Sur quelques aspects du Web sémantique

Philippe Laublet*, Chantal Reynaud**, Jean Charlet***

*Université de Paris-Sorbonne– CNRS (LaLICC)
ISHA 96 bvd Raspail 75006 Paris
Philippe.Laublet@paris4.sorbonne.fr
** Université Paris-Sud – CNRS (L.R.I.) & INRIA (Futurs)
L.R.I., Bâtiment 490, 91405 Orsay cedex, France
cr@lri.fr
Université Paris-X Nanterre
200, avenue de la République, 92001 Nanterre cedex, France
***Mission de recherche en sciences et technologies de l'information médicale - DPA/DSI/AP-HP
jc@biomath.jussieu.fr

Résumé. Le Web sémantique, proposé initialement par le W3C, est d'abord une nouvelle infrastructure devant permettre à des agents logiciels d'aider plus efficacement différents types d'utilisateurs dans leur accès aux ressources sur le Web (sources d'information et services). Différents *langages* de niveau de complexité croissante sont proposés afin de mieux exploiter, combiner et raisonner sur les *contenus* de ces ressources. Les connaissances utilisées, par exemple sous forme de marqueurs sémantiques, doivent s'appuyer sur des *ontologies* afin de pouvoir être partagées et munies d'interprétations opérationnelles. La notion de *méta-données* est au cœur de la démarche avec une grande diversité dans l'interprétation et l'utilisation de cette notion. *L'intégration* automatique d'informations provenant de sources hétérogènes est cruciale particulièrement pour des applications d'entreprise. Enfin la problématique des *services* Web enrichit d'une nouvelle dimension la perspective du Web sémantique. Mais cette perspective peut se heurter à un certain nombre d'obstacles qui devront être surmontés, la recherche de

solutions pouvant amener à différents points de vue sur l'avenir du Web sémantique, mettant plus ou moins l'accent sur l'automatisation ou au contraire sur l'utilisateur.

1 INTRODUCTION

L'expression Web sémantique, attribuée à Tim Berners-Lee [1] au sein du W3C, fait d'abord référence à la vision du Web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Espace virtuel, il devrait voir, à la différence de celui que nous connaissons aujourd'hui, les utilisateurs déchargés d'une bonne partie de leurs tâches de recherche, de construction et de combinaison des résultats, grâce aux capacités accrues des machines à accéder aux *contenus* des ressources et à effectuer des *raisonnements* sur ceux-ci.

Le Web sémantique, concrètement, est d'abord une *infrastructure* pour permettre l'utilisation de connaissances *formalisées* en plus du contenu informel actuel du Web, même si aucun consensus n'existe sur jusqu'où cette formalisation doit aller. Cette infrastructure doit permettre d'abord de localiser, d'identifier et de transformer des ressources de manière robuste et saine tout en renforçant l'esprit d'ouverture du Web avec sa diversité d'utilisateurs. Elle doit s'appuyer sur un certain niveau de consensus portant, par exemple, sur les langages de représentation ou sur les ontologies utilisés. Elle doit contribuer à assurer, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ontologies. Elle doit faciliter la mise en œuvre de calculs et de raisonnements complexes tout en offrant des garanties supérieures sur leur validité. Elle doit offrir des mécanismes de protection (droits d'accès, d'utilisation et de reproduction), ainsi que des mécanismes permettant de qualifier les connaissances afin d'augmenter le niveau de confiance des utilisateurs.

Mais restreindre le Web sémantique à cette infrastructure serait trop limitatif. Ce sont les applications développées sur celle-ci qui font et feront vivre cette vision et qui seront, d'une certaine manière, la preuve du concept. Bien sûr, de manière duale, le développement des outils, intégrant les standards du Web sémantique, doit permettre de réaliser plus facilement et à moindre coût des applications ou des services développés aujourd'hui de manière souvent ad-hoc.

Les recherches actuellement réalisées s'appuient sur un existant riche venant, par exemple, des recherches en représentation ou en ingéniérie des connaissances. Mais leur utilisation et leur acceptation à l'échelle du (ou d'une partie du) Web posent de nouveaux problèmes et défis : changement d'échelle dû au contexte de déploiement, le Web et ses dérivés (intranet, extranet), nécessité d'un niveau élevé d'interopérabilité, ouverture, standardisation, diversités des usages, distribution bien sûr et aussi impossibilité d'assurer une cohérence globale. *Comme l'écrit, en substance, Tim Berners-Lee, le Web sémantique est ce que nous obtiendrons si nous réalisons le même processus de globalisation sur la représentation des connaissances que celui que le Web fit initialement sur l'hypertexte.*

Les propositions faites autour de l'infrastructure du Web sémantique doivent permettre aussi bien la réalisation d'outils généralistes avec des utilisateurs mal définis (un exemple pourrait être des moteurs de recherche prenant plus en compte le contenu sémantique de documents) que la réalisation d'applications pour des tâches plus complexes comme la gestion de connaissances au service des membres d'une entreprise[1]. On soulignera, dans le premier cas, surtout l'utilisation de méta-données (section 4) et dans le deuxième, la nécessité de systèmes d'intégration de données hétérogènes (section 5) ou bien encore d'utilisation et de combinaison de services Web (section 6). Les langages proposés pour le Web sémantique sont au cœur de la démarche, même si l'infrastructure ne se réduit pas à ceux-ci. Nous les présentons en section 2. La section 3 discute, elle, du rôle important, pour la réalisation du Web sémantique, des ontologies et des langages permettant de les représenter. Celles-ci peuvent être utilisées et même s'avérer parfois être indispensables pour l'ensemble des recherches évoquées dans ce papier[2].

