

Introduction aux technologies du Web Sémantique

ROGER ROBERTS

Starting from the theoretical basis of general linguistics, the author proposes an overview of the technologies used to manage information and the documentary content on the World Wide Web. The main focus of his attention is on the semantic Web, outlining its technologies and potentiality.

Keywords: HTML – Protegé – RDF – Semantic Web – XML

La sémantique est une question de distance..., distances géographique, temporelle et culturelle. Elle permet de construire des ponts entre mon interlocuteur et moi lorsque nous sommes éloignés géographiquement, entre celui que je serai dans 10 ans et moi quand il s'agit d'accéder à nouveau à ce que j'ai conservé, entre un interlocuteur inconnu et moi qui n'évoluons pas dans le même environnement culturel... Il est donc tout à fait normal de retrouver cette composante majeure dans l'univers de la toile qui tisse des liens entre des individus qui échangent de l'information au sein d'un village virtuel!

Dans l'ensemble des langages véhiculés par les humains, il en est un auquel Jacques Derrida accorde la primauté sur les autres: la langue. Il désigne ce système métaphysique comme *logocentrisme*. Derrida élabore une différence proche de celle qui, chez Ferdinand de Saussure <fr.wikipedia.org/wiki/Ferdinand_de_Saussure>, donne sens aux éléments signifiants, sous forme de trace. La "trace" <[fr.wikipedia.org/w/index.php?title=Trace_\(philosophie\)&action=edit&redlink=1](http://fr.wikipedia.org/w/index.php?title=Trace_(philosophie)&action=edit&redlink=1)>, cependant, pleinement en eux-mêmes, il n'y a aucune vérité première, aucune différence transcendante à poursuivre.

Or, la *différance* (qui du fait de sa représentation graphique différente par la présence d'un "à") est précisément le mouvement "producteur" de ces différences: elle est le "processus" par lequel diffèrent les concepts.

Ces analyses offrent un intérêt majeur pour tous les chercheurs qui sont à l'étude de théories cherchant à établir des bases opérationnelles pour une sémiologie accessible à des machines. Dans l'univers audiovisuel, l'hypothèse de Jacques Derrida est pleine de sens dans la mesure où pour communiquer, pour indexer un objet ou pour échanger des objets, le recours à la langue et plus précisément à sa représentation textuelle, est le seul outil praticable. Une représentation particulièrement accessible pour des outils informatique!

Au début du XX^e siècle, les travaux menés dans le cadre de la linguistique ont ouvert des portes extrêmement précieuses à toutes les personnes essayant de comprendre l'univers particulièrement hermétique de la signification. On estime en Europe que c'est Ferdinand de Saussure qui a fondé la linguistique <fr.wikipedia.org/wiki/Linguistique> moderne et établi les bases de la sémiologie <fr.wikipedia.org/wiki/Sémiologie>.

Dans son Cours de linguistique générale <fr.wikipedia.org/w/index.php?title=Cours_de_linguistique_générale&action=edit&redlink=1> (publié *post-mortem* en 1916), c'est lui qui a défini entre autre la distinction entre langage, langue <fr.wikipedia.org/wiki/Langue> et parole <fr.wikipedia.org/wiki/Parole>, le caractère arbitraire du signe linguistique... <fr.wikipedia.org/wiki/Signe_linguistique>. L'idée fondamentale de Saussure est que le langage est un système clos de signes. Tout signe est défini par rapport aux autres, par pure différence (négativement), et non par ses caractéristiques propres ("positives"): c'est pourquoi de Saussure parle de «système»:

- **Sémantique**: l'étude du «sens» des symboles et expressions. Il s'agit de considérer le «sens» de façon opérationnelle
- **Langage**: un moyen de communication avec un ensemble de signes (vocaux, gestuels, graphiques, tactiles, olfactifs, etc.) doté d'une sémantique, et le plus souvent d'une syntaxe.
- **Langue**: un système de signes linguistiques, vocaux ou graphiques ou gestuels, qui permet la communication entre les individus doté d'une syntaxe précise et d'une grammaire.
- **Métadonnées**: une métadonnée est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique).

