

# Applications du Web sémantique

Alain Léger<sup>1</sup> et Jean Charlet<sup>2</sup>

<sup>1</sup> France Telecom R&D, 4, rue du Clos Courtel, 35512 Cesson  
alain.leger@rd.francetelecom.com

<sup>2</sup> Mission de recherche STIM, AP-HP & INSERM ERM 0202  
Jean.Charlet@spim.jussieu.fr

## **Résumé**

*Les technologies du Web sémantique sont de plus en plus appliquées à un large spectre d'applications au sein desquelles une connaissance de domaine est modélisée et formalisée (ontologie) afin de servir de support à des traitements très diversifiés (raisonnements) effectués par des machines. En outre, ces représentations peuvent être rendues compréhensibles par l'homme pour assurer un couplage optimal entre raisonnements humains (cognitifs) et mécaniques (sémantique formelle) confiant à l'homme et à la machine des tâches complémentaires. Pour citer quelques-unes de ces applications : Portails d'entreprises et Mémoire d'entreprises, E-Commerce, E-Work, Traitement Automatique des Langues et Traduction Automatique, Recherche d'Information, Intégration d'Entreprises et E-Work, Communautés d'Intérêts, Data Mining, etc.*

*Au carrefour d'une maturité technologique émergente et d'une pression économique pressant des gains potentiels et l'élargissement ou la création de nouveaux marchés, se manifeste un intérêt croissant pour l'évaluation des technologies du Web sémantique sous l'angle des coûts et bénéfices mesurables qu'offre cette nouvelle technologie. Une première étape dans la mesure objective de l'intérêt de cette nouvelle technologie est d'en présenter simplement de premiers résultats pré-industriels pour des applications prototypes les plus prometteuses. C'est bien l'objectif de ce document qui essaie de tracer les résultats les plus significatifs et les plus lisibles à ce jour.*

**Mots-clés :** *Web sémantique, ontologies, services Web, applications, évaluation.*

#### **Abstract**

*Semantic web technology is more and more applied to a large spectrum of applications within which domain knowledge is conceptualized and formalized (Ontology) as a support for very varied processing (Reasoning) operated by machines. Moreover, those representations can be rendered understandable by human beings so that a subtle coupling between human reasoning (cognitive) and mechanical (logic-based) is possible for sharing complementary tasks to human and machine. To name few of those applications areas: Corporate Portals and Knowledge Management, E.Commerce, E.Work, Natural Language Processing, Automated Translation, Information Research, Data and Services Integration, Social network and collaborative filtering, Data Mining, etc.*

*From a societal and economic perspective, this emerging technology should contribute to economic wealth growth, in particular in the key area of agile enterprise integration, but also in the public domain to ease work, leisure, administrative and everyday activities. At the crossroad of a maturing technology and a pressing industry anticipating real benefits for cost reductions and market expansion, needs to objectively evaluate the expectations via benchmarking is more and more expressed. A first concrete step towards this evaluation is to present the most prominent prototypical applications either deployed or simply fielded. To that end this chapter tentatively traces some most significant applications.*

**Key-Words :** *Semantic Web, Ontologies, Applications, Web services, Evaluation.*

## **1. INTRODUCTION**

Via la pénétration profonde des technologies numériques au sein de la société de l'information, le contenu du Web est multiforme, incertain et très dynamique. Cela conduit naturellement à tenter d'abstraire cette complexité apparente, en fournissant des nouveaux services capables de *raisonner sur des représentations conceptuelles (sémantiques) via des automates – ex. Web services*. Cette couche sémantique, fait l'objet d'une très forte activité de R&D mondiale dénommée « Web sémantique »

(DARPA, IST), OntoWeb [16], Semantic Web [22], ERCIM News [7], dont les applications premières sont évidentes et dont les prolongements semblent potentiellement très riches de retombées technologiques et de services pour tous les acteurs de la chaîne de traitement de l'information.

Cette nouvelle technologie est à la croisée de nombreuses disciplines telles les sciences cognitives, l'apprentissage symbolique, le traitement automatique des langues, les systèmes multi-agents, l'ingénierie des connaissances, les sciences du raisonnement et de la décision, qui adoptent une approche formelle, normative et algorithmique des raisonnements de sens commun et leurs traces fortes dans la langue via la rencontre machinerie-computationnelle/facteur-humain.

Nous présentons ici quelques applications phares ayant fait l'objet de travaux importants dans différents domaines applicatifs, ainsi que leurs résultats.

## **2. E-COMMERCE**

### **2.1. Quels usages des ontologies pour le E-Commerce ?**

Le commerce électronique doit permettre un échange plus fluide d'information et de transactions entre tous les acteurs économiques, depuis l'offreur de produits ou services jusqu'aux clients finals. On distingue usuellement deux scénarios : des offreurs aux clients (B2C – *Business-to-Customer*) et entre offreurs et grossistes (B2B – *Business-to-Business*).

Les applications du B2C permettent aux offreurs de produits et services de propager et présenter leurs offres, et aux clients, de trouver et de commander l'offre(s) sélectionnée(s). En fournissant un accès unique à une large collection d'articles ou de services fréquemment mise à jour, une place de commerce électronique facilite la rencontre entre l'offre et la demande grâce à des outils de médiation commerciale.