Il est clair que la diversité des recherches relevant aujourd'hui du Web sémantique rend illusoire toute volonté d'exhaustivité pour cette brève présentation dont l'objectif est simplement de souligner un certain nombre de points importants pour la réalisation de la vision du Web

[1] Cette opposition a surtout valeur argumentative. Tous les intermédiaires peuvent se présenter. Si l'on prend l'exemple du *e-learning*, on peut avoir aussi bien des outils qui permettent de trouver des offres de cours que des applications dédiées utilisant, par exemple, des méta-données pour personnaliser les parcours.

[2] Le département STIC du CNRS soutient une action spécifique sur le thème du Web sémantique depuis Novembre 2001 (http://www.lalic.paris4.sorbonne.fr/stic/). Nous remercions les participants pour leurs différents apports à cette action qui ont influencé certains aspects de ce papier même si les points de vue (et les erreurs) présentés ici sont de la seule responsabilité des auteurs.

sémantique. Pour une information plus complète on pourra consulter les premiers livres ou actes de conférence parus [3], [8], [7], [5].

2 LES LANGAGES POUR LE WEB SEMANTIQUE

2.1 Les langages du W3C

Les travaux visant la réalisation du Web sémantique se situent à des niveaux de complexité très différents. Les plus simples utilisent des jeux plus ou moins réduits de méta-données dans un contexte de recherche d'information ou pour adapter la présentation des informations aux utilisateurs. Dans ce cas, des langages de représentation simples sont suffisants. Dans les travaux plus complexes mettant en œuvre des architectures sophistiquées, pour permettre par exemple l'exploitation de ressources hétérogènes, des langages plus expressifs et plus formels issus des travaux en représentation et en ingéniérie des connaissances, sont nécessaires.

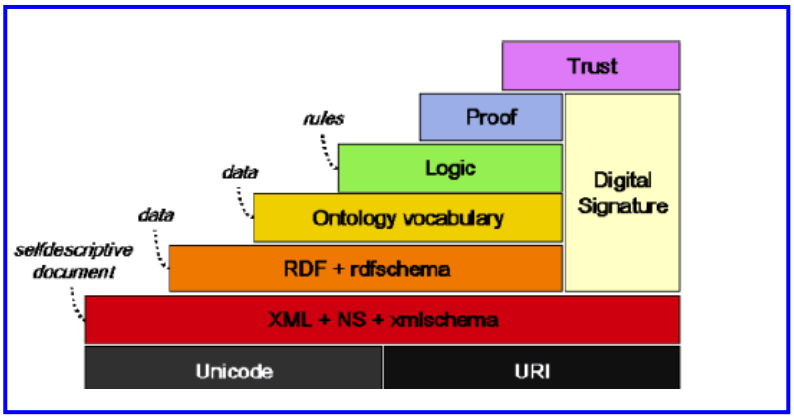

FIG. 1 – Les couches du Web sémantique

La proposition du W3C s'appuie au départ sur une pyramide de langages dont seulement les couches basses sont aujourd'hui relativement stabilisées. La figure 1 montre une des versions [8] de l'organisation en couches proposée par le W3C. Deux types de bénéfices peuvent être

attendus de cette organisation. (1) Elle permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs. (2) Par ailleurs, si elle est bien conçue, elle doit permettre de disposer du langage au bon niveau de complexité, celle-ci étant fonction de l'application à réaliser.

Un aspect central de l'infrastructure est sa capacité d'identification et de localisation des diverses ressources. Elle repose sur la notion d'URI (Uniform Resource Identifier) qui permet d'attribuer un identifiant unique à un ensemble de ressources, sur le Web bien sûr mais aussi dans d'autres domaines (documents, téléphones portables, personnes, etc.). Cette notion connaît aujourd'hui de nombreuses extensions, en cours de standardisation, à d'autres entités que les URLs. Elle est à la base même des langages du W3C.

Une autre caractéristique de tous ces langages est d'être systématiquement exprimables et échangeables dans une syntaxe XML. Ceci permet de bénéficier de l'ensemble des technologies développées autour d'XML : XML Schemas, outils d'exploitation des ressources XML (bibliothèques JAVA, etc.), bases de données gérant des fichiers XML, même si des langages de requêtes spécifiques [6] sont nécessaires pour les langages construits sur XML comme RDF.

2.2 RDF et RDFS

Le premier de ces langages est RDF (" Resource Description Framework) auquel s'est ajouté rapidement RDF Schema (RDFS). Les objectifs initiaux de RDF étaient la représentation et une meilleure exploitation des méta-données. Mais, de manière plus générale, RDF permet de voir le Web comme un ensemble de ressources reliées par les liens étiquetés " sémantiquement ". RDF a permis aussi d'exprimer de larges vocabulaires, comme le catalogue de produits UNSPSC, surtout quand il est complété avec RDFS qui permet d'offrir un niveau supérieur de structuration.

Les énoncés RDF sont des triplets ressource-attribut-valeur (la valeur est une ressource ou chaîne de caractères). Une ressource doit disposer d'une URI. Les triplets sont interprétables comme sujet-prédicat-objet. On notera que le modèle de données n'est pas celui de la structure d'arbres d'XML même si une syntaxe XML existe. On est plutôt proche des premiers réseaux sémantiques. La simplicité du modèle, critiquable pour certains, peut être une des clés de son acceptation et de la relative simplicité de la réalisation d'outils. Certains ajouts comme les containers

et la possibilité de considérer un énoncé (triplet) RDF comme un nœud du graphe lui-même, peuvent augmenter l'expressivité du langage, particulièrement dans un contexte discursif ou de méta-données (qui a affirmé tel énoncé, …) même si [12] constate le peu d'utilisation de ces ajouts dans des applications réelles.

