règles de calcul permettant de combiner les niveaux de preuve, le statut des propositions du système n'est pas toujours évident à évaluer. Une proposition thérapeutique n'aura pas la même fiabilité selon qu'elle est *evidence-based* et basée sur des preuves scientifiques objectivées par un niveau de preuve, ou sur une recommandation issue des RPC mais sans preuve scientifique associée et fondée sur le consensus, ou enfin dans le cas où elle n'est pas explicitement mentionnée dans les RPC et résulte au final d'un avis d'expert hors recommandations.

L'objectif de ce travail est de (i) s'appuyer sur le texte des RPC et la modélisation réalisée pour ASTI-MG pour caractériser le statut en terme de niveau de preuve d'une proposition thérapeutique attachée à un profil patient, (ii) d'analyser la distribution « théorique » de ces statuts dans les propositions thérapeutiques issues de la BC d'ASTI-MG, enfin (iii) d'analyser comment cette distribution s'instancie en pratique au sein d'un échantillon de patients réels. Cette étude utilise les RPC françaises sur la *prise en charge des patients atteints d'hypertension artérielle essentielle* (Haute Autorité de Santé, 2005) ainsi que la BC d'ASTI-MG associée.

2 Modélisation des recommandations

2.1 Les RPC pour la prise en charge de l'HTA

Le document propose une stratification du niveau de risque cardiovasculaire en fonction de la pression artérielle, du nombre de facteurs de risque, de l'atteinte des organes cibles, et de l'existence de maladies cardiovasculaires et rénales. La prise en charge se décline en traitements non pharmacologiques avec la description des mesures hygiénodiététiques et en traitements pharmacologiques avec une présentation des 5 classes d'antihypertenseurs à utiliser, (les diurétiques thiazidiques (DT), les bêta-bloquants (BB), les inhibiteurs calciques (ICa), les inhibiteurs de l'enzyme de conversion (IEC), et les antagonistes des récepteurs de l'angiotensine II (ARAII)). La stratégie thérapeutique générale est décrite ainsi que la prise en charge des situations particulières.

2.2 Principes de modélisation

Nous avons procédé à la modélisation des connaissances selon les principes énoncés par Shiffman *et coll.* (2004) pour la formalisation des RPC sur la prise en charge de l'HTA. Nous avons ainsi réalisé une première étape d'« atomisation » au cours de laquelle les concepts élémentaires ont été identifiés et caractérisés en termes de variables de décision pour décrire les profils patient, et de variables d'action pour décrire les traitements. La seconde étape, l'étape de « désabstraction » a permis d'ajuster le niveau de généralité utilisé dans le texte des RPC à celui plus concret nécessaire à l'opérationalisation des recommandations. Ainsi, on réalise la quantification systématique

des variables qualitatives, et on traduit, par exemple, *sujet âgé* par *sujet de plus de 75 ans*. L'étape de « désambiguisation » a permis de contraindre l'interprétation et d'imposer une spécification du contexte de certains énoncés. Par exemple, la proposition des ARAII à la place des IEC en cas d'intolérance est explicite avec niveau de preuve dans le contexte d'une *insuffisance cardiaque par dysfonction systolique* et explicite sans niveau de preuve dans le contexte de l'*insuffisance rénale*. Au niveau de la phase de désambiguisation, nous avons considéré que cette substitution était licite dans tous les contextes.La dernière étape, l'étape de « complétion », est réalisée lors de la construction de l'arbre de décision de sorte que la contrainte d'exhaustivité et d'exclusivité des modalités des variables de décision soit vérifiée à tous les niveaux de profondeur (Bouaud & Séroussi, 2005). Elle permet ainsi d'expliciter des variables implicites, ou, et c'est le plus souvent le cas, des modalités implicites de variables explicites.

La BC est finalement représentée sous la forme d'un arbre de décision à 2 étages : l'étage clinique qui caractérise l'état clinique du patient sous la forme d'un ensemble de critères cliniques, et l'étage thérapeutique qui permet, pour chaque situation clinique, d'explorer la séquence thérapeutique recommandée afin d'identifier le traitement à prescrire en fonction des traitements déjà reçus par le patient, et de sa réponse aux traitements antérieurs en termes d'efficacité et de tolérance. La figure 1 illustre par un exemple un des chemins de l'arbre de décision.