La théorie linguistique de Ferdinand de Saussure est nettement sémiotique <fr.wikipedia.org/wiki/Sémiotique> dans la mesure où elle interprète le langage comme un ensemble de signes: le linguiste distingue dans le signe <fr.wikipedia.org/wiki/Signe> deux éléments: le signifiant <fr.wikipedia.org/wiki/Signifiant> et le signifié <fr.wikipedia.org/wiki/Signifié>:

- **Le signifiant** désigne la représentation graphique d'un mot, d'une image, d'un son... Ce qui importe dans une représentation, ce sont les différences qui les distinguent les uns des autres. La valeur d'une représentation découle de ces différenciations. Chaque langage construit des lexiques à partir d'un nombre limité de caractères (phonèmes pour la langue), et d'une syntaxe qui définit l'ordre dans lequel ces caractères doivent être organisés. Pour l'univers informatique, le HTML est à tous points de vue, un signifiant d'une représentation en évidence!
- **Le signifié** désigne le concept, c'est-à-dire une représentation sémantique associée à un signifiant. Cette observation conduit Saussure à distinguer également **signification** et **valeur** puisque l'existence de langues différentes introduit néces-

sairement des significations et des valeurs différentes. Ainsi le signifié est un concept défini du fait de l'existence ou de l'absence dans une langue d'autres concepts qui lui sont opposables.

Pour certains psychanalistes, et notamment Jacques Lacan, tout l'intérêt de l'analyse de Saussure se situe dans cette ligne qui représente à la fois la différence mais également le lien entre le signifiant et le signifié. Pour le monde informatique, elle exprime la différence entre le HTML et les autres langages issus du XML et de ses dérivés!

Tous les travaux entrepris dans le cadre du «web sémantique» visent en fait à permettre de rapprocher au travers d'outils sophistiqués le signifiant du signifié afin de rendre les différences accessibles à des machines. Il ne faudrait pas oublier tous ceux qui ont contribué au premiers pas de l'informatique et rendu possible ce débat machine/sémantique. Comme Alan Turing, considéré comme un des pères fondateurs de l'informatique <fr.wikipedia.org/wiki/Informatique> moderne. Il est à l'origine de la formalisation des concepts d'algorithme <fr.wikipedia.org/wiki/Algorithmique> et de calculabilité <fr.wikipedia.org/wiki/Calculabilité>. La théorie des classes, qui permet de parler de collections d'objets qui ne sont pas nécessairement des ensembles chère à von Neumann ou encore Gödel avec le premier théorème d'incomplétude (dans n'importe quelle théorie récursivement axiomatisable, on peut construire un énoncé arithmétique qui ne peut être ni prouvé ni réfuté dans cette théorie).

La création et le développement rapide d'un outil de communication à vocation mondiale ne pouvaient que rebondir sur toutes les avancées proposées puisqu'en tant que outil gérant de l'information, elle sous-tend l'ensemble des activités liées à l'encodage d'information dans des bases de données et à des outils de recherche ouverts.

Le World Wide Web représente aujourd'hui une avancée technologique de première importance par l'influence qu'il exerce sur la majorité des aspects de nos économies et de nos sociétés. Cependant, en l'état actuel, il est peu satisfaisant en raison du nombre élevé d'activités souhaitées qui ne sont pas bien prises en charge par les outils automatiques. Ainsi, l'outil principal servant à la récupération d'informations est constitué de moteurs de recherche basés sur des mots clés. Ces outils, mêmes s'ils sont indispensables, présentent de fortes restrictions en termes de récupération, précision et contenu à partir de pages du web. Il est en effet assez aberrant de pratiquer au niveau de la recherche un outil basé sur le "protocole Gutenberg": en clair, le moteur de recherche analyse la syntaxe du signifiant en n'ayant aucun outil capable de resituer cette représentation dans un contexte de signifié. Pour la machine EBU signifie tout autant European Broadcast Union, European Boxing Union que European Bhuddist Union qui sont effectivement des organisations pleinement reconnues, mais que tout distingue dans leur raison sociale et morale ou sportive.

L'essentiel du contenu actuel est destiné à une interprétation par l'homme; la machine n'étant capable de le capturer et de le manipuler qu'au niveau syntaxique et donc de proposer un nombre incalculable de réponses à une requête non définie au niveau contextuel.