RDFS ajoute à RDF la possibilité de définir des hiérarchies de classes et de propriétés dont l'applicabilité et le domaine de valeurs peuvent être contraintes à l'aide des attributs rdfs:domain et rdfs:range. A chaque domaine applicatif peut être ainsi associé un schéma identifié par un préfixe particulier et correspondant à une URI[3]. Les ressources instances sont ensuite décrites en utilisant le vocabulaire donné par les classes définies dans ce schéma. Les applications peuvent alors leur donner une interprétation opérationnelle. On peut noter que RDFS n'intègre pas en tant que tel de capacités de raisonnement. Par contre, apparaissent des solutions de base de données dédiées à RDF(S), comme l'architecture Sesame [2] à laquelle est associé le langage de requête RQL.

Pour résumer, XML peut être vu comme la couche de transport syntaxique, RDF comme un langage relationnel de base. RDFS offre des primitives de représentation de structures ou primitives ontologiques.

2.3 Les Topic Maps

Une proposition concurrente à RDF(S) pour représenter les méta-données, par exemple pour les ressources Web, est celle des Topic (Navigation) Maps dont un des buts originaux était de gérer et de fusionner des index de livres. Une première standardisation, fondée sur une DTD SGML, en a été donnée par l'International Organization for Standardization (ISO). L'approche des Topic Maps repose sur les notions de *topics* qui peuvent être n'importe quel sujet ou entité, *d'associations* qui étiquettent des relations entre topics et *d'occurrences* qui sont des ressources, disposant d'une URI, qui peuvent être liées à des topics. Les associations peuvent être elles-mêmes des topics. Elles peuvent avec de bons outils de visualisation offrir, par exemple, des possibilités de navigation dans un ensemble de ressources. Une syntaxe XML existe depuis 2001 sous le nom XML Topic Maps (XTM). De même que pour RDF(S), des langages de requête existent, par exemple TMQL.

---

[3] On trouvera de nombreuses références à des schémas RDF incluant jusqu'à plusieurs milliers de classes dans [12].

Comme souvent signalée, une des particularités des Topic Maps de l'ISO, par rapport à RDF(S) du W3C, est la notion *de " scope "* qui permet de définir des contextes différents dans lesquels les éléments nommés identiquement peuvent avoir des significations différentes. Par contre, la notion de hiérarchies de classes n'existe pas dans les Topic Maps. Quoiqu'il en soit, les Topic Maps, de même que RDF(S) permettent aussi bien l'expression de méta-données que l'exploitation des relations entre éléments pour différentes tâches, par exemple l'aide à la navigation sur le Web sémantique.

2.4 Autres langages du W3C

La couche suivante pour le W3C propose de standardiser un langage, dont les spécifications sont bien avancées, pour la représentation et l'utilisation d'ontologies permettant certains mécanismes inférentiels (voir section 3 sur les ontologies).

Dans une vision plus prospective, la pyramide du W3C inclut aussi des couches supérieures. La possibilité de faire des déductions plus complètes pourra s'appuyer sur la standardisation d'un langage de règles, comme RuleML (Rule Markup Language) encore en cours de maturation. De manière plus ambitieuse encore, la capacité de produire des preuves des déductions faites pourra augmenter le niveau de confiance des utilisateurs dans ces déductions. Notons que ce problème de confiance ne se réduit évidemment pas à cette capacité. Elle doit reposer aussi sur des méthodes de qualification de l'origine de l'information, par exemple par des méta-données et des annotations, éventuellement certifiées par des signatures électroniques.

3 ONTOLOGIES

Sans revenir sur les différentes définitions données des ontologies en ingéniérie des connaissances, il est clair que les recherches sur celles-ci sont essentielles pour la réalisation du Web sémantique. En effet, d'une part, une fois construite et acceptée par une communauté particulière, une ontologie doit traduire un certain *consensus explicite* et un certain niveau de *partage* qui sont essentiels pour permettre l'exploitation des ressources du Web par différentes applications ou agents logiciels. D'autre part, la *formalisation*, autre facette des ontologies, est nécessaire pour que ces outils puissent être munis de capacités de raisonnement permettant de

décharger les différents utilisateurs d'une partie de leur tâche d'exploitation et de combinaison des ressources du Web.

Pour représenter les ontologies[4], le W3C cherche à proposer un standard, connu actuellement sous le nom d'OWL (Ontology Web Language). Il s'appuie sur le langage DAML+OIL, produit de la combinaison de l'américain DAML (Darpa Agent Markup Language) et OIL (Ontology Inference Layer) provenant de projets européens. Le langage OWL est actuellement construit sur RDFS, et apporte ainsi aux langages du Web sémantique, l'équivalent d'une logique de description tout en disposant aussi d'une syntaxe XML. Sans être exhaustif, il ajoute à RDF la possibilité de définir des classes de manière plus complexe correspondant aux connecteurs de la logique de description équivalente (intersection, union, restrictions diverses, etc.), les classes disjointes, les propriétés inverses ou transitives ou bien encore les restrictions de cardinalité sur les propriétés. Ces descriptions peuvent être utilisées ensuite par un raisonneur comme FaCT (Fast Classification of Terminology), pour inférer par exemple la subsumption de concepts. En se fondant sur une logique de description, un tel langage a une sémantique formelle claire, ce qui permet de le doter de services inférentiels.

