ONTOLOGIES ET WEB SÉMANTIQUE...
L'ère du documentaliste a-t-elle sonné ?

Sébastien DECLERCQ
Étudiant en Master Information, Communication & Document, Université Lille 3 - Charles de Gaulle

▪ Alors que le web sémantique apparaît de plus en plus fréquemment dans la littérature professionnelle, certaines notions restent dans le flou et le non-dit. L'objectif du présent article est de détailler ces éléments et de leur associer un sens pratique. Tout d'abord, il abordera les grandes thématiques associées au web sémantique, que cela soit au niveau de sa définition intrinsèque ou de la mise en évidence des ontologies. Ensuite, il présentera un nouveau langage informatique, le Web Ontology Language (OWL), permettant la création de celles-ci. Enfin, cet article mettra en évidence les différents apports que ces nouvelles notions peuvent entraîner dans le domaine infodocumentaire, de même que la place que pourra prendre le documentaliste dans la mise en place et la gestion de ces nouveaux éléments informationnels.

▪ Het semantisch web komt steeds meer ter sprake in de vakliteratuur. Bepaalde begrippen blijven eerder vaag en onverklaard. Bedoeling van het artikel is deze elementen te omkaderen en een link naar de praktijk te vinden. De auteur begint met de omgeving van het semantisch web te bekijken, een intrinsieke definitie uit te tekenen, het realiseren van een ontologie etc. Verder wordt dan de Web Ontology Language (OWL) als nieuwe informaticataal voorgesteld dit om de ontologie te kunnen ondersteunen. Uiteindelijk zal het artikel nagaan welke weerslag deze nieuwe terminologie heeft binnen het I&D-domein en de rol die de I&D-professional kan vervullen bij het toepassen en het beheer van deze nieuwe informatietool.

À l'aube de l'avènement de ce que de plus en plus de professionnels de l'information nomment le "web 3.0", ou "web sémantique", les informations techniques concernant sa mise en œuvre restent vagues. Il est en effet fréquemment cité dans la littérature, mais seuls des articles de théorisation sur ses tenants et aboutissants nous parviennent. Néanmoins, loin d'être uniquement un système de pensée, le web sémantique s'offre à nous sous de (plus en plus) nombreuses applications **pratiques**, même si peu de témoignages prennent en considération les éléments techniques mis à présent à la disposition des scientifiques de l'information.

Plusieurs raisons peuvent expliquer cela.

D'une part, la définition de cette nouvelle mouture du web est des plus variables et chacun l'adapte à sa propre vision ; même Tim Berners-Lee, le fondateur du *World Wide Web*, est revenu sur sa propre définition, introduisant du même coup la notion de "web des données".

D'autre part, la notion d'"ontologie", inhérente au développement du web sémantique, n'est que trop peu souvent décrite. Or sa définition - et sa compréhension ! - est vitale à la bonne assimilation de ce que ce sera le futur du web.

Définitions et recadrage des termes

Comme signalé en introduction, l'unique notion de "web sémantique" se décline en de nombreux termes qui, bien que fort proches dans les esprits, représentent différentes approches techniques d'un même sujet. Ainsi, "web 3.0", "web sémantique" et "web des données" ne désignent pas spécifiquement la même chose !

Tout d'abord, la notion de "web 3.0", souvent associée au web sémantique, est de plus en plus contestée : agissant à la manière des mises à jour d'un logiciel, le chiffre associé au web augmente en fonction de ses évolutions. Tabler sur le fait que la version 3.0 sera celle du sémantique est une erreur. Tim Berners-Lee, ayant lui-même fait cet amalgame, s'est rétracté.

Le nombre croissant d'opposants à l'association entre les notions de "web 3.0" et de "web sémantique" s'explique par le fait que le web sémantique, à l'heure actuelle, n'en est qu'à ses balbutiements et que d'autres évolutions façonnent elles aussi le web. Ainsi, le web 3.0 pourrait tout aussi bien être le "web géolocalisé en temps-réel"[1] ou encore le "web des objets"[2].