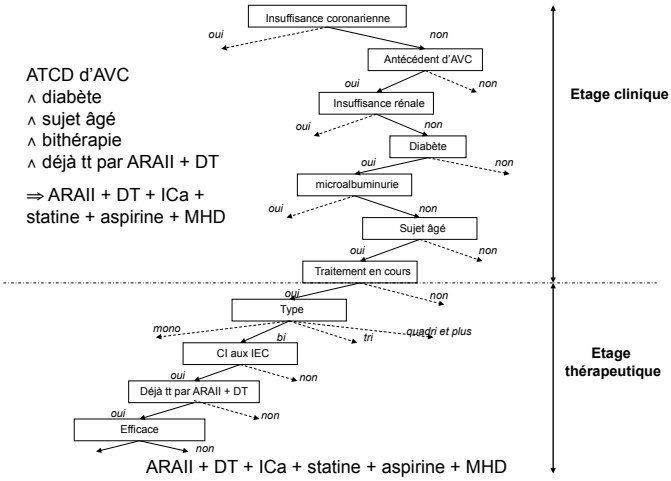


FIG. 1 – Un chemin de l'arbre de décision.

2.3 Construction de l'étage clinique

Les critères cliniques correspondent aux facteurs de risque (*âge*, *tabagisme*, *antécédents familiaux d'accident cardiovasculaire précoce*, *diabète*, *dyslipidémie*), aux critères d'atteinte des organes cibles (*hypertrophie ventriculaire gauche*, *microalbuminurie*), aux

maladies cardiovasculaires et rénales (*insuffisance rénale*, *protéinurie*, *accident vasculaire cérébral*, *insuffisance coronarienne*, *artériopathie des membres inférieurs*), au niveau de la pression artérielle, et à un ensemble de critères permettant de caractériser les situations particulières qui sont considérées une à une en association avec l'HTA : *sujet âgé*, *antécédent d'accident vasculaire cérébral*, *pathologie cardiaque* (*angor stable*, *antécédent d'infarctus*, *insuffisance cardiaque par dysfonction systolique*), *insuffisance rénale*. Ils comprennent également les critères de la prévention secondaire qui apparaissent en fin de document sous la forme d'une figure non référencée dans le texte et qui coïncident avec les maladies cardiovasculaires et rénales.

L'organisation des critères pour construire des profils cliniques « pertinents » est une tache complexe dans la mesure où trois points de vue sont pris en compte dans les RPC pour proposer des stratégies thérapeutiques : l'évaluation du risque cardiovasculaire, les situations particulières, et le niveau de prévention (secondaire *vs* primaire). Ces trois points de vue ne sont pas indépendants mais utilisent au contraire certains critères communs selon le contexte. La stratégie adoptée pour ordonnancer les critères dans l'arbre consiste à déterminer le niveau de prévention, l'atteinte des organes cibles, le diabète, les facteurs de risque et la situation particulière d'un patient donné. Par construction, l'expansion de l'étage clinique représente un catalogue exhaustif de toutes les situations cliniques, médicalement pertinentes, construites par la combinatoire des critères mentionnés dans les RPC. Les profils cliniques de la BC incluent des situations avec HTA seulement (*HTA non compliquée*), avec HTA + une comorbidité, avec HTA + 2 comorbidités, jusqu'au cas extrême de la situation associant l'HTA à toutes les comorbidités. Pourtant, seules l'HTA non compliquée et les situations cliniques associant l'HTA à une seule comorbidité, explicites dans le document des RPC, font l'objet de recommandations avec niveau de preuve. Les autres situations, construites essentiellement au cours de l'étape de complétion, n'existent pas explicitement dans le texte des RPC et ne sont donc pas associées à des « recommandations » au sens strict du terme.