Une des idées force du Web sémantique est ainsi de munir les langages du Web d'une sémantique formelle à l'aide d'une interprétation en terme d'un modèle. Elle permet une caractérisation précise des opérations applicables et par exemple de pouvoir affirmer la correction des algorithmes comme des algorithmes de recherche. Des discussions sont encore en cours, mi 2002, sur comment doivent être (re)définis RDFS et OWL afin de disposer chacun d'une sémantique formelle cohérente [14], sémantiques qui devront être rendues compatibles. Pour certains, comme Peter Patel-Scheider [15], OWL devrait être capable de décrire et d'organiser les nombreuses connaissances exprimées en XML, tout en étant évidemment plus expressif. A cette fin, la sémantique d'OWL devrait tenir compte des modèles de données sous-jacents à XML plutôt que de RDF. De manière plus générale, le débat reste largement ouvert sur les niveaux de complexité nécessaires pour l'OWL [17], complexité algorithmique des mécanismes d'inférence, complexité technique du point de vue des constructeurs d'outils et complexité conceptuelle pour l'utilisateur moyen (pour l'apprentissage du langage et pour son intégration aux pratiques de cet utilisateur).

[4] On trouvera dans [17] un ensemble de points de vue intéressants sur le rôle des ontologies pour le Web sémantique ainsi que sur les problèmes de leur représentation et de leur standardisation.

Des outils pour développer et maintenir des ontologies existent comme Protégé 2000 et OilEd. Une synthèse est faite dans [6]. Leur interface d'édition permet généralement de modéliser sous une forme proche des frames. Un générateur de différents langages comme DAML + OIL est souvent associé ou intégré à ces outils.

Du point de vue des ontologies, seront cruciales pour le Web sémantique les méthodes et des outils contribuant à :

- construire les ontologies, que ce soit à partir de sources primaires, particulièrement les corpus textuels, ou en recherchant une certaine réutilisabilité. La construction d'ontologies à partir de l'analyse de corpus textuels est un domaine en forte évolution où un certain nombre de méthodologies et d'outils sont testés par une communauté très active[5]. La question de la réutilisabilité qui a suscité de longs débats dans la communauté Ingénierie des connaissances a permis de progresser vers la recherche d'une certaine généricité mais reste un enjeu majeur pour le Web sémantique ;
- gérer l'accès aux ontologies, leur évolution, avec gestion des versions, et leur fusion. Les ontologies sont souvent riches de plusieurs milliers de concepts et ne restent alors directement appréhendables que par leur concepteur. Leur accès par des utilisateurs, mêmes professionnels, nécessite de gérer le lien entre les concepts des ontologies et les termes du langage naturel, que ce soit pour une simple compréhension ou pour l'indexation et la construction de requêtes destinées à des tâches de recherche d'information. Les solutions mises en œuvre à ce jour passent par des méthodologies séparant explicitement les termes et les concepts d'un domaine et des outils de visualisation et de navigation recherchant des proximités conceptuelles dans les termes d'un domaine et permettant d'appréhender intuitivement la complexité de ce domaine ;
- assurer l'interopérabilité des ontologies en gérant les hétérogénéités de représentation et les hétérogénéités sémantiques. Ces dernières sont les plus dures à gérer et elles nécessiteront des réflexions conjointes à la problématique de l'accessibilité des ontologies.

Notons que tous ces problèmes sont plus ou moins difficiles suivant le champ d'application et l'étendue recherchée de l'ontologie. Ils semblent plus résolubles pour des domaines où les professionnels (les êtres humains) travaillent déjà de façon coopérative et partagent des

[5] http://www.biomath.jussieu.fr/TIA/. Voir aussi l'article de A. Condamines *et al.* dans ce volume.

conceptualisations communes dans des activités fortement interconnectées. Un exemple pourrait être le domaine du tourisme. Ces problèmes sont d'expérience largement plus complexes dans des domaines comme la médecine ou le droit, où le partage de conceptualisations s'avère alors être un travail de longue haleine passant souvent par des ontologies réalisées de novo et partagées de manière assez locale.

Quoiqu'il en soit, des ontologies de taille importante sont aujourd'hui proposées pour de nombreux domaines (culturels, scientifiques, commerciaux, etc..). L'objectif de ce papier n'est pas d'en faire une recension. On pourra consulter [6]. L'accès à ces ontologies et leur rôle pour le Web sémantique reste encore largement à explorer, même si on peut partager les perspectives (traduites par nous) de Patel-Schneider : *" les utilisateurs d'un serveur Web d'ontologies auront accès à de nombreux services puissants pour stocker et analyser la connaissance disponible sur le Web et faire des inférences à partir de cette connaissance, fournissant aux programmeurs beaucoup plus que les données contenues explicitement dans les documents XML "* [17].

4 META-DONNEES

L'utilisation des infrastructures du Web sémantique par différentes applications sera progressive. On peut parier que les plus nombreuses, à court terme en tout cas, s'appuieront essentiellement sur l'exploitation de méta-données, un des principes de base du Web sémantique étant de décrire les ressources du Web à l'aide de marqueurs exploitables par différents logiciels.

Comme exemples de premières réalisations utilisant pleinement RDF pour des méta-données ou des annotations, on peut citer le système *Annotea* et le logiciel *RDFPic*. Annotea est un système client-serveur collaboratif pour l'annotation de documents sans modification de ceux-ci. Ces annotations en RDF peuvent être ajoutées, modifiées et consultées par une communauté d'utilisateurs qui ont accès à un même serveur d'annotation. XPointer et XLink sont utilisées pour associer les méta-données avec différentes parties des documents. [11] donne des exemples de travail collaboratif facilité par ce type d'outil. RDFPic, lui, est utilisé pour attacher des méta-données à des photos numériques dans le but de faciliter la recherche d'images. Il s'appuie sur une combinaison d'éléments descriptifs propres avec ceux du Dublin Core qui propose un ensemble de champs descriptifs relativement limités (auteur,…). Les

descriptions RDF sont imbriquées dans les fichiers JPEG. D'autre part, différentes initiatives ont déjà produit des spécifications de méta-données en RDF, par exemple LOM (Learning Object Metadata) pour l'apprentissage à distance.