Les applications du B2B ont une plus longue histoire et utilisent les échanges informatisés via des structures de messages et de protocoles très codifiées, pré-établies et normalisées (EDI – *Electronic Data Interchange* ou Échange de Données Informatisés) récemment assouplies via des standards basés sur XML (*eXtensible Markup Language*).

Une nouvelle génération de services d'échange de messages compatible XML est en développement : ebXML (*electronic business in XML*), UN/CEFACT, OASIS et de nombreux acteurs du commerce électronique ont uni leurs efforts pour concevoir un nouveau standard pour le commerce électronique. Loin de devoir remplacer EDIFACT, ebXML se positionne dans la complémentarité et dans la continuité. EDIFACT est particulièrement adapté aux échanges de gros volumes avec des partenaires stables, alors que ebXML doit répondre, en plus, à la problématique des petits échanges entre partenaires épisodiques.

Actuellement, les systèmes à base d'ontologies apparaissent comme une technologie clé pour le développement de solutions d'E-Commerce efficaces, ouvertes et profitables. Cependant, par manque de normes de modèles de domaine et de processus métier dans les plus larges secteurs économiques, le E-Commerce peine à décoller.

En effet, la variété d'entreprises et de solutions de commerce électronique déployées faisant usage de configurations d'échanges très diversifiées, associée au manque de fiabilité et de sécurité sur Internet, rendent impossible le passage à l'échelle par l'intégration et l'interfonctionnement de ces différentes solutions.

Par ailleurs, dans une situation de marché où coopérations et compétitions interfèrent, l'adoption de standards de domaines et de transactions économiques est très difficile à atteindre. De plus :

- les pratiques commerciales sont très variées et rendent très difficiles les alignements normatifs ;
- les entreprises sont complexes : la description des produits et services (seuls ou associés), et leurs interactions sont difficiles à modéliser ;
- les règles du jeu économique sur des places de marché sont très opportunistes ;
- l'adoption de standards pourrait limiter la créativité commerciale.

Malgré toutes ces difficultés, de réels bénéfices pourraient être tirés de l'usage d'ontologies dans les domaines suivants : catégorisation de produits dans des catalogues, catégorisation de services (dont les Web services), pages Jaunes des sociétés de services, identification des pays, régions et monnaies, identification des organisations, de personnes et d'entités légales, identification de containers de transport (type, situation, routes et contenus) ou classification de données statistiques.

Quelques applications B2B font usage de références codées dans des classifications (ex. UNSPSC, OTA) pour réduire la taille des données à transmettre entre acteurs économiques. De tels codes s'affranchissent des

ambiguïtés inhérentes de la langue naturelle (polysémie sur les noms de produit et polymorphisme sur les noms propres). Enfin, pratiquement aucun des schémas de classification utilisés n'est décrit formellement comme le devrait être une ontologie.

Nous présentons dans la suite deux exemples de tentatives d'usage des technologies du Web sémantique au contexte du *E-Commerce*.

## **2.2. Le E-Commerce à base de connaissances : ONTOSEEK® et MKBEEM®**

Pour des services de pages jaunes ou des catalogues de produits, une représentation structurée des contenus couplée à des ontologies linguistiques<sup>1</sup> améliore de manière notable le rappel et la précision des outils de recherche marchands. Le système ONTOSEEK (1996-1998) a couplé une représentation des connaissances du domaine (langage à pouvoir d'expression très limité de la famille de graphe conceptuel – GC) à une large ontologie linguistique multilingue (SENSUS basé sur WORDNET) pour une recherche de produits en langue naturelle multilingue [12].

ONTOSEEK combine un mécanisme de recherche par le contenu sémantique (ontologie) avec un formalisme de représentation assez pauvre (GC). A la différence des systèmes connus, l'utilisateur n'est pas supposé connaître le vocabulaire de codage des produits mais grâce à l'ontologie linguistique SENSUS peut s'exprimer avec les termes de son vocabulaire.

Les principaux choix d'architecture fonctionnelle d'ONTOSEEK :

- usage d'une *Ontologie linguistique généraliste* pour représenter finement les produits ;
- grande flexibilité terminologique dans l'expression des requêtes, grâce à un mécanisme d'*intersection sémantique* entre les requêtes et la description des produits ;
- assistance interactive pour la formulation de la requête par généralisation et spécialisation.

---

<sup>1</sup> Le terme d'« ontologie linguistique » peut faire penser à un oxymoron dans la mesure où une ontologie à trait aux concepts, pas aux mots de la langue. En pratique, les utilisateurs principaux de ce terme [24] décrivent plutôt l'association d'une ontologie et de lexiques de différentes langues dans le but de créer une représentation pivot. Voir le chapitre 4 sur les ontologies (ce volume) pour de plus amples développements sur les ontologies).

Ils font usage d'un formalisme de représentation basique des GC pour représenter les requêtes et les descriptions des produits. Le mécanisme d'*intersection sémantique* est basé sur un simple calcul de subsomption sur les arcs et les nœuds du graphe et ne met pas en œuvre un calcul complet d'endomorphisme de graphe. ONTOSEEK n'a pas fait l'objet de déploiement commercial mais a très bien montré, à son époque, les gains potentiels que les prémisses de technologie du Web sémantique pouvaient apporter aux moteurs marchands pour le commerce électronique.