Face à des protestations de plus en plus nombreuses, Tim Berners-Lee propose alors une nouvelle notion, qui est une étape intermédiaire vers le web sémantique : le **web des données** (ou "*web of data*")[3]. Ce web des données a l'avantage d'offrir une vision à plus court terme de ce que sera le web sémantique : Tim Berners-Lee propose de quitter le web actuel, qui traite des **pages**, des **documents**, et de se tourner vers un web qui gère des **données.**

Cette transition qui peut paraître anecdotique est pourtant une véritable progression. Imaginez : alors que, actuellement, les moteurs de recherche nous transmettent comme résultats des pages où se trouvent les termes demandés indépendamment de leur localisation dans la page, le web des données soumettra des documents dans lesquels les termes recherchés seront liés.
La qualité des résultats sera donc d'autant plus grande.

Cela ne résout pas la question du web sémantique, toujours non défini ! Ne vous inquiétez pas, on y arrive.

Le web sémantique a pour objectif de "*permettre aux machines d'exploiter automatiquement les contenus de sources d'information accessibles par le Web pour réaliser des tâches variées*"[4]. La machine sera donc capable d'interpréter et d'exploiter le contenu des données web. Cette vision de la machine interprétant le contenu d'un texte est née d'un syllogisme simple :

Les documents web ont **un sens** *; les documents web sont* **compréhensibles par un ordinateur** *; un ordinateur est capable de* **comprendre le sens** *d'un document.*

Naturellement, ce syllogisme n'apparaît simple que si l'on possède des bases en web sémantique. Toutefois, sa compréhension peut vous mener à une véritable assimilation de ce qu'est réellement le web sémantique. Je vais donc tenter de vous l'expliciter.

La première affirmation est logique : chaque document, qu'il soit numérique ou non, textuel ou vidéo, *etc.*, possède un sens ; c'est sa raison d'être. Néanmoins, ce sens n'est pas accessible à tout un chacun : en fonction de la langue dans laquelle il est retranscrit par exemple, la compréhension du contenu du document est compromise. C'est actuellement le problème que rencontrent les ordinateurs : nous, nous parlons en langage naturel ; eux, ils ne comprennent que le binaire. Le sens du document leur est donc inaccessible.

Néanmoins, les ordinateurs **comprennent** les documents : ils en connaissent la nature (enregistrement sonore, photographie,...) et savent comment les traiter. Ceci n'implique toutefois pas qu'ils en comprennent le contenu ! Il en va de même pour nous : une monographie écrite en russe nous est incompréhensible si l'on n'en connaît pas la langue et l'alphabet, mais cela ne nous empêche pas d'en déterminer la nature, la langue et de lui trouver une place dans nos rayonnages.

C'est à ce dernier aspect que s'attaque le web sémantique : grâce aux efforts développés, les machines seront capables de passer outre les problèmes linguistiques et de **réellement** comprendre le sens du document. En exploitant les données fournies "au format sémantique", l'ordinateur assimile une chaîne de caractères comme "chat" à une entité numérique compréhensible pour lui.

Ce "format sémantique" - dont les implications purement informatiques ne seront pas traitées ici - consiste en la création de liens sémantiques puissants entre les termes. Ces liens pouvant être pluridirectionnels, ils tissent une véritable toile[5] sémantique, nommée **ontologie**.
Une ontologie est ainsi un "super-thésaurus" offrant la possibilité de multiplier les relations sémantiques à souhait.

De ce fait, on peut tenter de définir le web sémantique comme un web offrant la possibilité d'exploiter l'entièreté des relations sémantiques entre des termes usuels. Bien que souvent décrit comme l'évolution naturelle du web, le web 3.0 est une réelle innovation : le web sémantique devra, en théorie, permettre aux ordinateurs de comprendre le sens d'un document.

Afin d'arriver à ce résultat, le développement de nouveaux langages informatiques et de nouveaux outils est effectué depuis de nombreuses années, principalement par les chercheurs du W3C (*World Wide Web Consortium*), chargés de la standardisation du web.

