Application du Web sémantique: vers l'avènement du balisage sémantique et des modélisations des connaissances évolutives?

Lise VERLAET

CERIC-LERASS
Route de Mende
34 199 Montpellier Cedex 5
e-mail: lise.verlaet@univ-montp3.fr

Résumé. *L'être humain aime à ranger, ordonner, répertorier, classifier, organiser le monde qui l'entoure. Ce que l'on pourrait qualifier d'obsession de l'homme vis-à-vis de ce qui l'environne n'avait jusqu'à présent pas véritablement atteint l'espace virtuel. Voilà qui est fait, reste à savoir comment le structurer. Car il ne suffit pas de le conceptualiser pour rendre le Web moins opaque, il faut également l'organiser pour mieux s'y retrouver. L'intentionnalité sous-jacente à cette organisation du Web est de conférer aux systèmes informatiques les capacités de «reconnaître» les ressources en les qualifiant, de les mettre en relation afin de pouvoir «raisonner» sur ces informations, de sorte à assurer une meilleure qualité de service à l'homme.*

C'est là la mission que tente de relever les acteurs du Web sémantique: conceptualiser et organiser les ressources numériques et leur contenu afin d'assurer aux utilisateurs des recherches d'information plus pertinentes. Appliquer les principes du Web sémantique n'est pas chose facile. En effet, il est à la fois nécessaire de procéder à la sémantisation des corpus pour en révéler les concepts, fait qui demande une très bonne connaissance du domaine d'application. Mais il faut également avoir de solide connaissance en informatique pour respecter l'architecture prônée par le Web sémantique. N'étant ni expert ni informaticien, nous nous sommes néanmoins prêtées au jeu et proposons, à travers cet article, une approche différente

ubb media centre
Journal of Media Research, 8/2010, pp. 12-24

de l'indexation des documents via le balisage sémantique, lequel va nous permettre de constituer des modélisations des connaissances évolutives.

Mots-clés: *Web sémantique, balisage sémantique, XML, ontologie, réseau sémantique, modélisations des connaissances.*

1. Rappel sur le projet d'un Web sémantique

Le désormais célèbre projet de Tim Berners-Lee et du W3C, la création d'un Web sémantique, a été mis en place pour pallier aux désagréments du Web hypertexte au sein duquel la recherche d'information s'avère génératrice de surcharge cognitive et de désorientation dans le cyberespace. En effet, la masse devenue informe des documents opacifie la Toile et rend les recherches d'information délicates car celles-ci sont essentiellement basées sur les mots énoncés au sein des ressources. De fait, le Web sémantique repose sur une idée simple: donner davantage de renseignements sur le contenu des documents pour permettre une meilleure organisation des ressources et faciliter ainsi leur traitement par les systèmes informatiques de sorte à répondre plus justement aux requêtes des internautes. Pour Berners-Lee et Al. (Berners-Lee et Al., 2001) «*le Web Sémantique est une extension du Web actuel dans lequel les informations sont définies de manière précise, afin de permettre une meilleure coopération de travail entre la machine et l'homme*». Clarifier les informations sur le Web, c'est assurément apporter de la clarté quant à leur utilisation prétendue et/ou possible. Donner aux systèmes d'exploitation informatique des renseignements qualitatifs sur les informations qu'ils ont à traiter, c'est offrir un service de recherche plus pertinent aux usagers du Web. Mais c'est également permettre aux machines, via une infrastructure spécifique, de partager et de réutiliser les informations qu'elles contiennent, et subséquemment d'assurer l'interopérabilité entre les systèmes informatiques.

Le Web Sémantique recommande un ensemble d'outils syntaxiques spécifiques aux traitements des ressources numériques. Il repose sur une architecture pyramidale composée de différents langages de représentation des connaissances contenues dans le Web. Cette architecture propose une combinaison de principes, eux-mêmes convertis en langages informatiques. Chaque langage de couche supérieure vient spécifier les couches de niveau inférieur. En effet, chacun de ces langages de représentation vient compléter les informations du langage précédent, ajoutant au fur et à mesure divers formalismes pour le traitement des données. Ces couches sémantiques de complexités croissantes viennent ainsi graduellement spécifier les informations inhérentes aux ressources par rapport à un domaine et à un usage donné.