MKBEEM<sup>2</sup> (*Multilingual Knowledge-Based eCommerce*), projet IST du 5<sup>e</sup> programme cadre avait pour objectif de développer une plate-forme de *commerce électronique multilingue et multiculturelle* principalement centré vers des services pour le B2C. Les résultats finaux ont clairement indiqué que pour des domaines commerciaux bien délimités – mais totalement réalistes – les technologies de la connaissance couplées à des technologies du traitement automatique des langues (TAL) fournissent des services de traduction et d'interprétation de grande qualité et opérationnels à très court terme.

L'innovation clé réside dans ce couplage du TAL et de représentation des connaissances qui offre à ce jour les services suivants :

- représentation de la requête langue naturelle dans une représentation sémantique (ontologie) ;
- maintenance aisée de catalogues de produits et services multilingues ;
- création aisée d'offres composites de produits et de services ;
- recherche en langue naturelle de produits ou de services par le contenu sémantique ;
- catégorisation et indexation automatiques des produits ou des services décrits en langue naturelle ;
- intégration aisée et rapide de nouvelles offres de produits ou de services dans un contexte multilingue et pluriculturel.

La qualité des résultats a été jugée sur un prototype pan-européen pour le Finnois, le Français, l'Espagnol et l'Anglais dans les domaines du tourisme (SNCF) et de la vente par correspondance (Redoute-Ellos). Cette technologie fait maintenant l'objet de transferts technologiques vers la Réunion des Musées Nationaux (RMN – toujours en cours) et vers des opérateurs de plateformes de services multilingues. Elle a aussi des prolongements scientifiques dans des projets IST du 6<sup>e</sup> programme cadre – *e.g.* AceMedia<sup>3</sup>.

---

<sup>2</sup> <http://www.mkbeem.com>

<sup>3</sup> <http://www.acemedia.org/>

### 3. APPLICATIONS MÉDICALES

La médecine est un des domaines d'applications privilégiés du Web sémantique comme elle l'a été, à une autre époque, des techniques de l'Intelligence artificielle, en particulier les systèmes experts. C'est en effet un domaine complexe où les informations à partager sont nombreuses et où il n'y a pas ou peu de solutions algorithmiques à ce partage comme à l'usage des connaissances, en particulier cliniques. Ainsi, un des principaux mécanismes du Web sémantique qui est la description de ressources via des annotations est de la plus grande importance en bio-informatique, plus particulièrement autour des questions de partage des ressources génomiques. Dans le contexte, plus ancien, de la recherche d'information, la médecine a une longue tradition de développement de thésaurus comme le MeSH (*Medical Subject Heading*) ou UMLS (*Unified Medical Language System* – <http://www.nlm.nih.gov/research/umls/umlsmain.html>) et les utilise maintenant dans le cadre des mécanismes du Web sémantique. Enfin, et plus récemment, les services Web proposent des solutions à la problématique récurrente et non résolue de l'interopérabilité en médecine, en particulier dans le contexte des systèmes d'information hospitaliers (SIH). C'est dans ces 3 champs de l'informatique médicale que nous allons décrire les travaux de recherche qui se développent, les résultats et les perspectives attendues<sup>4</sup>.

#### 3.1. Le partage de ressources

Dans le domaine de la génomique fonctionnelle, il est nécessaire d'accéder à une multitude de bases de données et de connaissances accessibles via le Web, mais hétérogènes dans leur structure et leur terminologie. Parmi ces ressources, citons les bases de données comme Swissprot, où les produits de gènes sont annotés par GENEONTOLOGY, GENBANK, etc. En comparant ces ressources, on s'aperçoit qu'elles proposent de l'information identique – *e.g.* des références à des articles – sous des formats extrêmement différents, bien que XML soit mis en avant comme langage de description.

Dans un autre domaine que la génomique mais en utilisant les mêmes mécanismes du Web sémantique (ontologies, médiateurs), le projet

---

<sup>4</sup> Ces descriptions doivent beaucoup au *workshop* organisé par le Laboratoire d'informatique médicale (LIM) de Rennes en collaboration avec l'AS Web sémantique <<http://videostream.univ-rennes1.fr/~wsm/>>.

NEUROBASE<sup>5</sup> est un projet soutenu par le ministère français de la Recherche (MENRT) qui a pour objectif de fédérer au travers d'Internet des bases d'informations en neuroimagerie, situées dans différents centres d'expérimentation, cliniques neurologiques ou de recherche en neurosciences. Ce projet consiste à spécifier comment relier et accéder à ces bases d'informations par la définition d'une architecture informatique permettant l'accès et le partage de résultats d'expérimentations ou bien encore de méthodes de traitement des données au sein d'un même site ou entre sites différents. Cette architecture repose sur le concept de médiateurs (Cf. chap. 5, ce volume). Cela permettra, par exemple et au sein de ces bases d'informations, la recherche de résultats similaires, la recherche d'images contenant des singularités ou encore des recherches transversales de type « fouille de données » pour mettre en évidence d'éventuelles régularités. Le médiateur de NEUROBASE devrait être expérimenté sur une application clinique d'aide à la décision en chirurgie de l'épilepsie.