Ces deux exemples, Annotea et RDFPic, sont intéressants parce qu'ils permettent de souligner des distinctions importantes dans les rapports entre méta-données et Web sémantique. Une première est entre une représentation imbriquée dans les ressources qu'elles qualifient (les balises META dans des pages HTML en sont la forme la plus primitive avant même le Web sémantique) et une représentation externe, par exemple un catalogue ou un fichier Topic Maps associé à un portail sur Internet ou Intranet. L'utilisation de RDF est compatible avec les deux alors que l'approche Topic Maps fonctionne à l'aide de fichiers externes. L'approche externe présente l'avantage de laisser tels quels les documents et permet de plus d'exploiter les mêmes documents par des communautés ou pour des tâches différentes, comme souvent signalé dans le contexte des systèmes hypermedia ouverts.

Une autre différence, moins souvent soulignée [13] est entre des méta-données considérées comme objectives, même si elles peuvent être erronées, (comme dans le Dublin Core, l'auteur, la date de création de la ressource, etc…) et des annotations qui représentent des points de vue qui peuvent subjectivement dépendre de l'auteur. Dans le premier cas, l'utilisation d'XML peut être suffisante. Dans le deuxième cas, la flexibilité des outils proposés par le Web sémantique semble être plus essentielle. D'une part, RDF permet de combiner beaucoup plus naturellement des descriptions venant de fichiers différents (simples ajouts de triplets) que dans le cas des DTD ou des schémas XML. De plus la réification en RDF, même si elle est souvent critiquée d'un point de vue théorique, peut s'avèrer dans ce cas particulièrement utile. Elle permet par exemple de représenter qui est l'auteur d'une annotation (notion de méta-méta-données) ou d'utiliser des mécanismes d'authentification ou de " confiance ". On verra alors les annotations comme la base d'un processus de construction de consensus où chacun exprime son point de vue [13]. RDF offre aussi une architecture souple qui permet d'ajouter des descriptions sous forme de triplets dans une vision dynamique et évolutive des méta-données. La vision élargie [13] des méta-données, que peut proposer le Web sémantique, les voit ainsi selon nous comme *subjectives*, apportant différents points de vue sur la même ressource, *évolutives,* permettant d'évoluer par des processus de consensus, *distribuées*, par exemple avec des architectures peer-to-peer.

En terme d'applications, il est clair que la recherche d'information (et donc l'indexation) est la première utilité des méta-données. Mais leur rôle ne se limite pas à cela. On peut citer aussi, l'assistance à la navigation, le travail collaboratif à base d'annotations, l'aide à la certification et de manière aussi importante la personnalisation et l'adaptation à des utilisateurs particuliers, par exemple pour la construction de parcours particuliers.

5 L'INTEGRATION DE SOURCES D'INFORMATIONS

5.1 Distribution et hétérogénéité des sources

La diversité des sources d'information distribuées et leur hétérogénéité est une des principales difficultés rencontrées par les utilisateurs du Web aujourd'hui. Cette hétérogénéité peut provenir du format ou de la structure des sources (sources structurées : bases de données relationnelles, sources semi-structurées : documents XML, ou non structurées : textes), du mode d'accès et de requête ou de l'hétérogénéité sémantique : entre les schémas conceptuels ou ontologies implicites ou explicites sous-jacentes. Il est en effet illusoire de penser qu'une même ontologie " universelle " sera largement utilisée sans oublier que les termes sont parfois également exprimés dans des langues différentes.

La prise en compte de ces problèmes est une des clés de la mise en place d'applications sur l'infrastructure proposée par le Web sémantique. Elle s'avèrera encore plus fondamentale si l'on adhère à la vision, à plus long terme, d'agents logiciels capables de raisonner en accédant à des ressources variées. *On peut affirmer que le Web sémantique doit d'abord être une infrastructure dans laquelle l'intégration des informations d'une variété de sources peut être réalisée et facilitée.* Le Web sémantique devrait donc tirer largement bénéfice des recherches déjà effectuées pour la réalisation de *systèmes de médiation* et des résultats déjà obtenus, ces travaux devant contribuer à son acceptation en montrant son utilité. Inversement on peut penser que la réalisation de tels systèmes sera facilitée par l'infrastructure proposée.

L'aide apportée par les systèmes de médiation peut recouvrir différentes formes : découvrir les sources pertinentes étant donnée une requête posée, puis aider à accéder à ces sources pertinentes, évitant à l'utilisateur d'interroger lui-même chacune d'elles selon leurs propres modalités et leur propre vocabulaire, enfin combiner automatiquement les

réponses partielles obtenues de plusieurs sources de façon à délivrer une réponse globale. De tels systèmes de médiation offrent à l'utilisateur une vue uniforme et centralisée des données distribuées, cette vue pouvant aussi correspondre à une vision plus abstraite, condensée, qualitative des données et donc, plus signifiante pour l'utilisateur. Ces systèmes de médiation sont, par ailleurs, très utiles, en présence de données hétérogènes, car ils donnent l'impression d'utiliser un système homogène. Parmi les différentes grandes catégories d'applications de ces systèmes de médiation, on peut citer les applications de recherche d'information, celles d'aide à la décision en ligne (avec entre autres l'utilisation d'entrepôts de données) et celles, de manière plus générale, de gestion de connaissances au sens large.