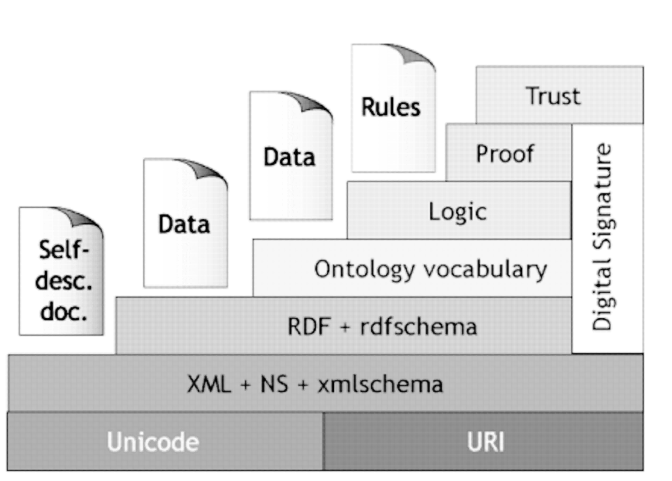

Figure 1. Architecture du Web Sémantique selon Tim Berners-Lee (2000)

A la base de l'infrastructure du Web Sémantique se trouve l'URI (*Uniform Resource Identifier*) qui permet d'attribuer un identifiant unique à une ressource assurant ainsi la localisation des différentes ressources sur la Toile. Ce protocole d'identification des ressources est commun au Web hypertexte, c'est la couche la plus stable de l'architecture.

Nous pouvons d'ailleurs considérer que c'est véritablement avec le langage XML (*eXtensible Markup Language*) que débute l'action de sémantisation du Web. XML est la syntaxe sur laquelle repose l'ensemble de l'architecture ascendante du Web Sémantique. Ce langage va permettre d'apposer aux documents des métadonnées via des balises sémantiques (également appelées marqueurs ou annotations sémantiques). Ces métadonnées décrivent le contenu des documents, les rendant ainsi interprétables par les systèmes informatiques. Les balises XML vont permettre de construire un index, une base de données formulée à partir des concepts contenus dans les ressources. En ce sens, le balisage sémantique est comparable à l'indexation documentaire, comme elle, la qualification des ressources demande un important effort d'analyse et de catégorisation des informations pertinentes.

Le W3C recommande d'enrichir les informations existantes à l'aide du langage RDF (*Resource Description Framework*). Ce troisième langage de l'architecture du Web Sémantique est un modèle de métadonnées permettant de référencer et de lier les ressources. Il permet de décrire de manière formelle le contenu d'une ressource. Le modèle RDF garantit la représentation des métadonnées, facilitant ainsi leur exploitation par les systèmes informatiques. Ce faisant, RDF assure une certaine opérabilité entre des ressources du Web, y compris de celles qui ne sont ni formalisées ni structurées. Il permet donc à une même communauté d'utilisateurs de partager et d'échanger des informations via ce modèle de métadonnées. Car ce modèle de

métadonnées est une première représentation des connaissances liée à un domaine ou une activité de recherche d'information. Bien que critiqué par certains puristes, RDF offre une solution simple qui engendre de plus en plus d'émules.

Puis, nous trouvons la couche «Ontology vocabulary», dont le langage le plus usité à l'heure actuelle est OWL (*Ontology Web Language*). Le langage OWL consiste à conceptualiser de manière formelle un monde de connaissance de sorte à pouvoir générer des règles d'inférences exploitables par les machines. OWL étend les capacités de représentation de RDF au moins tout autant que sa complexité.

L'avant-dernière couche de l'architecture du Web Sémantique est la couche «logique» en référence à la logique de traitement des informations sous-jacentes au projet informatique. Cette dernière s'appuie sur les règles d'inférences de la couche inférieure, c'est-à-dire sur les ontologies, pour traiter les données et ainsi produire des résultats en accord avec la requête de l'utilisateur. Cette couche est généralement opérée à l'aide d'algorithmes pilotant l'agent de recherche au sein des informations. Cela peut également être toutes sortes de programmes experts dont l'objectif est de traiter les couches inférieures afin de présenter à l'utilisateur les informations pertinentes en fonction de sa recherche. C'est là tout l'intérêt porté par le Web Sémantique: clarifier et ordonner les ressources et plus largement les informations afin que ces dernières soient transférables et exploitables par les systèmes informatiques. Mais c'est aussi contribuer à une amélioration de la qualité des recherches pour les usagers du Web.

Enfin, au sommet de la pyramide, les couches «preuve» et «validité» qui bien que séparées dans les strates de l'architecture du Web Sémantique sont souvent associées. Ces couches permettent l'authentification et la validation des ressources du Web. Le Web étant un espace virtuel où chacun peu déverser des ressources de tout ordre, il est nécessaire de pouvoir contrôler la validité de leur contenu, notamment à l'aide de la signature électronique. L'utilité de ces couches peut paraître quelque peu superflue, cependant la démocratisation des principes du Web Sémantique peut conduire des sites peu scrupuleux à biaiser les métadonnées. Il suffit pour cela d'ajouter des métadonnées décrivant des informations populaires chez les utilisateurs afin de s'assurer une meilleure visibilité sur la Toile. C'est d'ailleurs une pratique courante. A travers ces dernières couches le Web Sémantique fait en quelque sorte signer un contrat de bonne conduite aux diffuseurs des ressources numériques.