### 3.2. L'indexation et le catalogage

Le site PubMed <<http://www.ncbi.nlm.nih.gov/PubMed/>> de la NLM (*National Library of Medicine*) donne accès à la plus grande base d'articles scientifiques dans le domaine de la bioinformatique. Ces articles sont indexés à l'aide des termes du MeSH <<http://www.nlm.nih.gov/mesh/meshhome.html>>, un thésaurus contenant près de 22 000 descripteurs. La maintenance de PubMed met en lumière un des problèmes de l'indexation, le travail que représente le choix d'index pertinent pour représenter les articles. Cela rejoint, comme le projet suivant, la question de la mise en place des annotations (Cf. chap. 3, ce volume), difficile d'autant plus qu'elle est effectuée a posteriori. La NLM a ainsi un gros projet d'indexation automatique des ressources fondée sur l'analyse du titre, du résumé de l'article et des index déjà posés sur les articles cités en référence »[1].

Le site CISMef du CHU de Rouen, reconnu en France comme étant le site de référence en informatique médicale, « catalogue » et indexe l'ensemble des sites médicaux francophones de qualité (environ 13 500 en 2004, 50 ressources de plus chaque semaine – <<http://www.chu-rouen.fr/cismef/>>). En dehors du fait que cela ne se fait pas sans méthode et sans une certaine force de travail, il est intéressant d'explorer d'un peu plus près les modes d'indexation des sites [6] : ainsi, quand une page

---

<sup>5</sup> <http://www.irisa.fr/visages/neurobase>

Web est cataloguée, elle est indexée pour pouvoir être retrouvée et reproposée aux intéressés. RDF, les balises du *Dublin Core* pour décrire des informations de type bibliographique, des éléments du *Learning Object Metadata* (LOM) pour les ressources pédagogiques et des éléments du *HIDDEL metadata* [8] pour décrire la transparence et la qualité de l'information de santé sur Internet, sont alors utilisés au sujet de toutes les pages indexées. Des balises nécessaires aux ressources médicales comme la gratuité de la ressource ou son *niveau de preuve* s'y ajoutent. Comme pour Medline, c'est le thésaurus MeSH qui est utilisé pour indexer le contenu médical des ressources. Parmi l'intense activité de recherche autour de CISMéF, un travail a été fait pour formaliser la terminologie utilisée dans CISMéF (principalement le MeSH) en utilisant OWL [24] : cela permet d'améliorer les mécanismes de raisonnement de CISMéF, principalement la subsomption, mais peut se heurter à la difficulté « d'ontologisation » du MeSH (Cf. chap. 4, § 2.2.3).

Ces applications, en particulier CISMéF, nous interrogent sur l'utilisation des thésaurus pour l'indexation par rapport à la possibilité qu'offriraient les ontologies. Si les thésaurus montrent ici parfois leur limite avec une organisation des concepts médicaux parfois ambiguë ou incohérente, la mise en place d'ontologies a un coût (en temps en particulier) non négligeable et dont la rentabilité n'est pas évidente. De plus, une ontologie manipule des concepts à une telle granularité qu'ils ne sont pas facilement accessibles dans le contexte du travail courant du praticien. Des solutions semblent se mettre en place en reliant les concepts de l'ontologie aux termes des thésaurus dans un *serveur de terminologie* comme en propose le projet GALEN [18] qui rejoint les propositions de *thésaurus sémantiques* dans d'autres domaines [20]. Ainsi, le nouveau thésaurus des soins médicaux, la *Classification commune des actes médicaux* (CCAM) a été développée en utilisant les représentations de GALEN, comme squelette conceptuelle du thésaurus [19].

### 3.3. Des services Web pour l'interopérabilité

Les services Web abordés au § 2.1.1 permettent de proposer des solutions au problème de l'interopérabilité en médecine. C'est ce type d'usage que cherche à promouvoir l'association Edisante <<http://www.edisante.org/>> dans le cadre de son groupe de travail GT11 au sein d'un projet « EDI données cliniques » soutenu par le MENRT. La proposition consiste à utiliser les propositions de E-Commerce et les langages des services Web, en particulier ebXML et SOAP (*Simple Object Access Protocol*) en les augmentant d'éléments spécifiques à la

santé pour proposer une norme d'échange de données cliniques entre praticiens ou institutions [5].

Les propositions du GT11 portent sur une structure permettant de transporter des données et des documents hétérogènes mais avec des informations associées à ce transport, renseignant sur la finalité du message et son contenu, et en permettant la gestion et le traitement – c'est le concept d'enveloppe. Ce concept rejoint totalement le concept récent d'enveloppe ebXML. Il s'en différencie par deux points principaux :