A titre d'illustration très simple du premier type d'applications, supposons qu'un utilisateur pose la requête suivante : quels sont les films de Woody Allen à l'affiche à Paris ce soir ? où ? leurs critiques ? Supposons l'existence de deux sources d'information. La première, Internet Movie Data Base, utilise un SGBD relationnel et contient une liste de films, précisant pour chacun le titre, les acteurs et le directeur. La seconde Pariscope, qui peut utiliser des fichiers XML contient, par film, les salles où le film peut être vu et, pour chaque salle, le nom de la salle et l'adresse. La réponse à la requête devra être construite en interrogeant chacune d'elle et en combinant les résultats de l'interrogation de façon à offrir à l'utilisateur une réponse globale.

Plus récemment de nouvelles applications ont vu le jour au service des entreprises : eCRM, Business Intelligence, eERP , eKM, etc. Ces applications, que l'on désigne parfois sous le vocable de WebHouse [9] si elles sont menées dans le contexte du Web, s'appuyent sur la construction d'entrepôts de données à partir du Web. Elles se trouvent également confrontées au problème de la médiation puisqu'elles mettent en œuvre un processus d'acquisition de données provenant de sources multiples, distribuées et hétérogènes. Un des objectifs est de permettre une acquisition ciblée des données, éventuellement de " nettoyer " ces données, de les structurer et de les organiser de façon à être capable de retrouver ensuite facilement une information préalablement stockée ou déduite par des outils d'extraction de connaissances. La conception d'outils de médiation intelligents entre les utilisateurs et les sources d'informations, accessibles via le Web ou stockées localement, est nécessaire. Ils aident l'utilisateur à spécifier facilement les données qu'il recherche, celui-ci ayant l'impression d'utiliser un système unique et homogène.

5.2 Différentes approches de la médiation

Les solutions à l'intégration d'information proposées dans le cadre du Web sémantique doivent s'appuyer sur les travaux déjà existants. L'approche médiateur [18] consiste à définir, de manière généralement *centralisée*, un schéma global, ou ontologie qui regroupe l'ensemble des prédicats modélisant le domaine d'application du système. L'utilisateur pose ses requêtes dans les termes du vocabulaire structuré fourni par l'ontologie.

L'ontologie établit également la connexion entre les différentes sources accessibles. En effet, dans cette approche, l'intégration d'information est fondée sur l'exploitation de vues abstraites décrivant de façon homogène et uniforme le contenu des sources d'information dans les termes de l'ontologie. Les sources d'information pertinentes, pour répondre à une requête, sont calculées par réécriture de la requête en termes de ces vues. Le problème consiste à trouver une requête qui est équivalente ou implique logiquement, la requête de l'utilisateur mais n'utilise que des vues. L'interrogation effective des sources se fait via des adaptateurs, appelés des wrappers en anglais, qui traduisent les requêtes réécrites en termes de vues dans le langage de requêtes spécifique accepté par chaque source.

Les différents systèmes d'intégration d'informations à base de médiateurs se distinguent par : d'une part, les langages utilisés pour modéliser le schéma global, les schémas des sources de données à intégrer et les requêtes des utilisateurs (schémas à base de règles, à base de classes ou combinaison des deux) et d'autre part, la façon dont est établie la correspondance entre le schéma global et les schémas des sources de données à intégrer. Concernant ce second point, on distingue l'approche Global As Views (GAV), qui consiste à définir le schéma global en fonction des schémas des sources de données à intégrer, et l'approche duale Local As Views (LAV). Les systèmes HERMES, TSIMMIS, MOMIS suivent GAV alors que Razor, Internet Softbot, Infomaster, Information Manifold, SIMS, OBSERVER et PICSEL suivent l'approche LAV. Les avantages et inconvénients de ces deux approches sont inverses [16]. GAV facilite l'expression des requêtes, LAV l'ajout de sources.

Plus récemment, sont apparus des médiateurs au dessus de données semi-structurées ayant le format de documents XML (C-Web, Xyleme), en attendant peut-être RDF. Ces systèmes sont fondés sur un schéma global à base d'arbres. Ils relèvent à la fois de l'approche GAV et LAV, la correspondance entre le vocabulaire du médiateur et celui des sources

étant exprimée par de simples correspondances de chemins. Rappelons qu'une autre approche est celle des entrepôts de données dont nous ne détaillerons pas les aspects techniques ici.

5.3 Intégration de sources hétérogènes et Web sémantique

L'intégration de sources d'information hétérogènes dans le cadre du Web sémantique pourra s'appuyer sur de multiples systèmes de médiation. Certains pourront suivre une approche centralisée telle l'approche décrite dans la section précédente. D'autres pourront suivre une approche décentralisée consistant à considérer une coalition de serveurs d'information. Chaque serveur peut jouer indifféremment le rôle de serveurs de données ou de médiateurs avec ses pairs en participant de manière distribuée et collective au traitement des requêtes des utilisateurs. Les connexions entre systèmes de médiation donneront au Web toute sa puissance, autorisant la recherche de données dans des sources non directement connectées aux sources du serveur interrogé.