2.4 Construction de l'étage thérapeutique

Construit au-dessous de l'étage clinique, l'étage thérapeutique est spécifique d'un profil clinique. Ainsi, il y a autant d'arbres thérapeutiques que de feuilles de l'étage clinique. Pour chacun, il faut explorer la séquence thérapeutique théoriquement recommandée pour le profil clinique afin de détecter le premier traitement non déjà administré et qui n'est pas contre-indiqué. Soit P un profil clinique, deux situations peuvent donc se présenter : (i) il existe une séquence thérapeutique recommandée S^P pour ce profil clinique dans les RPC, ou (ii) il n'en existe pas. Dans le premier cas, idéalement, S^P doit se présenter sous la forme d'une stratégie de prise en charge : S^P = (T_1^P, $T_1^P + T_2^P$, $T_1^P + T_2^P + T_3^P$), avec pour principe sous-jacent, un traitement initial par monothérapie T_1^P, puis en cas d'inefficacité, le passage à une bithérapie $T_1^P + T_2^P$, et toujours en cas d'inefficacité, le passage à une trithérapie $T_1^P + T_2^P + T_3^P$. Ce principe, recommandé dans les RPC, permet de gérer la stratégie thérapeutique dans sa dimension « efficacité », mais ne permet pas de gérer les problèmes de tolérance. Aussi, il faut pouvoir disposer de traitements de substitution en cas de contre-indication ou d'intolérance, c'est-à-dire, disposer de T_i^P et $sub(T_i^P)$.

Pour un profil clinique donné P dont on connaît la séquence recommandée S^P et les candidats à la substitution en cas d'intolérance, la première étape consiste à détecter si le patient est déjà traité ou non. Si le patient n'est pas déjà traité, il s'agit de proposer un traitement initial et on propose le traitement de première intention de la séquence, soit T_1^P sauf si T_1^P est contre-indiqué, auquel cas il faut proposer $sub(T_1^P)$. Si le patient est déjà traité, il s'agit du suivi d'un traitement en cours et il faut pouvoir se positionner dans la séquence des traitements déjà reçus par le patient. On explore alors le niveau d'association du traitement courant, afin de déterminer si on est en mono, bi, tri, ou quadrithérapie et plus, et dans chaque cas on établit si la « bonne » mono, bi, ou trithérapie a déjà été administrée, et si oui quelle était la réponse au traitement, sinon on la propose sauf si elle est contre-indiquée et on propose alors le traitement de remplacement.

Les séquences thérapeutiques recommandées avec niveau de preuve ne sont disponibles, au niveau du texte des RPC, que dans les situations particulières, et souvent restreintes à la donnée du traitement de première intention (monothérapie initiale). Par ailleurs, des règles d'association explicites existent dans le document mais sans niveau de preuve, sur la base du consensus du groupe de travail avalisé par le groupe de lecture. Ces règles permettent à partir d'une monothérapie initiale de construire la séquence recommandée en déterminant la bithérapie et la trithérapie adaptées. En revanche, pour de nombreuses situations générées par la combinatoire des variables cliniques, il n'existe pas de séquence thérapeutique préconisée dans les RPC, pas même le traitement de première intention qu'il suffirait de compléter. Ainsi, les séquences thérapeutiques proposées par ASTI-MG, par exemple pour le sujet âgé avec *diabète* ET *insuffisance rénale*, ont été établies « hors recommandation » en collaboration avec un expert du domaine.

Sur le plan des conduites à tenir lorsque les traitements théoriquement recommandés pour un profil patient ne conviennent pas à un patient donné du fait d'intolérance ou de contre-indications, seules des solutions alternatives aux IEC sont parfois proposées avec niveau de preuve (substitution par un ARAII). On estime que cette substitution recommandée avec niveau de preuve uniquement dans le cas d'une *insuffisance cardiaque par dysfonction systolique* est vraie dans les autres situations sur la base d'un consensus (étape de désambiguisation). On sait par ailleurs qu'il existe des contre-indications aux BB (asthme) ; elles ne sont pas explorées dans les RPC, et c'est conformément à l'avis de l'expert, que des solutions thérapeutiques de remplacement ont été élaborées et proposées (étape de complétion).

3 Caractérisation du statut des propositions ASTI-MG

Nous avons proposé de caractériser le statut d'une proposition thérapeutique du système vis à vis des RPC à partir de l'analyse de la situation particulière pour laquelle la proposition est faite. Calquée sur la modélisation de la base de connaissance, cette caractérisation s'effectue selon 3 dimensions : la description clinique, le niveau d'association thérapeutique, la gestion des contre-indications.

3.1 Sélection des variables pertinentes

Sur le plan clinique, nous avons retenu 10 variables intervenant dans la caractérisation des situations cliniques particulières pour lesquelles il existe des recommandations avec niveau de preuve : *sujet âgé, antécédent d'accident vasculaire cérébral, pathologie cardiaque, diabète, insuffisance rénale*, ainsi que les critères spécifiant les situations cliniques *pathologie cardiaque* et *diabète* pour lesquelles il existe des recommandations thérapeutiques spécifiques : *angor stable, antécédent d'infarctus, insuffisance cardiaque par dysfonction systolique, microalbuminurie, hypertrophie ventriculaire gauche*.