- *Le patient comme unique objet de la transaction.* Un tel échange ne saurait être anonyme du point de vue du couple émetteur-récepteur. Il concerne donc un émetteur et un destinataire, qui tous deux sont impliqués et responsabilisés dans l'échange. Le seul moyen de permettre à un émetteur de signer un envoi réservé à un récepteur précis, concernant un patient qui a le droit d'exiger d'en connaître le contenu, aboutit à une structure nécessairement unique pour le triplet {émetteur, récepteur, patient}.
- *Le caractère multimédia des informations transportées.* Une analyse de l'existant fait apparaître clairement l'existence chez la majorité des acteurs de santé de sources multiples d'informations concernant un même patient. Ces informations médicales ne sont pas nécessairement liées entre elles, notamment sur le plan informatique, et se présentent sous des formes et sur des supports divers (bases de données, documents textuels formatés ou pas, propriétaires ou pas, images, etc.). Si ces informations ne sont pas toujours gérées de manière centralisée chez l'émetteur, il peut être fondamental de les réunir à l'occasion d'un échange avec un autre acteur de santé, qui, lui, saura éventuellement les intégrer dans sa base de données. Même sans lien informatique structuré, le fait de les envoyer ensemble a un sens sur le plan médical, par rapport au contexte précis de l'échange, comme par exemple dans le cas de l'échange d'une image et de son compte rendu. Sur le plan de la traçabilité de l'échange, il est donc fondamental pour l'émetteur et le récepteur de pouvoir prouver que ces informations ont été transmises ensemble.

L'intérêt d'une telle approche est qu'elle trace un chemin vers l'interopérabilité plus facilement que des normes spécifiant précisément les items d'information échangés comme le propose le consortium américain HL7 (*Health Level 7*) ou l'organisme de normalisation européen CEN TC251 [3]. Elle permet une certaine interopérabilité, loin de l'*interopérabilité sémantique* que devraient offrir les ontologies mais plus réaliste dans le contexte de l'informatique médicale à ce jour.

### **3.4. Et dans le futur ?**

Les différents projets et applications reflètent bien un usage majeur du Web attendu par les communautés médicales, le partage ou l'intégration d'informations ou connaissances hétérogènes et proposent d'explorer des méthodes ou architectures différentes pour y répondre : approche médiateur, architecture type système à base de connaissance reposant sur les langages standards RDF et OWL. Les méthodes, langages, outils en cours de développement pour le Web Sémantique doivent prendre en compte ces attentes. Notons enfin que, dans ce cas comme dans d'autres domaines, le Web sémantique est une vision intégratrice et cohérente de problèmes pour lesquels des solutions sont réfléchies depuis longtemps.

## **4. PORTAILS ET MEMOIRES D'ENTREPRISE**

### **4.1. Les services offerts**

Depuis quelques années, la capitalisation des connaissances est vue comme un sujet stratégique pour les entreprises. C'est ainsi que se sont développées tant du point de vue méthodologique que technologique les activités de « Mémoire d'entreprise » ou de « gestion des connaissances de l'entreprise » (KM – *Knowledge Management*). Très clairement le KM est interdisciplinaire et fait appel à la gestion des ressources humaines, à l'organisation et à la culture de l'entreprise, et enfin aux technologies NTIC qui peuvent y jouer un rôle très fort de mutation des usages.

Van Heijst *et al.* [13] définissent la « mémoire d'entreprise » comme la « représentation explicite, persistante, et désincarnée, des connaissances et des informations dans une organisation ». Elle peut inclure par exemple, les connaissances sur les produits, les procédés de production, les clients, les stratégies de vente, les résultats financiers, les plans et buts stratégiques, etc. . La construction d'une mémoire d'entreprise repose sur la volonté de « préserver, afin de les réutiliser plus tard ou le plus rapidement possible, les raisonnements, les comportements, les connaissances, même en leurs contradictions et dans toute leur variété » [17]. Le processus de capitalisation des connaissances permet de réutiliser, de façon pertinente, les connaissances d'un domaine donné, précédemment stockées et modélisées, afin d'accomplir de nouvelles tâches [23]. Le but est de « localiser et rendre visible les connaissances de l'entreprise, être capable de les conserver, y accéder et les actualiser,

savoir comment les diffuser et mieux les utiliser, les mettre en synergie et les valoriser » [11].

Dans un passé récent, les solutions de KM se sont principalement tournées vers les silos de documents textuels produits par l'entreprise comme lieu privilégié de la connaissance. Dans un futur proche, les technologies du Web sémantique, et tout particulièrement les ontologies et les raisonnements sémantiques associés offrent de nouvelles perspectives aux solutions de KM.

Bien que les premières tentatives aient déjà clairement montré tout le potentiel que l'on pouvait en tirer, de nombreux champs d'investigation restent ouverts avant que le Web Sémantique tienne ses promesses, par exemple :

- *Une intégration « sans couture »* des savoirs de l'entreprise est absolument nécessaire, pour éviter toutes les redondances et surcharges superflues ;
- *Une méthodologie, un outillage et une stratégie* de mise en place sont indispensables pour soutenir l'effort de création et de capitalisation des connaissances. Par exemple des outils de maintenance (semi-) automatisée des ontologies pour suivre l'évolution dynamique des savoirs ;
- *L'accès et la présentation* de la connaissance doivent tenir compte du *contexte des tâches courantes* ;
- *La personnalisation* (Cf. Chap. 6, ce volume) doit tenir compte des attentes des utilisateurs pour éviter la surcharge cognitive et pour délivrer l'information au bon niveau de granularité.

Le développement de portails des savoirs servant les besoins de l'entreprise ou de communautés est plus ou moins à ce jour une tâche essentiellement manuelle. Dans un contexte économique très versatile et opportuniste, Ontologies et Outils d'inférence, TAL, devraient faciliter la maintenance évolutive des portails qui doivent être à jour et de plus en plus pertinents.