S'il est vrai que l'architecture du Web Sémantique se décompose en plusieurs couches interreliées, les langages XML, RDF et OWL résident au cœur des préoccupations. En effet, les différents travaux que nous avons étudiés (Bach, 2006; Iksal, 2002; Ranwez, 2000; Ta, 2005; Villanova-Olliver, 2002) nous ont permis de révéler deux principes forts pour la sémantisation des informations, d'une part la conceptualisation des ressources et de leur contenu; et d'autre part l'organisation de cette conceptualisation.

2. XML et le concept de balisage sémantique

L'indexation analytique documentaire est un concept fondamental sous-jacent au Web Sémantique. Ce processus permet de qualifier les ressources et ainsi permettre aux systèmes informatiques de retrouver plus facilement l'information utile et pertinente pour les besoins de l'utilisateur. L'indexation analytique documentaire repose sur un vocabulaire contrôlé correspondant le plus souvent à un modèle de métadonnées préétabli. Ainsi, après analyse du contenu de la ressource, seuls les termes du vocabulaire pourront être utilisés pour qualifier les métadonnées la concernant. Cette technique d'indexation a largement fait ses preuves dans le monde de la documentation, elle permet de disposer d'un index commun pour la sémantisation des ressources. Cet index était le plus souvent rédigé en langage artificiel, ce, jusqu'à l'arrivée de XML et par là même de la possibilité de créer des métadonnées en langage naturel.

2.1. XML, le langage du Web sémantique

Comme le HTML, le XML est également un dérivé du SGML. Développé sous l'égide du W3C, le XML conserve la simplicité du HTML et la richesse sémantique du SGML. Le XML est un langage informatique à balises extensibles ou libres, elles permettent ainsi de créer autant de balises que nécessaire et ce, en langue naturelle. De fait, le contenu des balises est sémantiquement malléable de sorte à être adapté à la description des documents quel que soit leur domaine d'applications. C'est donc au créateur de documents XML de choisir le vocabulaire spécifiant le contenu des balises en fonction des ressources et/ou conformément à un répertoire de termes. Car si le HTML se concentre sur la mise en pages des ressources Web, le XML s'attache à leur contenu, en ce sens il est considéré comme un métalangage puisqu'il permet de qualifier le contenu des documents Web. Les balises XML sont des données sur les données portées par les ressources. Le XML est donc un langage pour les métadonnées. Il facilite ainsi l'élaboration de langages à balises spécialisées. De fait, il n'est plus nécessaire de recourir à des langages artificiels. Ceci a pour principal effet de considérablement faciliter la tâche d'indexation, et par conséquent, de pouvoir développer de nouvelles méthodes de description des contenus jusque là délaissées, étant donné le travail colossal qu'elles représentaient.

2.2. Le balisage sémantique

Le langage XML permet de créer librement des métadonnées sur le contenu des textes. Il favorise ainsi l'émergence de métadonnées type «annotation» ou «commentaire», qui sont alors considérées comme «interprétatives» ou «subjectives». Ces métadonnées peuvent baliser les ressources et leurs contenus en fonction des intentionnalités projetées lors de l'élaboration du système d'information. L'emploi des métadonnées n'est donc plus restreint à la formulation d'un index «objectif», car elles sont désormais en mesure de répondre à différents contextes d'utilisation. Muriel Amar

(Amar, 2004) constate que «*si les méthodes d'indexation, linguistique ou structurelle, bouillonnent avec l'expansion du Web, elles ne sont pas encore parfaitement adaptées aux besoins des utilisateurs*». Nous abondons dans son sens, qui plus est, nous sommes convaincus que satisfaire les besoins des utilisateurs passe par de nouvelles approches des ressources numériques et notamment par le balisage sémantique.

De notre point de vue, le balisage sémantique se situe au cœur du concept d'adaptation des informations aux lecteurs. Le balisage sémantique est un processus intentionnel qui varie selon les domaines d'application, selon le projet dans lequel il s'inscrit. Ainsi chaque concepteur de dispositif socio-technique doit mener une importante réflexion préalable sur le système de balisage à mettre en place, car le balisage sémantique se trouve également au centre du processus de médiation avec l'utilisateur. En d'autres termes, le balisage doit être pensé en vue d'offrir aux utilisateurs un outil qualitatif susceptible de les assister efficacement dans leurs recherches d'information et dans la construction de leurs connaissances, en leur proposant une lecture des informations adaptées à leurs besoins. Il s'agit donc de passer d'une indexation «objective et formelle» à la notion de balisage sémantique «contextuel» où la notion de sens contextuel est liée à une intention de recherche.