Les premiers pas et la naissance du standard W3C OWL

Après de nombreuses années de recherches et de perfectionnements continus dans le domaine du traitement et de la gestion numérique, un standard du web a fait son apparition : OWL, ou Web Ontology Language. Basé sur le RDF (Resource Description Framework), le OWL est un langage web proche de la machine. De ce fait, sa compréhension et son utilisation, pour un humain est ardue, mais est aisée pour un ordinateur. Son exploitation est donc on ne peut plus facile.

Le OWL comprend trois sous-langages, à utiliser en fonction des besoins du codeur et du client :
- *OWL Lite* : conçu pour les utilisateurs "basiques". Il reprend les idées principales du OWL mais de façon simplifiée pour être compréhensible par l'humain.
- *OWL DL* : ou Description Logic. Ce dernier est plus poussé que le Lite, mais n'offre pas toutes les possibilités du OWL.

- *OWL Full :* le OWL en lui-même.

Le OWL est basé sur un vocabulaire XML[6]. Son utilisation est donc proche d'un langage XML classique. De ce fait, la structure du document reste compréhensible pour la plupart des utilisateurs fréquents des langages web, même si le OWL diffère par sa complexité.
Il faut toutefois noter que, contrairement au XML, le OWL ne demande pas une structure particulière pour la rédaction des éléments : que la définition de la métadonnée "Class", qui définit le type de document, se trouve en tête ou en fin de document ne change rien pour la machine. Le W3C suggère néanmoins un système de rédaction standardisé afin que chacun puisse comprendre plus facilement le OWL développé par un autre codeur.

Le OWL a été promu au rang de standard du web en 2004 par le W3C et sa dernière mise à jour (OWL 2.0) date de 2009. Il s'agit donc d'un langage vivant, qui fera bientôt partie de nos standards en documentation.

Que permet le OWL ?

Le OWL est, comme expliqué précédemment, un langage du web. Ce standard permet la rédaction de pages web complètes, ce qui permettra une utilisation plus en profondeur des termes repris dans le document. La puissance du OWL réside dans sa possibilité de créer le nombre de métadonnées souhaitées, passant d'une simple Class - proche des "termes génériques" de nos thésaurus - à une métadonnée plus complexe comme "Property:a_pour_fille".

Ceci peut paraître anodin, mais il n'en est rien : de telles métadonnées permettent des liens plus en profondeur et plus précis entre les termes d'un document. Le OWL offre donc la possibilité de faire transparaître sur le web les relations sémantiques entre différents termes.

Les métadonnées principales sont les "Class" qui, combinées à la métadonnée "SubClassOf", permettent de créer des relations hiérarchiques entre les termes.

On retrouve également des métadonnées offrant de nombreuses possibilités de relations : ce sont les métadonnées "Property". Ces dernières permettent de démultiplier les relations sémantiques, car elles sont multipliables à l'infini, comme les métadonnées du XML. Ainsi, selon les documents qu'il faut traiter, une propriété "a_eu_lieu_en" peut être créée, tout comme "est_née_en", etc.

De ce fait, grâce à ces métadonnées, un terme est renseigné de manière multiple, en augmentant de ce fait sa **sémantisation**.

Afin d'illustrer le mode de fonctionnement d'une ontologie, et donc du OWL, on peut considérer la figure 1 : on signale à l'ordinateur que Paul a Julie pour fille. Julie, quant à elle, est née en 1989, année de la chute du mur de Berlin. Avec ces simples informations, la machine assimile que, de fait, Paul était vivant en 1989 et a donc connu la chute du mur, ceci **sans que ces affirmations soient renseignées**. La machine comprend donc les notions transmises[7].

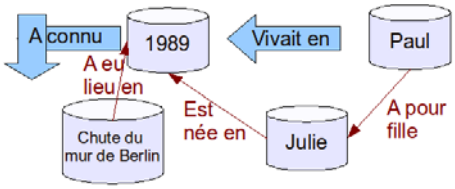

Fig 1 : Schéma du mode de fonctionnement d'une ontologie.