Les services classiques associés aux solutions de KM pour lesquelles les technologies du web sémantique seront fortement contributives sont :

- accès des employés en situation de mobilité à la mémoire de l'entreprise (Mobile KM) ;
- partage entre employés d'une même communauté (P2P – *Peer-to-Peer computing*) où la construction de la connaissance (Ontologie et annotations) s'opère de manière naturelle et consensuelle ;
- intégration des mémoires d'entreprises décentralisées et multinationales ;

- formation professionnelle continue (*e.Learning*) sur le portail de l'entreprise sur lequel l'employé se voit offrir des parcours de formation diversifiés et surtout personnalisés.

Le KM est évidemment un champ applicatif des technologies du Web sémantique très prometteur. Les technologies documentaires classiques ayant clairement montré leurs limites – très faible capitalisation des savoirs – l'introduction de ces nouvelles technologies laissent entrevoir de réelles avancées de l'offre et des usages.

#### **4.2. Des portails d'entreprise sémantiques : ONTOKNOWLEDGE® et COMMA®**

ONTOBROKER<sup>6</sup> est le premier exemple avancé de mise en œuvre des technologies du web sémantique au KM. L'architecture se compose d'une interface d'interrogation, d'un moteur d'inférence et d'un collecteur (*crawler*) de données sur le Web. Le formalisme d'interrogation est à base de « frame » et définissant la notion d'instances, de classes, d'attributs et de valeurs. ONTOKNOWLEDGE<sup>7</sup> est le projet qui a enrichi les résultats ONTOBROKER.

ONTOBROKER a été mis en œuvre avec succès sur les scénarios d'usage suivants :

- portails communautaires : Acquisition et partage de connaissances en communautés d'employés ;
- annotation de documents (projet (KA)<sup>2</sup> – *Knowledge Annotation Initiative*) ;
- gestion des ressources humaines.

CoMMA [4] est un projet IST subventionné par la commission Européenne visant à développer et tester un environnement de gestion de la mémoire d'entreprise. Le projet s'attache à préserver le contexte de l'existence et de l'utilisation de la mémoire d'entreprise en s'intéressant en particulier à deux scénarios :

- *Aide à l'insertion d'un nouvel employé* : Utiliser la mémoire d'entreprise pour permettre aux nouveaux employés de s'insérer rapidement, de comprendre la politique, le fonctionnement et l'organisation de l'entreprise et les rendre opérationnels le plus rapidement possible en leur permettant de trouver ou en leur suggérant pro-activement l'information dont ils ont besoin.

---

<sup>6</sup> <http://ontobroker.semanticweb.org/>

<sup>7</sup> <http://www.ontoknowledge.org>

- *Support de la veille technologique* : Utiliser la mémoire d'entreprise pour assister l'identification et l'évaluation de technologies émergentes concernant l'activité de l'entreprise, et diffuser l'information pertinente aux personnes concernées et compétentes.

CoMMA se distingue par son approche basée sur l'intégration de plusieurs technologies émergentes (langages du Web sémantique : XML, RDF-S, systèmes multi-agents, apprentissage symbolique, Ingénierie des connaissances). Chacune de ces technologies apporte des éléments de solution pour la réalisation, la gestion et l'exploitation d'une mémoire organisationnelle distribuée et hétérogène.

## 5. TRAITEMENT AUTOMATIQUE DES LANGUES

### 5.1. L'usage d'ontologies "linguistiques" dans les applications

« Ce qui concerne le sens est le point faible des études sur le langage, et le restera jusqu'à ce que nos connaissances aient avancé bien loin de leur état actuel » conjecturait Bloomfield [2].

Que peut apporter le Web sémantique au traitement automatique des langues (TAL) ? Le langage humain est construit de mots individuels (niveau lexical), qui peuvent avoir plusieurs sens, et parfois appartenir à plusieurs catégories lexicales ou parties du discours. Les textes en langue humaine sont des objets très structurés, présentant une cohésion inter et intra-phrased très forte [10].

La sémantique pour le traitement automatique s'intéresse à la modélisation des phénomènes sémantiques intervenant dans le langage humain (anaphore, ellipses, comparatif, références temporelles, attitudes, verbes, ...). Traditionnellement, les approches formelles se sont situées au niveau de la phrase. Elles ont été ensuite étendues au niveau du discours (FraCaS, *a framework for Computational Semantics* [9]).

Quand un auditeur reçoit un message d'un orateur, il essaie de comprendre ce que et pourquoi ce locuteur a produit ce message en faisant appel à ses compétences linguistiques, sa connaissance en général et en particulier celles de la situation d'énonciation, ses croyances, etc. L'auditeur construit donc une représentation (très probablement

sémantique) de ce qu'il comprend de la proposition du locuteur, afin de sélectionner une réaction en retour.

Pour construire cette représentation, il doit partager avec l'orateur quelques croyances et connaissances :

- reconnaissance phonétique et lexicale (si message vocal),
- connaissances lexicales,
- connaissances grammaticales,
- connaissances sémantiques du domaine du discours,
- règles conversationnelles et cohérence discursive,
- connaissances contextuelles.