Sur le plan thérapeutique, étant donné qu'il existe des recommandations avec niveau de preuve pour certaines monothérapies de première intention, des recommandations sans niveau de preuve pour certaines bi et trithérapies, et uniquement des avis d'expert pour les autres niveaux d'association (dont les quadrithérapies et plus), nous avons retenu les variables *traitement anti-hypertenseur en cours* et *niveau d'association* avec les valeurs *monothérapie, bithérapie, trithérapie, quadrithérapie et plus*.

Par ailleurs, comme la prise en compte des contre-indications a un effet sur le niveau de preuve des recommandations, les deux variables booléennes *CI aux BB* et *CI aux IEC* ont également été considérées.

3.2 Typologie des statuts

On considère un profil patient P caractérisé par les variables identifiées précédemment et on s'intéresse à la caractérisation du statut, en terme de preuve, de la proposition thérapeutique qui lui est attachée. On distingue trois statuts : la proposition thérapeutique est une recommandation avec niveau de preuve (RNP) explicitement mentionnée dans les RPC, la proposition thérapeutique est une recommandation sans niveau de preuve mais fondée sur le consensus (RFC), la proposition thérapeutique n'existe pas dans les RPC, c'est un avis hors recommandation (AHR) proposé par un expert.

RNP : Un profil patient P est globalement associé à une recommandation avec niveau de preuve (RNP), lorsque la partie clinique du profil est associée à une recommandation avec niveau de preuve (RNP_{cl}), la partie thérapeutique est associée à une recommandation avec niveau de preuve (RNP_{th}), et dans le cas où il existe des contre-indications à certains médicaments, la solution alternative est recommandée avec niveau de preuve (RNP_{ci}) :

$$RNP(P) = RNP_{cl}(P) \wedge RNP_{th}(P) \wedge RNP_{ci}(P)$$

On considère que $RNP_{cl}(P)$ est vraie lorsque la partie clinique du profil patient coïncide exactement avec une des 9 situations cliniques identifiées. Dans ces cas, il existe en effet des RNP dans le GBP.

On considère que $RNP_{th}(P)$ est vraie lorsque la partie thérapeutique du profil patient correspond à l'instauration d'un traitement anti-hypertenseur, c'est-à-dire, lorsque *traitement anti-hypertenseur en cours = non*.

On considère que $RNP_{ci}(P)$ est vraie s'il n'existe pas de *CI aux BB*, et pas de *CI aux IEC*, sauf dans le cas particulier d'une *insuffisance cardiaque par dysfonction*

systolique où il existe une RNP pour la substitution des IEC par les ARAII dans le GBP.

RFC : Un profil patient P est globalement associé à une recommandation fondée sur le consensus (RFC), s'il n'est pas déjà associé à une recommandation avec niveau de preuve (RNP), et si la partie clinique du profil est associée à une recommandation avec niveau de preuve (RNP_{cl}) ou fondée sur le consensus (RFC_{cl}), la partie thérapeutique est associée à une recommandation avec niveau de preuve (RNP_{th}) ou fondée sur le consensus (RFC_{th}), et s'il existe des contre-indications à certains médicaments, la solution alternative est recommandée avec niveau de preuve (RNP_{ci}) ou fondée sur le consensus (RFC_{ci}) :

$$\begin{aligned} RFC(P) = \quad & \neg RNP(P) \wedge (RFC_{cl}(P) \vee RNP_{cl}(P)) \\ & \wedge (RFC_{th}(P) \vee RNP_{th}(P)) \\ & \wedge (RFC_{ci}(P) \vee RNP_{ci}(P)) \end{aligned}$$

On considère que $RFC_{cl}(P)$ est vraie dans les 4 contextes particuliers suivants : *insuffisance rénale* seulement, *hypertrophie ventriculaire* seulement, *insuffisance rénale* $\wedge$ *diabète*, *ATCD d'infarctus* $\wedge$ *insuffisance cardiaque par dysfonction systolique*.