Le balisage sémantique consiste à analyser les corpus afin d'en faire surgir les passages porteurs de sens pour les utilisateurs compte tenu de la finalité de leur recherche. Cela nécessite de dégager, de conceptualiser et contextualiser les différents thèmes étudiés. Nous utilisons les termes de balisage sémantique au détriment de celui d'indexation pour trois raisons.

- D'une part parce que ce balisage est exprimé en langage naturel (XML), il est donc à distinguer des langages artificiels généralement employés pour l'indexation des ressources.
- D'autre part, on retrouvera le terme de balisage sémantique lorsque les balises sont placées au cœur même des documents, ni en prélude ni à la fin, mais bien au sein des textes. Nous ne nous concentrons plus seulement sur le document, mais bien sur les contenus de ces derniers, en effectuant une étude fragmentaire de l'ensemble des corpus. Grâce au balisage sémantique, les lecteurs peuvent consulter les corpus en fonction de leurs centres d'intérêts, sans être contraints à la relative linéarité des textes. Cette démarche de «lecture tabulaire» n'est pas récente, prenons l'exemple *La vie mode d'emploi* de Georges Perec, dont l'index très détaillé, permet au lecteur de suivre le récit du personnage de son choix à travers les différents chapitres. En d'autres termes, le lecteur a la possibilité de découvrir, de consulter les informations de son choix. Il en va de même avec le balisage sémantique. Le balisage sémantique permet de repérer et de qualifier les fragments de textes pertinents du point de vue de l'utilisateur, en fonction d'une intention de recherche afin qu'il puisse y avoir directement accès. De fait on ne se situe plus dans une étude globale des ressources mais dans une analyse fine de leurs contenus.

- Enfin le balisage sémantique diffère de l'indexation parce qu'il peut être intentionnel et donc interprétatif. En effet, il n'est pas purement descriptif et formel, état de fait lié à un vocabulaire contrôlé et à un langage artificiel. Mais il peut également être de l'ordre de l'annotation ou du commentaire, et par conséquent être subjectif et interprétatif. Le baliseur formule un découpage interprétatif du texte qui peut être dépendant du projet dans lequel s'inscrit la lecture supposée de l'utilisateur. Les termes de cette annotation sont traduits sous forme d'indices pour l'utilisateur, lesquels serviront à orienter ou guider ce dernier dans le corpus documentaire.

De fait, le balisage sémantique se situe au cœur du processus de médiation entre le texte et ses différents types de lecteurs. Le balisage sémantique étant relatif à une intention de recherche, il existe autant de balisages possibles que d'intentions de recherche formulables. Ainsi, les problématiques inhérentes au balisage sémantique sont intrinsèquement liées aux problématiques de lecture et de recherche d'information des utilisateurs.

3. Les modélisations des connaissances

Les recherches préalablement menées en Intelligence Artificielle sur les représentations des connaissances nécessitaient un important consensus sur la définition exacte de concepts communs pour le partage des données. Le Web Sémantique reprend les apports sur les représentations des connaissances de l'Intelligence Artificielle, et notamment le concept d'ontologie, tout en laissant aux concepteurs une grande souplesse quant à leurs formulations. Le Web Sémantique conçoit l'ontologie comme un outil permettant de clarifier un domaine de connaissance, de partager et de réutiliser ces connaissances, ces données, tant pour les utilisateurs que pour les systèmes informatiques.

Nous avons vu que RDF est un langage de représentation des connaissances qui permet de référencer, structurer et mettre en relation les ressources Web. Ce langage doit permettre aux agents de recherche d' *«avoir accès à des collections structurées d'informations et d'ensembles de règles d'inférence qu'ils peuvent utiliser pour parvenir à un raisonnement automatisé».* Berners-Lee et Al. (2001). Mais aussi que les capacités de RDF se voient étendues par la couche «ontology vocabulary» de l'architecture du Web Sémantique.

3.1. Ontologie

Le concept d'ontologie est au cœur des problématiques du Web Sémantique. Si de manière générale une ontologie peut être définie comme la modélisation conceptuelle partielle et partagée d'un monde de connaissance (Gruber, 1993; Guarino, 1997), il existe néanmoins des divergences quant à la formulation de cette dernière. Nous distinguerons deux méthodes de constitution d'ontologie, lesquelles diffèrent selon la rigueur, le formalisme porté à la définition des relations entretenues entre les concepts.