L'idée maîtresse du web sémantique est de rendre le contenu accessible et assimilable par la machine. Ceci ouvre la voie au développement d'outils sophistiqués susceptibles d'apporter un niveau bien supérieur de fonctionnalités pour assister les activités humaines sur le web.

Le web sémantique repose sur l'association des technologies suivantes:

- *les métadonnées explicites*: elles permettent aux pages web de comporter leur signification dans leurs balises.
- *les ontologies*: il s'agit des principes fondamentaux d'un domaine et leurs relations. *la logique*: elle permet de déduire des conclusions en combinant les métadonnées aux ontologies.

Brève introduction aux technologies du web sémantique:

HTML (*Hyper text Mark-up language*) est le langage à balises standard dans lequel sont écrites les pages web. Il repose sur une série de balises prédéfinies qui contrôlent l'édition d'une page (comme les caractères gras ou italiques d'une police, les listes numérotées ou non, les ruptures de ligne, etc.) Bref, dans le langage de Ferdinand de Saussure il s'agit bien de la représentation en évidence du signifiant.

Bien que le langage XML (*eXtensible Mark-up Language*) repose également sur des balises pour l'enrichissement du contenu web, ce langage permet aux utilisateurs de définir leurs propres balises. A cet égard, XML est donc un métalangage à balises indépendant du domaine (langage servant à définir un langage à balises). Les balises définies par l'utilisateur structurent la page qui ainsi devient assimilable.

Par contre, les balises XML ne décrivent pas la mise en forme des pages web. XML distingue donc le contenu de sa mise en forme, une caractéristique bien utile pour déterminer différentes présentations et vues sur la base des mêmes données et constitue la base pour une représentation du signifié.

XML fait en réalité partie d'une famille de langages destinés à diverses activités s'articulant autour du noyau du langage XML:

- DTD et XML Schema: deux langages permettant à l'utilisateur de définir son propre vocabulaire.
- XPath: langage fournissant l'accès à certaines parties des documents XML. Cet accès est une condition préalable et nécessaire pour adresser une requête de documents XML.

- XQuery: langage de requête destiné à XML.
- XSLT: langage déterminant les transformations de XML en HTML ou entre des représentations XML. On a ainsi XSLT comme outil essentiel pour la manipulation syntaxique des documents XML.

Dans l'élaboration du web sémantique, XML fournit la couche de base de la manipulation syntaxique. Bien que XML soit un langage universel pour définir des balises, il ne procure aucun moyen d'approcher la sémantique (le sens) des données. Il n'y a, par exemple, aucun signifié associée à l'encapsulation des balises. Il revient à chaque application d'interpréter l'emboîtement ce qui dans les faits, nécessite des outils supplémentaires pour pallier à ce manque.

Le RDF

RDF (*Resources Description Framework*) est un langage servant à décrire des ressources. Son élément constitutif de base est la formulation d'un triplet se composant d'une entité (appelée ressource en terminologie web), d'une propriété et d'une valeur (qui peut être une autre ressource). La formulation est essentiellement la définition d'un fait $P(a,b)$ où P est une propriété binaire, et (a,b) sont des ressources. Dans la perspective du web sémantique, RDF définit une couche située au-dessus de XML. De ce fait, RDF a été doté d'une syntaxe XML.

RDF est indépendant du domaine, en d'autres mots, il ne pose aucun présupposé quant à un domaine spécifique. C'est donc à l'utilisateur que revient le rôle de définir sa propre terminologie dans un langage schéma appelé RDF Schema (RDFS). Constitutivement, RDFS est un langage d'ontologies primitif (ou naturel) proposant les caractéristiques suivantes:

- Organisation des objets en classes (professeur, membre du personnel, étudiant, cours, cours pour étudiants) et propriétés binaires (enseigne, étudie, travaille).
- Sous-classes (tous les professeurs sont des membres du personnel) et relations des sous-propriétés (tout chef de département fait partie de ce département).
- Restrictions de domaine (seul le personnel académique peut enseigner) et d'étendue (une personne ne peut qu'enseigner) au niveau des propriétés.