On peut penser que dans le cadre du Web sémantique ces systèmes de médiation distribués seront mieux adaptés par leur flexibilité. Dans ce contexte de médiation décentralisée, apparaissent de nouveaux problèmes liés à la façon de connecter les différents systèmes mis en relation. Il faut trouver et définir des correspondances sémantiques entre les ontologies manipulées par chacun des systèmes. Il faut pouvoir disposer d'une approche simple et naturelle de description de telles correspondances sémantiques entre ontologies. Le passage à l'échelle du Web en matière de connexion entre ontologies n'est envisageable que si elle est en partie automatisée. Il est donc nécessaire d'étudier comment cette automatisation est possible, sachant qu'elle devra pouvoir être établie entre des ontologies qui sont locales à des sources et qui sont hétérogènes. Il est en effet illusoire de penser que toute ontologie ou schéma local à une source d'information sera toujours exprimée dans les termes d'un schéma global (ou ontologie globale) pré-existant(e). Enfin, il faut pouvoir définir et raisonner sur des correspondances entre ontologies de différentes sortes : égalité, inclusion, recouvrement.

6 SERVICES

La notion de services correspond à une approche spécifique des ressources disponibles sur le Web qui met l'accent sur les fonctionnalités

offertes par tel ou tel logiciel en terme d'un processus métier. Suivant la définition usuelle, un service permet à un utilisateur, non seulement d'obtenir de l'information, mais aussi d'effectuer des changements sur l'état du monde. Le commerce électronique est, bien sûr, un exemple privilégié de ces approches, mais l'exposition de tout processus fonctionnel sur le Web, comme la souscription d'une police d'assurance, relève également de cette problématique. Cette notion de services est aujourd'hui au cœur des stratégies d'un certain nombre des principaux acteurs du monde du logiciel comme Microsoft ou IBM.

De nombreuses techniques et plusieurs langages sont proposés comme des standards actuels ou futurs pour découvrir ou localiser les services Web (typiquement à travers un annuaire de services), pour invoquer ou activer ces services et les faire interopérer. De manière plus ambitieuse et moins aboutie aujourd'hui, il est proposé en plus de surveiller ou de suivre leur exécution (identifier les échecs, donner des traces) et surtout de les composer (sélection automatique, enchaînement et interopération). Parmi ces propositions, les plus connues sont SOAP (Simple Object Acess Protocol) qui décrit les modalités de l'échange d'information entre services en termes d'enveloppes d'échange, de fichiers contenus et principalement de routage, les annuaires UDDI (Universal Description, Discovery and Integration), WSDL (Web Services Description Language) qui décrit, de manière fonctionnelle et opérationnelle, comment utiliser un service et plus récemment WSFL (Web Services Flow Language) pour décrire des enchaînements de services comme un processus.

L'approche par les services Web, en terme applicatif, se construit sur les architectures à base de composants comme J2EE et COM puis .Net. Mais elle présente deux spécificités. D'une part, elle est plutôt orientée processus que programmes. Un service est le plus souvent un groupement logique de composants qui expose, pour ses clients, une fonction business, même s'il peut aussi faire appel aux applications développées avec les approches traditionnelles des systèmes d'information. D'autre part, elle repose sur les protocoles et les technologies standards d'Internet et du Web, avec différents niveaux d'utilisation, applications internes, applications développées pour des partenaires de confiance externes ou dans le cadre d'un marché ouvert, ce qui pose alors des problèmes de sécurité accrus.

Les approches services Web et celles du Web sémantique partagent le but commun de rendre l'information sur le Web plus accessible aux machines. Un certain nombre de recherches se proposent de les coupler. Une des idées partagées par ces travaux est de répondre aux limites de WSDL par l'ajout d'une couche sémantique (à base de marqueurs) au

dessus de WSDL décrivant le quoi et le pourquoi, pas seulement le comment.

En effet, WSDL fournit une description concrète mais de bas niveau d'un service Web, en termes de sa localisation, des opérations disponibles et des messages associés ainsi que des types de données et formats de leurs paramètres d'entrée ou de sortie. Ces descriptions sont insuffisantes pour qu'un agent logiciel puisse interpréter la signification réelle des opérations WSDL. Pour que cette interprétation soit possible, on peut annoter sémantiquement les services Web. L'ontologie de l'agent logiciel doit alors être commune (ou appariable) à celles utilisées pour l'annotation. Les capacités de raisonnement des langages du Web sémantique deviennent ainsi utilisables, en particulier pour la découverte et la composition de services Web. Parmi les recherches qui se fixent pour but d'enrichir l'approche des services Web, on peut citer DAML-S qui propose une ontologie de haut niveau des services Web sous forme d'un ensemble de classes. La classe SERVICE qui est le haut de l'ontologie DAML-S se voit ainsi associer un ensemble de connaissances par l'intermédiaire de deux classes. La première SERVICEPROFILE fournit, par un ensemble d'attributs et de propriétés, l'information nécessaire pour qu'un agent puisse découvrir un service. La deuxième SERVICEMODEL décrit comment utiliser le service en des termes plus abstraits que WSDL. De nombreuses autres classes sont associées à celles-ci afin de donner des descriptions conceptuelles des services Web permettant par exemple d'apparier une requête et une description de service en raisonnant à l'aide de l'ontologie.

7 CONCLUSION

Quelles seront les clés de la pénétration des technologies du Web sémantique ? Comme pour toute nouvelle technologie dont les usages potentiels sont nombreux, il est difficile de prévoir lesquels prévaudront et comment telle ou telle catégorie de professionnels ou d'utilisateurs trouveront un bénéfice réel aux nouvelles possibilités offertes. Il est néanmoins possible de repérer d'ores et déjà des obstacles à la diffusion du Web sémantique. Dans une vision prospective, [4] souligne tout un ensemble de recherches qu'il serait utile de développer pour contribuer à lever ces obstacles. Faute de place, nous nous contenterons d'insister, dans cette conclusion, sur deux de ces obstacles qui sont particulièrement cruciaux pour les débuts même du Web sémantique.