La première méthode inhérente à la formulation ontologique vise à organiser les concepts sous une structure hiérarchique arborescente via les relations de spécification de type «est un» ou «est une sorte de». Cette formulation ontologique, notamment soutenue par Sowa (Sowa, 2000), permet d'obtenir des ontologies dites formelles. L'ontologie peut donc être perçue comme une grille avancée de sémantisation de tous les contenus d'un domaine ou d'une application. Elle fournit les rubriques (concepts) et les catégories (sous-concepts) ainsi que les liaisons hiérarchiques entre ces éléments.

La seconde méthode répond à la constitution d'ontologies dites informelles. Les ontologies informelles ne sont pas restreintes à des relations de spécifications mais recouvrent l'ensemble des relations possibles entre les concepts dont la structure forme un réseau sémantique. A ce titre, nous retrouvons de plus en plus dans la littérature l'ontologie informelle sous la dénomination de réseau sémantique.

Par ailleurs, quel que soit le formalisme adopté pour la constitution des ontologies, nous pouvons observer différents types ou niveaux lesquels sont dépendants du projet à réaliser et donc des fonctionnalités données à ces modélisations des connaissances. Les travaux de Guarino (Guarino, 1997) nous permettent de répertorier quatre principaux niveaux:

- *Les ontologies de haut niveau ou supérieure*s qui formalisent des concepts très généraux tels que le temps, l'espace, la matière (…). Ces concepts n'appartiennent pas à un domaine en particulier, ils sont universels et collectivement partagés. L'objectif affiché de ce type d'ontologie est la création d'une modélisation conceptuelle universelle. Dolce, Sumo ou encore BFO sont des exemples d'ontologies de haut niveau.
- *Les ontologies de domaine* sont inhérentes à un domaine particulier et contiennent les concepts qui lui sont spécifiques. Les ontologies du domaine viennent spécifier les concepts d'une ontologie de haut niveau. En la matière, les domaines de la médecine et de la pharmacologie ont pris de l'avance et disposent de nombreuses ontologies.
- *Les ontologies de tâche* s'attachent à la qualification des tâches génériques réalisées dans un domaine donné, par exemple, évaluer, diagnostiquer (…). Les ontologies de tâches spécialisent les ontologies de domaine.
- *Les ontologies d'application* convoquent les concepts relatifs à l'ontologie du domaine et d'une tâche de sorte à décrire l'exécution d'une activité particulière comme apprendre à réparer un téléviseur.

Les ontologies sont des modèles d'organisation des données, ils fournissent des représentations partielles et formelles d'un monde de connaissances. Les ontologies assurent la couverture conceptuelle du domaine ou de l'application, mais aussi son agencement. C'est à partir du système de relation inhérent aux modèles de représentation, aux règles d'inférence qu'ils sous-tendent – lesquelles ont été préalablement établis par des communautés d'utilisateurs – que les machines vont s'appuyer pour traiter les ressources Web. Ces modélisations des données, que l'on peut assimiler à des modélisations des connaissances si l'on considère que les

connaissances sont des données ayant un sens (ce sont des balises sémantiques), sont autant d'informations structurées sur le raisonnement et l'utilisation des ressources de diverses communautés d'utilisateurs. La modélisation des données offre par là même un langage compréhensible par les machines sur l'exploitation attendue par les utilisateurs des informations.

Via les ontologies, le Web Sémantique propose à ces utilisateurs un Web clarifié et organisé, non plus basé sur une anarchie d'entités lexicales mais sur des concepts reliés entre eux en fonction de leur sens. Grâce aux données inhérentes aux modélisations des connaissances, il souhaite résorber le phénomène de surcharge cognitive en ne présentant à l'utilisateur que des ressources répondant à sa demande conformément à un modèle d'utilisation. De plus, la structure des modélisations, les liens qui organisent les données, permettent aux machines de construire des parcours pour l'utilisateur, des orientations possibles au sein des ressources. Et enfin, elles participent à l'interopérabilité entre les systèmes informatiques de par des modèles communs pour le transfert et le partage des informations.