Le OWL est un système puissant d'exploitation des ontologies, qui permet justement la création déductive de ces liens entre les termes. Qui plus est, le OWL offre l'avantage d'automatiser les associations entre différentes ontologies : la modélisation d'une nouvelle base de connaissance n'est donc plus un acte isolé, mais bel et bien une action partagée.

Toujours dans cette optique, des moteurs de recherche d'ontologies ont vu le jour, avec en tête de liste le moteur *Swoogle*[8] qui offre accès à plus de 10.000 ontologies.

Le futur du web sera-t-il sémantique ?

En découvrant les apports phénoménaux qu'apportent les ontologies au web et aux recherches documentaires ainsi que la "facilité" avec laquelle le OWL peut les intégrer, on ne peut qu'espérer que le web 3.0 sera sémantique.

Néanmoins, les difficultés liées à la non-spécialisation de la majeure partie des concepteurs de sites web en matière de linguistique et de sémantique sont telles qu'on ne peut envisager une réelle révolution dans la conception des pages web, sans implication des scientifiques de l'information.

Ceux-ci ayant déjà une base acquise en gestion de (méta)données et en automatisation d'outils descriptifs, ils sont les plus à même de servir

d'intermédiaires entre linguistes et informaticiens.

Dès lors, la place des documentalistes dans la gestion de l'information numérique devient prédominante : avec leur expertise dans la gestion documentaire, ils sont aptes à aider à la mise en application d'un système sémantique. Leur présence est plus que nécessaire au bon déroulement d'un projet de cette envergure.

Par ailleurs, avec la démultiplication des outils sémantiques, la recherche sur le web s'en trouvera certes plus performante, mais parfois bien plus complexe. Les habilités des documentalistes seront mises au service de l'institution, afin de retrouver l'information demandée et ce sans que le traditionnel "*tout le monde sait chercher sur Google*" ne vienne interférer dans nos démarches informationnelles : l'exploitation des ontologies nous sera attribuée.

Grâce au standard OWL, ces ontologies sont devenues plus qu'envisageables ; elles font partie intégrante de la réalité actuelle des TIC[9]. Toutefois, leur intégration sera lente et le web 2.0, ou collaboratif, a encore de beaux jours devant lui.

La transition du web 2.0 au web sémantique s'effectuera, par ailleurs, de manière assez douce : tout comme le passage du "web 1.0" au web 2.0, les utilisateurs ne verront pas une réelle différence avant que la "naissance" du nouveau web ne fasse l'objet d'un battage médiatique.

Les initiatives sémantiques fleurissent actuellement à travers la toile, que ce soit avec des moteurs de recherche comme ***owl:SameAs***[10] ou avec l'usage du RDF par l'agence de presse **Reuters**[11], mais il nous faut toutefois garder un œil critique sur ces solutions : vu l'intérêt grandissant pour le secteur sémantique, de nombreux centres se présentent comme des pionniers en la matière en profitant de la niche actuelle, mais il n'en est rien. De même, ces outils pseudo-sémantiques sont avant tout des outils du web des données : ils font donc bel et bien un pas vers le sémantique, mais n'en sont pas réellement.
Un regard critique sur ces technologies et outils est donc nécessaire.

Ainsi, oui, le web sera sémantique, mais il est encore loin de l'être...

Sébastien Declercq
Rue des Salanganes, 36
1428 Lillois
declercq_sebastien@yahoo.fr
http://www.tic-au-tac.eu

Mai 2010

Références

Charlet, J. ; Laublet, P. ; Reynaud, C. *Présentation de l'Action Spécifique Web Sémantique* [en ligne].
< http://enssibal.enssib.fr/autres-sites/RTP/websemantique/presentation.html>

Hermann, A. ; Ducloy, J. Ontologie. *Ticri* [en ligne], 8 novembre 2009 (consulté le 1er mai 2010).
<http://maquettewicri.loria.fr/fr.ticri/index.php5?title=Ontologie>