RDF et RDFS fournissent les langages de base pour le web sémantique.

OWL

La puissance d'expression de RDF et de RDFS est volontairement très limitée: RDF est (en gros) limité à des attributs binaires et RDFS est (toujours en gros) limité aux

hiérarchies de sous-classes et de sous-propriétés, avec restrictions de domaine et d'étendue des propriétés.

Il existe cependant plusieurs cas particuliers d'utilisation du web sémantique qui nécessitent une plus grande expressivité. Ce type d'extensions comprend:

- la *disjonction* (par ex. une personne ne peut être à la fois professeur et membre du personnel administratif).
- les *combinaisons booléennes des classes* (par ex. le personnel est l'ensemble du corps académique, du personnel administratif et du personnel d'assistance technique).
- les *restrictions de cardinalité* (par ex. un service ne peut avoir qu'un seul chef).
- les *caractéristiques spéciales des propriétés* (par ex. "supérieur de" est transitif, "enseigne" et "reçoit des cours de" sont des propriétés inverses)
- l'*étendue locale des propriétés* ("*range*") définit la portée d'une propriété, par exemple "mange" pour toutes les classes. Dans certains cas, on peut souhaiter réduire la portée en fonction de la classe. Par exemple, les vaches ne mangent que de l'herbe tandis que d'autres animaux mangent également de la viande.

OWL a été élaboré comme nouveau langage d'ontologies standard pour le web. Il repose sur RDFS et tente de trouver un équilibre entre puissance d'expression et support logique efficace. La logique (le raisonnement) est un facteur important parce qu'elle permet de

- (a) vérifier la cohérence d'une ontologie et des connaissances,
- (b) vérifier la présence de relations non voulues entre classes
- (c) classer automatiquement les instances en classes.

Logique

La création formelle du langage OWL est une partie de la représentation et du raisonnement des connaissances appelée logique descriptive. Cette création est riche de promesses; l'approche est différente pour la représentation et le raisonnement sur la base de règles. Ses principaux avantages sont:

- Les moteurs de règles existent et sont très puissants.
- Les règles sont bien connues et s'utilisent en informatique générale. Elles sont faciles à apprendre.

Les systèmes de règles peuvent être envisagés comme une extension ou une alternative à OWL. La première idée est de poursuivre les recherches actuelles en visant à intégrer les règles et la logique descriptive tout en maintenant un support logique assez efficace. Une idée plus récente étudie l'utilisation de RDF/S conjointement aux règles comme base d'un autre langage ontologique web.

Outre les systèmes classiques à base de règles, il est intéressant d'analyser ceux capables de prendre en compte des conclusions contradictoires. Ces systèmes sont utiles pour la modélisation des données ayant hérités de défauts et des règles comportant des exceptions. Ils sont aussi très pratiques pour l'intégration des connaissances où des incohérences peuvent évidemment se produire lorsqu'on assemble des connaissances de sources différentes.

Pour plus de plaisir (une partie du texte qui précède a été rédigé avec la complicité active de ce site web et de Wikipedia <www.ics.forth.gr/isl/swprimer>).

L'idée principale est de rendre le sens accessible et manipulable par un moteur de recherche et une organisation qui prennent en compte la dimension culturelle de l'activité humaine grâce aux nouvelles technologies développées au sein du Web sémantique.

L'industrie audiovisuelle a produit ces dernières années beaucoup de standards afin de contribuer à aider les diffuseurs dans leur quête sémantique:

1. Solutions disponibles: AAF / MXF / SMPTE / P-Meta/ TV anytime ... (pragmatique mais ...)
2. Solutions pour la représentation de métadonnées, structures and synchronisation: SMIL (*Synchronized Multimedia Integration Language*) / ...
3. Modèles spécifiques: SMEF (BBC) / RAI / FARAO (ORF) / INA / IMMIX (Pays-Bas) / DR-M / MPEG-7/ MPEG 21 <www.enthrone.org>, ...
4. Encapsuleurs: MXF / METS / PK-ZIP / SPK-ZIP / PDF/A
5. Ontologie and Sémantique: FRBR (*Functionnal requirements for Bibliographic records* - IFLA)
6. Ontologie: Dublin Core Metadata Initiative (DCMI) / RDF / MARC (*MAchine-Readable Cataloguing*).