Selon Zyl *et al.* [28], il y a eu quelques applications faisant usage d'ontologies linguistiques. En complément de l'usage traditionnel de ces ontologies pour la génération (Natural Language Generation, NLG) et la traduction, ces applications les mettaient en œuvre pour l'extraction de sens d'un texte, pour la recherche d'information, et pour l'intégration d'informations hétérogènes.

Une ontologie linguistique telle que définie dans [28] sert de format pivot entre applications ou entre interprétations possibles communes de différentes langues. Les ontologies linguistiques ont généralement pour objet de résoudre les questions suivantes : comment représenter les connaissances d'un univers donné et comment lier cette représentation à celles aujourd'hui classiques des grammaires et des lexiques ?

De nombreuses applications (toutes ?) du web sémantique devraient à l'avenir faire appel aux outils traditionnels du TAL enrichis des représentations et des traitements sémantiques associés.

## **5.2. La traduction Automatique : PANGLOSS® et MIKROKOSMOS®**

Une application de génération de langue naturelle fait traditionnellement appel à une représentation neutre (pivot) à laquelle on relie les différents termes d'une base lexicale multilingue. Ces applications sont des systèmes de traduction à base de connaissances (KBMT – *Knowledge-Based Machine Translation*), traduisant via le sens (sémantique) un texte d'une langue vers d'autres langues. La représentation du sens est modélisée dans une ontologie indépendante des langues qui joue le rôle « d'interlingua ».

Les principaux bénéfices attendus sont : de fournir un fondement pour représenter le sens de texte dans un « interlingua » ; pour permettre à des

lexiques de différentes langues de partager un même modèle. Le modèle ontologique<sup>8</sup> résultant est du coup partagé pour le TAL par l'analyse et la génération.

WORDNET et EUROWORDNET [26] en est un archétype. A la différence de WORDNET dédié à la langue anglaise, EUROWORDNET est une base multilingue (Allemand, Hollandais, Français, Italien, Espagnol, Tchèque et Estonien). Le réseau est organisé de manière identique à WORDNET en « synsets » (ensembles de mots synonymes) lié par des liens basiques de synonymie. Ces ensembles sont ensuite reliés à un interlingua (Inter-Lingual-Index) basé sur le Princeton WORDNET. Au travers de cet index, les langages sont interconnectés de telle sorte qu'il est possible de passer des mots d'une langue aux mots similaires d'une autre langue.

Le système PANGLOSS® [14] traduit des textes Espagnols en Anglais. L'ontologie linguistique utilisée dans Pangloss ® est SENSUS (identique à celle utilisée dans le système ONTOSEEK cité plus haut).

Le système MIKROKOSMOS® [25, 15] traduit des textes Espagnols et Chinois en Anglais. Il inclut un interlingua (TMR – *Text Meaning Representation*) qui produit une représentation sémantique pour les langues sources citées. Il propose aussi un outil d'édition et une API pour accéder à l'ontologie MIKROKOSMOS.

## 6. CONCLUSION

Nous avons résumé dans le paragraphe précédent quelques classes d'applications archétypes de l'usage immédiat et tangible des technologies du Web sémantique. Il ne fait aucun doute que cette technologie du sens doit apporter un saut qualitatif indiscutable si ce n'est une réelle rupture technologique. D'un point de vue économique et sociétal, cette technologie doit pouvoir contribuer à la croissance économique, en permettant aux entreprises d'inter-fonctionner plus aisément et de trouver plus rapidement de nouvelles et meilleures opportunités de marchés, mais également contribuer à la société civile dans sa vie quotidienne au travail et pour ses loisirs. Toutefois, la technologie est encore immature et de nombreuses questions scientifiques restent ouvertes telles que :

- le passage à l'échelle du Web,

---

<sup>8</sup> Qualificatif souvent utilisé mais dans le cas de WORDNET et autres, il est un peu abusif (Cf. Chap. 4, § 2.2.3).

- la tenue en contexte de forte hétérogénéité (modélisations et langages),
- la tenue en milieu fortement évolutif.

Le réel décollage des technologies du Web sémantique ne se fera que quand les technologies auront atteint un niveau de maturité et de conviction suffisantes (ce qui est déjà vrai pour quelques domaines comme le E-Commerce) et quand les modèles économiques feront apparaître de manière évidente les gains en terme de retour sur investissement et d'extension ou d'ouverture vers de nouveaux marchés. L'évaluation des coûts et bénéfices de ces technologies est ainsi un des objets du réseau d'excellence européen KnowledgeWEB<sup>9</sup>.