Institut Porphyre. *TOTh 08 : Terminologie et ontologie : théories et applications* [en ligne]. Annecy, 5 et 6 juin 2008 (consulté le 1er mai 2010). <http://ontology.univ-savoie.fr/toth/TOTh2008_actes.pdf>

Johanna. Les ontologies pour structurer votre terminologie. *Neodoc - repenser la documentation technique* [en ligne]. 23 octobre 2008 (consulté le 1er mai 2010). <http://blog.neodoc.net/2008/10/une-ontologie-pour-structurer-votre.html>

Lacot, Xavier. *Introduction à OWL, un langage XML d'ontologies web* [en ligne]. 2005 (consulté le 1er mai 2010). <http://lacot.org/public/introduction_a_owl.pdf>

Lapique, Francis. Le langage d'ontologie web OWL. *Flash Informatique* [en ligne]. 8-24 octobre 2006 (consulté le 1er mai 2010), n°8/06, p. 3-8.
<http://ditwww.epfl.ch/SIC/SA/SPIP/Publications/IMG/pdf/8-6-page3.pdf>

Molette, Pierre. *Apport du web sémantique à la recherche d'information* [en ligne]. In Groupement Français de l'Industrie de l'Information. I-expo. Paris, mai 2008 (consulté le 1er mai 2010).
<http://www.slideshare.net/jdeyaref/lapport-du-web-smantique-la-recherche-dinformations>

OWL : naissance d'un nouvel outil sur le terrain du web sémantique. *Journal du Net* [en ligne]. 22 août 2003 (consulté le 1er mai 2010).
<http://www.journaldunet.com/solutions/0308/030822_owl.shtml>

OWL, RDF, N3 : les langages de description des connaissances [en ligne]. Centre de Recherche en Informatique de

Lens, 2008-2009 (consulté le 1er mai 2010). <http://www.cril.univ-artois.fr/~parrain/web/coursOWL.pdf>

W3C. *OWL Web Ontology Language Reference : W3C Recommendation* [en ligne]. (consulté le 1er mai 2010). <http://www.w3.org/TR/owl-ref/>

Notes

1 Tendance qui se généralise avec l'apparition dans les mœurs d'outils comme *Twitter* <http://www.twitter.com/> et *Foursquare* <http://foursquare.com/>

2 Web où les objets tels les frigos seraient connectés et pourraient communiquer avec leurs propriétaires via Internet.

3 La première utilisation grand public du terme "*web of data*" a eu lieu au TED, salon américain très réputé. La vidéo est accessible à cette adresse : <http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html> (consulté le 15 mai 2010).

4 Charlet, J. ; Laublet, P. ; Reynaud, C. *Présentation de l'Action Spécifique Web Sémantique* [en ligne]. <URL : <http://enssibal.enssib.fr/autres-sites/RTP/websemantique/presentation.html>

5 Notez que l'on retrouve bien ici la notion de "web".

6 Métalangage du web permettant une grande flexibilité, d'où son nom : eXtensible Markup Language. Il offre la possibilité de créer d'autres langages web, ayant une ossature proche, ce qui en facilite l'assimilation.

7 Il est à noter que l'on se trouve ici dans le cas du web des données et non du web sémantique : en traduisant des données, la machine détecte différentes implications à même de faire émerger de nouvelles données. La distinction entre le web sémantique et le web des données s'opère en ce sens : les fonctionnements sont similaires mais on ne peut affirmer ici que l'ordinateur **sait** ce que veut dire "chute du mur de Berlin".

8 *Swoogle* [en ligne]. <http://swoogle.umbc.edu/> (consulté le 15 mai 2010).

9 Technologies de l'Information et de la Communication.

10 *SameAs* [en ligne]. <http://sameas.org/> (consulté le 15 mai 2010).

11 Informations complémentaires, voir :
Reuters wants the world more tagged. *Read Write Web* [en ligne], 6 février 2008 (consulté le 21 mai 2010).
<http://www.readwriteweb.com/archives/reuters_calais.php>