## Web sémantique et recherche d'information : où en est-on ?

Retour au sommaire de la lettre

*Le Web sémantique est dans toutes les bouches sans parfois que l'on sache réellement de quoi il s'agit. Dans cet article, nous tentons de faire le point sur sa définition, ses principes, ce qui a été réalisé aujourd'hui mais également ce qui reste à faire en termes de recherche d'information sur le Web dans ce domaine.*

L'environnement du web sémantique a beaucoup évolué depuis les premières directives du W3C en 2001. Essayons de retracer les différentes étapes qui ont mené à la conceptualisation d'un web sémantique basé sur les usages ; et surtout… centré sur l'usager. Cet article se situe volontairement dans une perspective des avancées qui concernent la recherche d'information.

### *Origine du web sémantique*

Le projet "semantic web" a vu le jour grâce au Consortium W3C à l'initiative de Tim Berners Lee. Le web sémantique, même si son application telle que ce dernier l'avait conçue en 2001 semble encore utopique aux yeux de bien des spécialistes, est cependant à l'origine de nombreux progrès. Son apport dans la création d'outils de recherche bien plus performants n'est plus à mettre en cause.

### *Tim Berners Lee*

Cet universitaire britannique est aujourd'hui titulaire de la chaire "*3Com Founder*" du Laboratoire d'Informatique et d'Intelligence Artificielle (CSAIL) au "*Massachusetts Institute of Technology*" (MIT). Il s'agit d'une organisation qui a pour mission de développer le web, de façon à ce que le gisement d'informations disponibles sur internet puisse être accessible au plus grand nombre. L'objectif premier était donc la création d'un système global de partage de l'information en réseau… mondial ! Une paille...

Depuis la conception de la notion de web sémantique, le W3C s'est attelé à un énorme travail dans le développement de normes, standards et définitions dans ce domaine.

### *Principes du web sémantique*

Le stockage des informations se fait sur la base de thesaurus ou ontologies et de façon intelligente. Ainsi des balises servent à tagger un mot et son domaine de rattachement, c'est-à-dire selon une logique de prédicats (relation « sujet », « objet », « prédicat ») qui fait appel à des définitions de type « *IS a* ». Ainsi la balise <peugeot> est rattachée au concept <voiture> dans la mesure où une *<peugeot> IS a (est une) <voiture>*, concept lui-même rattaché à la super classe de <véhicule> (*<voiture> is a <véhicule>*). Le web sémantique utilise ainsi un langage dérivé de la structuration XML : le langage RDF (*Resource Definition Framework*).

### *Langage OWL et Ontologies*

En plus du langage RDF, le W3C a créé le "langage" OWL. Ce langage est une extension spécialisée de RDF, servant à la création d'ontologies. Ainsi, le web sémantique fonctionne grâce au protocole HTTP d'une part, aux URL ou pages web, d'autre part. Ces URL utilisent le langage XML qui lui-même fait appel à deux sous-langages ou structurations supplémentaires, à savoir RDF et OWL. Mais le web sémantique, aujourd'hui ce n'est plus seulement les normes du W3C. C'est aussi toutes les avancées accomplies en faveur du sens et qui utilisent la langue naturelle comme vecteur de communication entre les usagers. A ce titre on trouve des applications de traduction automatique, la mise à disposition de dictionnaires multilingues sur le web, ou encore la cartographie d'information, qui fait appel à des processus cognitifs plus inédits que la simple analyse linéaire de texte.

### *Ce qui a été réalisé*

Sur le plan des applications grand public de meta données structurées, on trouve la première version de RSS, la structuration des contenus de Wikipedia ainsi que par exemple, Mozilla Firefox dans sa gestion des bookmarks.

Parmi les avancées récentes du web sémantique sur le plan de la recherche d'information et qui touchent le grand public, ou trouve notamment les swickis ou moteurs de recherche paramétrables par leurs utilisateurs, comme Eurekster (voir lettre R&R juillet 2006), par exemple.

Une autre application du web sémantique concerne les agents intelligents. Les standards du W3C sont à la source du filtrage d'information et de la diffusion sélective d'information. Ainsi, lorsque vous définissez un profil sur un moteur de recherche comme Widepress ou tout simplement Google, pour recevoir des alertes en fonction de vos centres d'intérêt, vous utilisez l'une des fonctionnalités que le web sémantique met à votre disposition.

On trouve également le foisonnement des blogs et wikis que se sont rapidement approprié les usagers de la toile. Les fils de News en structuration RSS découlent aussi directement des avancées du web sémantique.