## 7. REFERENCES

- [1] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindfleisch TC & Wilbur WJ (2000), The NLM Indexing Initiative, *Proc AMIA Symp* p.17-21.
- [2] Bloomfield L., (1933) *Language*, Holt, Rinehart and Winston, New York,
- [3] Charlet J., Cordonnier E. & Gibaud B. (2002) Interopérabilité en médecine : quand le contenu interroge le contenant et l'organisation. *Revue Information, interaction, intelligence* 2(2).
- [4] Gandon F ; et Dieng-Kuntz R. (2004) Ontologie pour un système multi-agents dédié à une mémoire d'entreprise, In : R. Teulier, J. Charlet et P. Tchounikine (éds) *Ingénierie des connaissances*, LHarmattan.
- [5] Cordonnier E., Croci S., Laurent J.-F., Gibaud B. (2003) Interoperability and Medical Communication Using "Patient Envelope"-Based Secure Messaging *Proceedings of the Medical Informatics Europe Congress*,
- [6] Darmoni S.-J., Leroy J.-P., Baudic F., Douyère M., Piot J. & Thirion B. (2000). CISMéF : a structured health resource guide. *Methods of Information in Medicine*, 39(1).
- [7] ERCIM News (2002),Special: Semantic Web, October <[http://www.ercim.org/publication/Ercim\\_News/enw51/](http://www.ercim.org/publication/Ercim_News/enw51/)>.
- [8] Eysenbach G, Yihune G, Lampe K, Cross P and Brickley D. A metadata vocabulary for self- and third-party labeling of health web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL). *Proc AMIA Symp* 2001 pp: 169-73.
- [9] FraCaS (1998) Survey of the state of the art in human language technology, Chapter 3 on *Language Analysis and Understanding*.

---

<sup>9</sup> <http://knowledgeweb.semanticweb.org>

- [10] IJCAI-97, Ontologies and Multilingual NLP, Kavi Mahesh, August 23-29, 1997, Nagoya, Japan.
- [11] Grunstein M. & Barthes J.-P. (1996) An Industrial View of the Process of Capitalizing Knowledge, *Proceedings of the 4th International Symposium on the Management of Industrial and Corporate Knowledge (ISMICK'96)*, p. 265-85.
- [12] Guarino N., Masolo C. & Vetere G., OntoSeek: (1999) Content-Based Access to the Web, *IEEE Intelligent System*.
- [13] Van Heijst G., Van Der Spek R. & Kuizinga E (1996) Organizing Corporate Memories, *Proceedings of the 10th Knowledge Acquisition for Knowledge-based Systems Workshop (KAW'96)*, p. 42/1-17.
- [14] Knight, K.; Chancer, I.; Haines, M.; Hatzivassiloglou. V.; Hovy, E. H.; Iida M.; Luk, S.K.; Whitney, R.A. & Yamada, K. (1995) Filling Knowledge Gaps in a Broad-Coverage MT System. *Proceedings of the 14th IJCAI Conference. Montreal (Canada)*.
- [15] Mahesh, K. & Nirenburg, S. (1995) A Situated Ontology for Practical NLP. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, Aug. 19-20, 1995, Montreal, Canada.
- [16] OntoWeb (2002). Web site of the EC project IST-OntoWeb <<http://www.ontoweb.org>> et SIG4 <<http://sig4.ago.fr>>.
- [17] Pomian J. (1996) Mémoire d'entreprise : techniques et outils de la gestion du savoir, Sapienta.
- [18] Rector *et al.* (1999) Terminology and concept representation languages: where are we? *Artificial Intelligence in Medicine*. Jan;15(1):1-4
- [19] RODRIGUES J.-M., TROMBERT-PAVIOT B., BAUD R., WAGNER J. & MEUSNIER F. (1998). Galen-In-Use : Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. In B. CESNIK, C. SAFRAN & P. DEGOULET, Coordinateurs, *Proceedings of the 9<sup>th</sup> World Congress on Medical Informatics*, Seoul.
- [20] Roussey C., Calabretto S. & Pinon J.-M. (2002). Le thésaurus sémantique : contribution à l'ingénierie des connaissances documentaires. In B. Bachimont, Coordinateur, *Actes des 6<sup>es</sup> Journées Ingénierie des Connaissances*, p. 209-20, Rouen, France.
- [21] Sabah G. (2000) Sens et traitements automatiques des Langues pp 77-108 in *Ingénierie des langues*, Jean-Marie Pierrel, Hermes..
- [22] Semantic Web (2001) <<http://www.ercim.org/EU-NSF/semweb.html>> Research Challenges and Perspectives of the Semantic Web, Sophia Antipolis, France, 3-5 October.
- [23] Simon G. (1996) Knowledge Acquisition and modeling for corporate memory: lessons learnt from experience, *Proceedings of the 10th Knowledge Acquisition for Knowledge-based Systems Workshop (KAW'96)*, p. 41/1-18.

- [24] Soualmia, L.F., Golbreich, C., Darmoni, S.J. Representing the MeSH in OWL: Towards a Semi-Automatic Migration. KR-MED 2004, International Workshop on Formal Biomedical Knowledge Representation, in press.
- [25] Viegas (1999), An Overt Semantics with a Machine-guided Approach for Robust LKBs. *The Proceedings of SIGLEX99 Standardizing Lexical Resources, as part of ACL99*. University of Maryland, USA, Maryland.
- [26] Vossen, P. (ed.) (1998) ; EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht .
- [27] Web sémantique Médical (2003) Premières journées Web sémantique médical organisées par le Laboratoire d'Informatique Médicale de Rennes et l'AS Web sémantique du CNRS. Présentations, résumés et compte rendu accessible en 2003 à <http://wsm2003.org/>
- [28] Zyl J.& Corbett D. (2000), A framework for Comparing the use of a Linguistic Ontology in an Application, *Workshop Applications of Ontologies and Problem-solving Methods, ECAI'2000*, Berlin Germany, August.