Comme le GBP recommande des stratégies thérapeutiques ne dépassant pas la trithérapie, $RFC_{th}(P)$ est vraie uniquement si le traitement courant est au plus une bithérapie, c.-à-d. quand *traitement anti-hypertenseur en cours* $=$ *non* ou lorsque *niveau d'association* $=$ *mono ou bithérapie*.

Seule l'intolérance aux IEC est explicitement prise en charge par le GBP, dans certains cas. Aussi, les autres contre-indications ne sont pas explicitement gérées, et $RFC_{ci}(P)$ est fausse lorsqu'il existe une *CI aux BB*.

AHR : Par construction, un profil patient P est associé à un avis d'expert hors recommandation (AHR), lorsqu'il n'est associé ni à une recommandation avec niveau de preuve (RNP), ni à une recommandation fondée sur le consensus (RFC) :

$$AHR(P) = \neg RNP(P) \wedge \neg RFC(P)$$

3.3 Algorithme de classification et expérimentations

Un algorithme a été développé sur la base des spécifications précédentes afin de déterminer le statut RNP, RFC, et AHR des propositions thérapeutiques associées à un ensemble de profils patient. Cet algorithme a été appliqué d'une part à l'ensemble des profils patient générés par l'expansion complète de la BC d'ASTI-MG et d'autre part à un échantillon de profils patient réels extraits d'une base de données d'un cabinet de médecins généralistes. L'objectif est de déterminer la distribution des profils théoriques de la BC et des profils effectivement rencontrés en pratique sur l'échantillon selon les statuts RNP, RFC, et AHR. Les médecins généralistes du cabinet étudié utilisent le dossier patient électronique éO (développé et commmercialisé par la société Silk Informatique à Angers). Nous avons interrogé la base des dossiers du cabinet avec la requête « présence du libellé HTA » dans la rubrique « Antécédents ». Une étude au cas par

cas des dossiers obtenus nous a permis d'éliminer les dossiers pour lesquels l'HTA était présente dans les antécédents familiaux mais pas personnels, et ceux pour lesquels la décision de prise en charge thérapeutique était antérieure à la diffusion des recommandations de prise en charge de l'HTA par la Haute Autorité de Santé (novembre 2005). Les deux distributions ont été analysées et comparées.

4 Résultats

4.1 Distribution théorique du statut

L'expansion de l'arbre de décision dans ses étages clinique et thérapeutique correspond au catalogue nosologique généré par complétion de la BC de tous les profils cliniques médicalement plausibles auxquels est associé l'ensemble des trajectoires thérapeutiques qui doivent être considéreés. La BC est ainsi constituée de 44 571 profils patient théoriques. La distribution des profils pour les 44 571 profils théoriques est représentée dans le tableau 1.

TAB. 1 – Distribution du statut dans la BC d'ASTI-MG.

Statut du profil	*n*	*%*
RNP	206	0,5 %
RFC	5 424	12,6 %
AHR	38 941	87,4 %
Total	44 571	100,0 %

On constate que les profils explicitement couverts par les RPC (avec et sans niveau de preuve) représentent dans la BC moins de 13 % du nombre de situations couvertes par le système. La grande majorité des profils (plus de 87 %) ne sont donc pas explicitement mentionnés dans le GBP et correspondent à des situations hors recommandations.

4.2 Distribution pratique du statut

Sur les 15 527 dossiers utilisés par le cabinet de médecins généralistes au 1er Janvier 2007, 669 présentaient le libellé « HTA » dans la rubrique « Antécédents ». Après élimination des faux positifs, le jeu de données patient est constitué de 435 enregistrements. L'algorithme de détermination du statut a été appliqué sur chaque profil. Le tableau 2 montre que seulement 8,3 % des patients de cet échantillon se trouvent dans une situation pour laquelle il existe des recommandations avec niveau de preuve. Ainsi, dans 91,7 % des situations rencontrées en pratique, la prescription ne peut être strictement basée sur des preuves scientifiques : 50 % sont prises en charge sans niveau de preuve par les RPC, 50 % sont hors recommandations.

TAB. 2 – Distribution du statut dans un échantillon de patients.