3.2. Construction des modélisations des connaissances

Si le Web sémantique propose plusieurs langages informatiques pour transférer les modélisations des connaissances aux machines, il n'en reste pas moins qu'il est au préalable nécessaire de les formuler, ce qui n'est pas une mince affaire. Généralement ces modélisations sont construites par un comité d'experts. La mise en place d'un comité d'expert pose toutefois un certain nombre de contraintes. Outre les difficultés liées à leurs emplois du temps surchargés en leurs qualités d'experts et donc à leurs rencontres, des divergences d'opinions sont inévitables. On retrouve ces mêmes problématiques lors de l'établissement de normes d'indexation. La constitution des modélisations des connaissances nécessite par là même un temps de réflexion et de mise au point important avec comme risques une formulation complexe, difficile d'appréhension et d'application. C'est pourquoi la plupart des projets font appel à un seul expert, voire deux. La principale critique pouvant être portée à cette solution est bien entendu la non-exhaustivité de la modélisation ainsi révélée puisqu'elle sera le reflet de la pensée, du système de pertinence de l'expert en question. Toutefois, suivant l'importance du projet à mettre en place, cette dernière solution est celle qui demande le moins de temps et d'investissement humain, ce qui n'est pas négligeable sachant que les modélisations des connaissances sont sujettes à une perpétuelle évolution.

3.2.1. Construction faite a priori

Les modélisations des connaissances telles que nous les rencontrons actuellement en informatique se sont largement inspirées des modèles type thésaurus ou taxonomie basés sur un vocabulaire contrôlé. Dans cette perspective, constituer une modélisation des connaissances nécessite de rassembler la liste des concepts inhérents au domaine étudié. Cette liste, si elle ne possède aucun support préalable (aucun modèle préexistant ou accessible), est le plus souvent publiée de manière intuitive et empirique pour

former la base du vocabulaire contrôlé à employer. Ce vocabulaire ne sera plus nécessairement rédigé en langage artificiel grâce aux balises XML. C'est ce vocabulaire conceptuel qui va être utilisé pour l'indexation ou le balisage des ressources mais aussi pour concevoir et échafauder les représentations des connaissances. Par ailleurs, construire les modélisations peut faire émerger un certain nombre de concepts absents du vocabulaire, il peut donc y avoir quelques allers-retours entre la modélisation des connaissances et la liste du vocabulaire pour corriger ses oublis.

Nous voyons cependant plusieurs inconvénients sous-jacents à cette méthode de construction des modélisations. Le premier réside dans le processus même d'indexation ou de balisage. En effet, seuls les termes inhérents à la liste du vocabulaire peuvent être employés pour qualifier les métadonnées ou les balises. Or un domaine ou une discipline active voit forcément son champ s'agrémenter de nouveaux concepts, lesquels ne pourront donc pas être pris en compte par le balisage puisqu'ils ne sont pas présents dans la liste du vocabulaire. Le contre-argument généralement émis à cet égard est que tout nouveau concept n'est pas foncièrement à même de figurer dans la liste du vocabulaire et donc dans les modélisations des connaissances. Et qu'en quelque sorte ce concept doit être éprouvé avant de rejoindre la liste des concepts officiels et reconnus. Certes c'est là un très bon argument, mais l'avancée scientifique ne se fait-elle pas aussi en contrecarrant, en appuyant ou encore en s'inspirant des concepts émergents? De fait, les occulter, ne serait-ce que temporairement pour certains, ne met-il pas un frein à la recherche scientifique? Comment évaluer l'impact d'un concept si on le rend invisible? Par ailleurs, si nous exagérons ce raisonnement «par restriction des concepts», quiconque ayant déjà eu à indexer un corpus documentaire a pu s'apercevoir que tous les termes du vocabulaire ne sont pas nécessairement utilisés. Doit-on pour autant supprimer de la liste du vocabulaire certains concepts reconnus mais non employés dans l'indexation des ressources? Il est évident que non, tout comme il nous est évident que tout nouveau concept doit figurer dans les modélisations des connaissances.

Le second inconvénient soulevé par l'utilisation de cette méthode d'élaboration des modélisations des connaissances, soit à partir d'une liste préétablie du vocabulaire, découle en partie de ce que nous venons d'exposer. Nous avons vu qu'il peut y avoir plusieurs allers-retours entre la liste du vocabulaire et la modélisation, dus à quelques oublis de concepts. Allers-retours d'autant plus fréquents qu'il sera inévitable d'ajouter les nouveaux concepts finalement jugés acceptables pour figurer dans la représentation du domaine. Seulement l'ajout de nouveaux concepts à la liste du vocabulaire implique nécessairement la reformulation de la liste ainsi que celle de la modélisation, mais aussi et surtout, de devoir réindexer l'ensemble du corpus documentaire. Tout ceci entraînant un travail considérable.