### *Les progrès accomplis*

La notion d'intelligence collective a également sa part dans la notion de web sémantique et tout particulièrement au travers des blogs et des wikis. Les interfaces centrées utilisateur se sont considérablement développées depuis six ans, et pas seulement à destination des utilisateurs individuels. Dans la droite ligne du développement du web 2.0 de nombreux sites de e-commerce, e-learning et e-administrations ont vu le jour depuis quelques années.

De plus en plus d'outils de recherche d'information intègrent des composants d'analyse ou d'interprétation du sens. Certains grands de la recherche d'information comme Google ont intégré à tout le moins des modules de correction orthographique. D'autres, comme le français Exalead sont allés plus loin et ont intégré des dictionnaires multilingues, des modules de morphologie, qui permettent de distinguer le singulier du pluriel ou encore les différentes formes conjuguées d'un verbe. D'autres comme le moteur français Mozbot, basé sur la technologie de Google, ont intégré les fonctionnalités d'analyse linguistique de Memodata, qui filtrent et rendent plus lisibles les résultats fournis par Google. D'autres encore ont intégré des modules de traduction automatique, comme Yahoo! ou Google.

Par ailleurs le tagging collaboratif, ou folskonomie a été à l'origine de la création de nouveaux moteurs de recherche depuis quelques années. On a ainsi vu naître des moteurs comme del.icio.us, qui ont joué un rôle important dans la dissémination du savoir au travers de réseaux sociaux.

### *Les progrès qu'il reste à faire*

Les résultats obtenus en termes de recherche d'information sont de plus en plus intéressants. Nous n'avons cependant pas encore eu l'occasion de trouver sur la toile des outils qui procèdent à une analyse sémantique des textes retournés par un moteur.

Il serait intéressant, en effet, de pouvoir disposer d'une réelle analyse des mots en contexte. Ainsi « louer un vélo » ne pourrait pas être confondu avec « louer dieu dans une église », car le contexte du verbe "louer" serait alors discriminant pour le choix du sens à privilégier, en cas de significations multiples. De tels outils existent et sont utilisées sur intranet par des entreprises du secteur privé, ou encore pour effectuer un tri *a posteriori* sur des résultats booléens retournés par un moteur de recherche. Mais aucun, à notre connaissance, n'est encore mis à la disposition du grand public sur internet. Parmi les fournisseurs de telles ressources on trouve des entreprises comme Lingway, Temis ou encore Memodata.

Par ailleurs, de nombreux progrès restent à accomplir dans le domaine du multimedia et notamment l'identification et le tagging des séquences audio et vidéo. Une grande contribution dans ce domaine est apportée par des projets européens comme PHAROS, qui allie plusieurs entités européennes travaillant de concert à la réalisation d'un moteur de recherche multimedia.

La vocation première du "web sémantique" est de mettre à disposition de l'utilisateur des données structurées sous forme de thesaurus ou ontologies, afin de lui permettre de mieux s'y retrouver parmi la masse d'informations disponibles sur la toile. L'autre objectif est que ces informations soient directement réutilisables par un ordinateur, quel que soit le contexte et au final, par les utilisateurs, afin de développer les recherches participatives en réseau.

### *En conclusion*

Toutes les avancées réalisées depuis six ans dans le cadre du W3C et à partir de la notion de web sémantique figurent à la base d'une navigation plus fluide et plus "intelligente" sur internet et aussi plus adaptée à l'utilisateur. L'un des défis du web sémantique est lié à son actuelle limitation. En effet, toutes les données disponibles sur internet ne sont pas forcément structurées sous format RDF. Par ailleurs, les utilisateurs qui mettent à disposition des contenus, *via* des blogs et des wikis fournissent des contenus en langue naturelle et donc, par nature, non structurés.

Le véritable langage du web sémantique est donc la langue naturelle, puisqu'Internet est un media de communication entre humains et à destination des humains. Ainsi, il y a fort à parier que le principal défi que devra relever le web sémantique dans les années à venir est la mise en place de modules réellement intelligents, capables de comprendre le langage des hommes dans un contexte multilingue, afin de réellement pouvoir répondre à leurs questions et partager non plus des données mais effectivement des idées.

Aujourd'hui, l'avenir du web sémantique est probablement dans la mise en place d'un web 2.0 qui répondra réellement aux attentes de l'utilisateur, en plaçant celui-ci au cœur de processus cognitifs effectivement participatifs, pour la création d'une intelligence collective vivante et partagée.

**Marianne Dabbadie**
*Directrice Innovation i-KM*
*Laboratoire GERIICO – EA 1060*