Statut du profil	*n*	*%*
RNP	36	8,3 %
RFC	204	46,9 %
AHR	195	44,8 %
Total	435	100,0 %

4.3 Contribution des critères au statut

La détermination du statut final d'un profil patient fait intervenir 3 sortes de critères (cf. section 3) : des critères cliniques, le niveau d'association thérapeutique et l'existence de contre-indications. Nous nous sommes intéressés à la contribution incrémentale de chacune de ces catégories dans la détermination du statut des stratégies thérapeutiques proposées sur le jeu de données. La figure 2 illustre ces contributions. Si on limite la description des profils patient aux critères cliniques, plus des 2/3 des profils se trouvent dans des situations (cliniques) pour lesquelles il existe des recommandations avec niveau de preuve (RNP). En revanche, la prise en compte du niveau d'association médicamenteuse courant, fait chuter cette proportion à moins de 10 %. Cette proportion est ensuite peu modifiée par la prise en compte des contre-indications et l'on retrouve au final que 8,3 % de patients sont en situation d'être pris en charge conformément à des RNP.

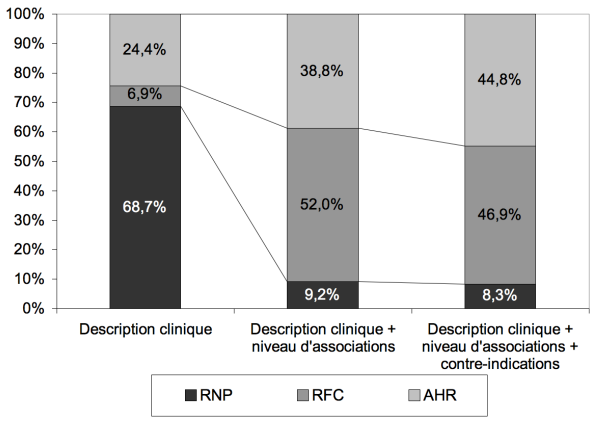


FIG. 2 – Distributions successives du statut en fonction des types de critères pour les 435 patients de l'échantillon (de gauche à droite).

5 Discussion et conclusion

L'analyse précédente a montré que 55,2 % des patients venus consulter se trouvaient dans une situation couverte par le GBP, en incluant les 8,3 % couverts par des recommandations avec niveau de preuve. On constate ainsi que la notion d'« observance des RPC » par les médecins ne fait sens que pour ces 55,2 % des patients. Par ailleurs, pour les 44,8 % de patients restants, la réponse à la question de la meilleure prescription n'est pas donnée par les RPC. Les médecins doivent alors composer eux-mêmes avec les connaissances dont ils disposent, dont les RPC, et leur expérience. Ce phénomène pourrait pour partie expliquer la variabilité observée dans les prescriptions.

La comparaison des distributions théorique au sein de la BC d'ASTI-MG et pratique au sein de l'échantillon de patients montre que 0,5 % de la BC fournit des RNP pour 8,3 % des patients, 12,6 % de la BC fournit des RFC pour 46,9 % des patients, et 87,4 % de la BC fournit des propositions AHR, fondée sur des avis d'experts, pour 44,8 % des patients. On serait en droit de penser que les médecins généralistes suivent moins volontiers les propositions du système qui représentent des avis d'experts que celles qui correspondent à des recommandations. Aussi, on pouvait s'attendre à ce que les généralistes soient plus observants sur des profils RNP que sur les profils RFC et enfin AHR. Nous avons comparé les prescriptions effectivement réalisées pour les patients du jeu de données aux propositions d'ASTI-MG, puis calculé le taux d'observance des pratiques aux propositions du système[1]. Le taux global de conformité des pratiques réelles aux propositions d'ASTI-MG est de 33,5 % (sur 435 patients). En fonction du statut du profil patient, il est de 69,4 % pour les profils RNP (36 patients), de 38,7 % pour les profils RFC (204 patients), et de 21,5 % pour les profils AHR. Il est également à noter que sur le sous-échantillon des patients couverts par les RPC (RNP et RFC), le taux d'observance est globalement de 43,3 %.