3.2.2. Construction à partir des concepts extraits du balisage sémantique

Une autre méthode de construction des modélisations des connaissances, certes bien moins fréquente, consiste à extraire la liste du vocabulaire directement depuis

les corpus à baliser. C'est donc le baliseur qui, après analyse et balisage des ressources documentaires, dresse la liste du vocabulaire conceptuel à partir duquel va être établie la représentation des connaissances du domaine. Cette méthode présente le principal avantage de diminuer les allers-retours entre vocabulaire, modélisation et balisage, et subséquemment de réduire le nombre d'interventions humaines. En effet, procéder au balisage permet d'obtenir un balisage précis des ressources mais aussi de répertorier tous les concepts relatifs au corpus. De fait, si nouveau concept il y a, ce concept sera automatiquement comptabilisé dans la liste du vocabulaire à employer pour concevoir la modélisation des connaissances. L'utilisation de cette méthode nécessite toutefois un travail plus rigoureux de la part du baliseur qui doit à la fois repérer les concepts mais également veiller à l'homogénéisation du vocabulaire en respectant à la lettre les règles de qualification des balises telles qu'énoncées par Hensens (Hensens, 2002). Par ailleurs, la liste du vocabulaire ainsi obtenue devra certainement être complétée pour la formulation de la modélisation. Tous les concepts ne sont pas toujours employés dans les ressources balisées et certains manqueront à l'appel, il reviendra donc aux experts chargés de la constitution de la représentation des connaissances de les ajouter pour assurer sa cohérence. Cependant cette tâche ne bouleversera en rien le travail de balisage, et viendra compléter la liste du vocabulaire à utiliser pour les prochains balisages à effectuer. Cette démarche est donc de notre point de vue beaucoup plus constructive que celle généralement utilisée.

Malheureusement quelle que soit la méthode employée pour constituer les modélisations des connaissances, aucune ne peut se passer de l'intervention des experts pour leur élaboration. En effet, ces méthodes assurent l'établissement d'une liste de vocabulaire à employer pour construire les modélisations, cependant elles ne fournissent pas de repères quant à la formulation de ces dernières. La mise en relation des différents concepts devant composer la modélisation reste donc à la charge de l'expert ou du comité d'experts. La construction des représentations des connaissances est donc réalisée *a priori* en fonction du système de pertinence des experts. Et le travail inhérent à la modélisation des connaissances, à l'assemblage, à la mise en relation des concepts est long et fastidieux. Il est de fait essentiel pour démocratiser l'utilisation des modélisations des connaissances de trouver une méthode facilitant le travail de l'expertise.

4. Vers des modélisations des connaissances évolutives

Mais comment faciliter le travail des experts pour la construction des modélisations des connaissances? Comme vous l'aurez compris, une première solution consiste à élaborer ces dernières via les concepts émergeants du balisage sémantique des ressources numériques. Poussons plus avant ce raisonnement. Si nous répertorions les concepts à partir du balisage sémantique, ne pouvons-nous pas extraire des textes des indices quant à la constitution des ontologies et réseaux sémantiques?

Nos travaux relatifs à la conception-réalisation d'une revue scientifique adaptative, soit une revue dont le contenu a été balisé de sorte à répondre aux intentionnalités de lecture des utilisateurs, nous permettent d'affirmer que les articles scientifiques regorgent de fragments d'information élaborée mettant en évidence les différents liens qui unissent les concepts. Par fragments d'information élaborée nous entendons des bribes de textes ayant un sens hors contexte, c'est-à-dire une signification en dehors de son article d'origine. En effet, les auteurs utilisent des «faits scientifiques» connus et reconnus par la communauté pour introduire ou assoir leurs raisonnements. Ce faisant, ils mettent ou tissent des relations entre les concepts, fragments d'information élaborée qui vont nous permettre de construire petit à petit les modélisations des connaissances du domaine. Ces dernières pourront être considérées comme évolutives car chaque auteur viendra apporter une pierre à l'édifice, le transformant au fil des avancées scientifiques.

Se pose alors la question de la validité scientifique des ontologies et des réseaux sémantiques si l'on utilise ce mode de construction. Nous n'affirmons pas que toutes les ressources du Web puissent faire l'objet d'un tel traitement, d'ailleurs là n'est pas non plus l'objectif du Web sémantique. Cependant les articles scientifiques semblent être des ressources toutes indiquées pour une telle exploitation puisque leur contenu a d'ores et déjà contrôlé par un comité scientifique faisant autorité dans le domaine. S'il est vrai que via ce procédé nous pouvons nous passer de l'intervention des experts du domaine, rien n'empêche de les solliciter pour garantir la cohérence des modélisations des connaissances. Mais nul doute que leur tâche en sera grandement facilitée puisque ces modélisations des connaissances seront le fruit d'une intelligence collective. Enfin, un autre avantage offert par cette méthode est que le balisage de ces fragments d'information élaborée nous permet d'obtenir des explications quant aux relations entretenues par les concepts. Nous sommes ainsi plus à même de comprendre les modélisations des connaissances du domaine et donc de raisonner sur celui-ci et par là même de le faire avancer.