Le premier, indéniable, est la diversité et la complexité des langages tels qu'ils sont actuellement proposés par le W3C. Cela est sans doute inévitable dans cette phase initiale. Mais il est sans doute bon de rappeler que des raisons du succès d'HTML sont la diversité de ses utilisations et sa simplicité ainsi que celle des outils permettant sa mise en œuvre. De même, XML reste relativement simple pour la réalisation d'applications dans différents métiers et surtout est maintenant bien maîtrisé par de nombreux développeurs. Même si on n'adhère pas complètement à ce que James Hendler[6] écrit, on peut affirmer que la convivialité des outils pour la mise en œuvre des langages du Web sémantique sera ainsi une des principales clés.

Le deuxième obstacle provient du fait que la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes. Les expériences des chercheurs et des praticiens de la documentation sont éclairantes de ce point de vue. La difficulté dans le cas de connaissances plus formalisées est évidemment accrue. Les expériences dans la construction d'ontologies sont, ici aussi, instructives et pourraient contribuer à lever quelques illusions.

Comme le souligne [4], dans l'idéal les méta-données et les annotations sémantiques devraient être un sous-produit automatique des activités usuelles des différents types d'utilisateurs. Même si on ne peut penser atteindre cet objectif en toute généralité, des avancées dans cette direction doivent être l'objet de recherches. On peut, à ce propos, se poser la question de savoir si le Web sémantique se généralisera ou restera cantonné dans des communautés réduites de professionnels.

La voie semble, en tout cas, ouverte pour deux visions complémentaires du Web sémantique. La première met plus l'accent sur la réalisation d'outils logiciels utilisant des représentations munies de sémantique formelle et des mécanismes inférentiels puissants, avec un coût souvent élevé de construction et de maintenance des connaissances. La deuxième met plus l'accent sur des représentations semi-formelles et repose plus sur l'utilisateur pour leur exploitation opérationnelle. Elle peut, à court terme, être plus souple à réaliser et finalement correspondre mieux aux fonctionnements cognitifs de ces utilisateurs. Le débat, non contradictoire, est ouvert.

---

[6] « Sur le Web, l'expressivité est le baiser de la mort, les langages et les solutions plus simples vont plus loin que les plus complexes » in [17].

8 RÉFÉRENCES

[1] BERNERS-LEE Tim, HENDLER James and LASILLA Ora, The Semantic Web, Scientific American, May 2001.

[2] BROEKSTRA Jean, KAMPMAN Arjohn and VAN HARMELEN Frank, Sesame: A Generic Architecture for Storying and Querying RDF and RDF Schema, in [7], p.54-68.

[3] CRUZ Isabel, DECKER Stefan, EUZENAT Jérôme and MCGUINESS Deborah (eds), The emerging Semantic Web, Selected papers from the first Semantic web working symposium, IOS press, Amsterdam (NL), 2002.

[4] EUZENAT Jérôme (ed), Research challenges and perspectives of the Semantic Web, Report of the EU-NSF strategic workshop, Sophia-Antipolis, Octobre 2001

[5] FENSEL Dieter, HENDLER James, LIEBERMAN Henry and WAHLSTER Wolfgang (eds), Spinning the Semantic Web : Bringing the World Wide Web to Its Full Potential, The MIT Press, 2002.

[6] GOMEZ PEREZ Asuncion (ed), A Survey of Ontology Tools, Deliverable 1.3, Ontoweb, May 2002.

[7] HORROCKS Ian and HENDLER James (eds), The Semantic web – ISWC 2002, Proceedings of the "First International Semantic Web Conference", Sardinia, June 2002, LNCS 2342, Springer 2002.

[8] HYVÖNEN Eero (ed), Semantic Web Kick-Off in Finland, Vision, Technologies, Research and Applications, HIT Publications, 2002.

[9] KIMBALL Ralph, MERZ Richard, The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse, John Wiley & Sons, January 2000.

[10] KOIVUNEN Marja-Ritta and MILLER Eric, W3C Semantic Web Activities, in [8], p.27-76.

[11] KOIVUNEN Marja-Ritta, Annotea: Applying Semantic Web Technologies to Annotations, in [8], p.213-226.

[12] MAGKANARAKI Aimilia et al, Benchmarking RDF Schemas for the Semantic Web, in [7], p.132-146.

[13] NILSSON Mikael, PALMÉR Matthias, NAEVE Ambjörn. (2002), Semantic Web Meta-data for e-Learning, Some Architectural Guidelines, Proceedings of the 11th World Wide Web Conference, Hawaii, 2002.

[14] PATEL-SCHNEIDER Peter and FENSEL Dieter, Layering the Semantic Web, in [7], p.16-29.

[15] PATEL-SCHNEIDER Peter and SIMÉON Jérôme, Building the Semantic Web on XML, in [7], p.147-161.

[16] ROUSSET Marie-Christine, BIDAULT Alain, FROIDEVAUX Christine, GAGLIARDI Hélène, GOASDOUE François, REYNAUD Chantal, SAFAR Brigitte,

Construction de médiateurs pour intégrer des sources d'intégration multiples et hétérogènes : le projet PICSEL, revue I3, vol.2, n°1, p.5-59, 2002.

[17] STAAB Stefen (ed), Ontologies'KISSES in Standardization, IEEE Intelligent Systems, March-April 2002, p.70-79.

[18] WIEDERHOLD Gio, Mediators in the architecture of future information systems, Computer, 25(3): p.38-49, 1992.