Lorsque l'on considère uniquement le statut des sous profils cliniques (figure 2), on constate que plus des deux tiers des patients (68,7 %) sont dans des situations RNP, suggérant ainsi que l'hypertension artérielle dans cet échantillon est peu ou pas compliquée. Cependant, lors que l'on rajoute la contrainte du niveau d'association médicamenteuse, les sous-profils RNP chutent à 9,2 %. La prise en compte des contre-indications modifie peu ce résultat en donnant au final les 8,3 % de profils complet RNP. La perte de ce statut RNP illustre que l'on se trouve rarement dans le cas de primo prescription, seule configuration pour laquelle des recommandations avec niveau de preuve sont disponibles. Parce que l'HTA est une maladie chronique, donc évolutive, les prescriptions doivent être adaptées en cas d'inefficacité ou d'intolérance hors d'un cadre EBM. Toutefois, les RPC fournissent tout de même des solutions RFC pour la gestion du niveau d'association thérapeutique puisqu'au final 46,9 % des profils sont RFC.

Au delà des limites intrinsèques de l'étude (représentativité de l'échantillon, cas particulier de l'HTA et des RPC françaises), ce travail suggère que pour la prise en charge de l'hypertension, l'EBM au sens strict fournit peu d'aide aux médecins. Si les RPC étendent la couverture des situations strictement prises en compte par l'EBM (ajout des RFC aux RNP de l'EBM), il reste environ 45 % de situations où le médecin est livré à

[1] Les prescriptions recueillies dans les dossiers médicaux ont été réalisées en pratique courante, en dehors de toute intervention, en particulier de l'utilisation ASTI-MG.

lui-même, ce qui pourrait partiellement expliquer la variabilité des pratiques observées. Des SADM, tels ASTI-MG, utilisés au moment de la décision pourraient être en mesure de fournir une aide à la prescription, mais dans ces situations hors recommandation, la question de l'acceptabilité des propositions se pose.

Remerciements

Les auteurs remercient leurs partenaires du projet ASTI 2, en particulier M. Christian Simon de la société Silk Informatique pour son aide dans la récupération des données issues du logiciel éO. Le projet ASTI 2 a reçu un financement de la C.N.A.M.T.S.

Références

BOUAUD J. & SÉROUSSI B. (2005). OncoDoc : modélisation d'un guide de bonnes pratiques, mise en œuvre et évaluation d'un système d'aide à la décision médicale. In R. TEULIER, J. CHARLET & P. TCHOUNIKINE, Eds., *Ingénierie des connaissances*, p. 229–250. Paris : L'Harmattan.

GARG A. X., ADHIKARI N. K. J., MCDONALD H., ROSAS-ARELLANO M. P., DEVEREAUX P. J., BEYENNE J., SAM J. & HAYNES R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes : a systematic review. *JAMA*, **293**(10), 1223–1238.

GEORG G., SÉROUSSI B. & BOUAUD J. (2003). Does GEM-encoding clinical practice guidelines improve the quality of knowledge bases ? a study with the rule-based formalism. In M. MUSEN, Ed., *Actes AMIA 2003*, p. 254–258, Washington, DC : AMIA.

HAUTE AUTORITÉ DE SANTÉ, Ed. (2005). *Prise en charge des patients atteints d'hypertension artérielle essentielle - actualisation 2005*. HAS/Service des recommandations professionnelles.

MATILLON Y. & DURIEUX P. (2000). *L'évaluation médicale, du concept à la pratique*. Paris : Flammarion.

NIES J., COLOMBET I., DEGOULET P. & DURIEUX P. (2006). Determinants of success for computerized clinical decision support systems integrated in cpoe systems : a systematic review. In *Actes AMIA 2006*, p. 594–598, Washington, DC : AMIA.

SACKETT D. L., ROSENBERG W. M., GRAY J. A., HAYNES R. B. & RICHARDSON W. S. (1996). Evidence based medicine : what it is and what it isn't. *Br Med J*, **312**(7023), 71–2.

SHIFFMAN R. N., LIAW Y., BRANDT C. A. & CORB G. J. (1999). Computer-based guideline implementation systems : a systematic review of functionality and effectiveness. *JAMIA*, **6**(2), 104–114.

SHIFFMAN R. N., MICHEL G., ESSAIHI A. & THORNQUIST E. (2004). Bridging the guideline implementation gap : a systematic, document-centered approach to guideline implementation. *J Am Med Inform Assoc*, **11**(5), 418–426.

SÉROUSSI B., BOUAUD J., DRÉAU H., FALCOFF H., RIOU C., JOUBERT M., SIMON C., SIMON G. & VENOT A. (2001). ASTI, a guideline-based drug-ordering system for primary care. In V. L. PATEL, R. ROGERS & R. HAUX, Eds., *Medinfo*, p. 528–532.