Références

AMAR M., «L'indexation aujourd'hui», *La fonction documentaire au cœur des TICE*, Les Dossiers de l'Ingénierie Educative, n° 49, décembre 2004.

BACH T.L., *Construction d'un Web sémantique multi-points de vue*. Thèse de doctorat 2006, en Informatique, ENSMP, Centre de Mathématiques Appliquées, France. URL: http://pastel.paristech.org/1989/01/These_BACH-Thanh-Le.pdf.

BACHIMONT B., «Interfaces et outils d'analyse et d'indexation», Atelier INA-Recherche, n°4, séance du 21 juin 1999, compte rendu. URL: http://www.ina.fr/inatheque/activites/ateliers /atelier4/A4_19990621.fr.html

BAGET J-F. & Al., «Les langages du Web sémantique». In *Revue Information, Interaction, Intelligence (I3)*, Hors série 2004 «Web Sémantique». URL: http://www.revue-i3.org/hors_serie /annee2004/revue_i3_hs2004_01_02.pdf

BERNERS-LEE T., HENDLER J., LASSILE O., «The Semantic Web», Scientific America, mai 2001.

BLANDIN B., «Technique, normes et standards», EDUCNET, 2003. URL: http://www.aix-mrs.iufm.fr/formations/tice/cd_tice/educnet/www.educnet.educatio n.fr /tech/default.htm_

BROUILETTE C., «Vers une définition de la lecture professionnelle», Périodique électronique étudiant *Cursus*, vol. 1 n°2, 1996. URL: http://www.ebsi.umontreal.ca/cursus/ vol1no2/brouillette.html

CRAMPES M., RANWEZ S., PLANTIE S., VAUDRY C., «Qualités d'une indexation portée par XML et une ontologie au regard d'un standard» In *XML et éducation*, Hors série 2003 STE.

GRUBER T., «A Translation Approach to Portable Ontology Specifications », *Knowledge Acquisition*, 5, 1993.

GUARINO N., «Understanding, Building and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development"», *International Journal of Human and Computer Studies*, 1997.

GUINCHAT C., SKOURI Y., *Guide pratique des techniques documentaires*, Edition Edicef, Vanves, 1996.

HENSENS H., «L'analyse documentaire, qu'est-ce que c'est?», *GESIST*, 2002. URL: http://www.mpl.ird.fr/documentation/indexation/index.htm

LAUBLET P., CHARLET J., REYNAUD C., «Introduction au Web sémantique», *Revue I3 Information-Interaction-Intelligence*, Numéro Hors-série Web sémantique, 2004. URL: http://www.revue-i3.org/hors_serie/annee2004/ revue_i3_hs2004_01_01.pdf

LE COADIC Y.F., *Usages et usagers de l'information*, collection Information Documentation 128, Edition Armand Colin, Paris, 2004.

MORIZIO C., *La recherche d'information*, collection Information Documentation 128, Edition Armand Colin, Paris, 2004.

RANWEZ S., *Composition Automatique de Documents Hypermédias Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur*, Thèse de doctorat, 2000, Université de Montpellier II, Montpellier, France. URL: http://www.lgi2p.ema.fr/Local/lgi2p/biblio /these_ranwez.pdf

SOWA J.F., «Ontology, Metadata and Semiotics», *Conceptual Structure: Logical, Linguistic, and Computational Issues*, p. 55-81, B. Ganter & G.W. Mineau, Berlin, 2000.

TA T.A., Web sémantique et réseaux sociaux- Construction d'une mémoire collective par recommandation mutuelle et (re-)présentations. Thèse de doctorat, 2005, EDITE-ENST, Paris, France. URL: http://pastel.paristech.org/1312/01/these-ta_finale.pdf

VERLAET L., Modèle communicationnel de balisage générique pour la sémantisation des informations: le cas d'une revue scientifique, thèse de doctorat en sciences de l'information et de la communication, CERIC, Montpellier, 2008.

VILLANOVA-OLIVER M., Adaptabilité dans les systèmes d'information sur le Web: Modélisation et mise en œuvre de l'accès progressif, Thèse de doctorat, 2002, Institut National Polytechnique de Grenoble, France. URL: http://www-lsr.imag.fr/Les.Personnes/Marlene.Villanova/THESE /TheseMVO.pdf