Une ontologie décrit les concepts et les rapports qui sont importants dans un domaine particulier, fournissant un vocabulaire pour ce domaine aussi bien que des spécifications automatisées de la signification des termes utilisés dans le vocabulaire. Les Ontologies traitent de taxonomies et de classifications, schémas de base de données, et de théories entièrement axiomatisées. Ces dernières années, des ontologies ont été introduites dans beaucoup de communautés scientifiques de manière à partager, réutiliser et traiter la connaissance d'un domaine spécifique. Les Ontologies sont devenues vitales pour beaucoup d'applications telles que des portails de la connaissance scientifique, des systèmes de gestion d'information et d'intégration, du commerce électronique, et de services de Web sémantique. L'absence d'une véritable solution générique au niveau de la norme MPEG 7 de l'ISO a constitué un handicap majeur!

Cerise sur le gâteau, les scientifiques de certaines universités ont développé des technologies nouvelles qui sont de véritables moteurs d'ontologie qui vont grandement simplifier la vie de tous ceux qui rêvent de produire des outils en OWL, RDF, ...

Un moteur d'ontologies est un processus applicatif qui possède une dimension sémantique (définition et mise en relation de concepts); une dimension logique (validation de la création de relations entre concepts ou déduction de relations non explicites entre concepts) et enfin une dimension usage avec la construction d'un vocabulaire («outil», «utilisateur», «traitement»), la construction d'une syntaxe (sujet, verbe, complément) et d'une grammaire: «l'utilisateur» «accorde» «l'outil» pour réaliser «un traitement».

La finalité du moteur d'ontologie est de:

- réduire la distance entre le langage de la machine (logique) et le langage de l'utilisateur (lisible) en intercalant entre l'une et l'autre un langage compréhensible par l'un et l'autre;
- réduire la distance entre des langages pratiqués dans différentes cultures (métier, région, temps, etc.) en intercalant entre l'une et l'autre un langage interprétable par l'un et l'autre;
- améliorer la qualité, l'efficacité et l'efficience des échanges entre les acteurs d'un projet via un socle sémantique commun;

Un moteur d'ontologies est un processus applicatif qui possède:

Une dimension sémantique:	Définir des concepts Mettre en relation des concepts
Une dimension logique:	Valider la création de relations entre concepts Déduire des relations non explicites entre concepts
Une dimension usage:	Construction d'un vocabulaire («outil», «utilisateur», «traitement») Construction d'une syntaxe (sujet, verbe, complément)
Construction d'une grammaire:	«l'utilisateur» «accorde» «l'outil» pour réaliser «un traitement».

Un exemple: «Protege»

«Protege» est une plate-forme ouverte développée par l'université de Stanford et qui fournit à une communauté d'utilisateur une série d'outils logiciels pour construire des modèles de domaine et des applications basées sur la connaissance des ontologies. En son sein, «Protege» met en application un ensemble riche de structures de «connaissance-modélisation et actions» qui soutiennent la création, la visualisation, et la manipulation des ontologies dans divers formats de représentation. «Protege» peut être

adapté aux besoins d'un client pour fournir l'appui logistique à la création de modèles de la connaissance de saisie de données. «Protege», basé sur Java, est extensible, et fournit un environnement prêt à l'emploi qui en fait une base flexible pour le prototypage et le développement d'applications rapides. De plus, «Protege» peut être utilisé en mode Plug-in avec une interface de programmation API pour construire des outils basés sur la connaissance et le développement d'applications.

La plate-forme Protégé propose deux manières principales de modéliser des ontologies:

- L'éditeur «Protege - Frames» permet à des utilisateurs d'établir et peupler des ontologies basées sur le protocole ouvert de connectivité de base de connaissance (OKBC). Dans ce modèle, une ontologie se compose d'un ensemble de classes organisées dans une hiérarchie pour représenter les concepts fondateurs d'un domaine, un ensemble de connecteurs associés aux classes pour décrire leurs propriétés et rapports, et un ensemble d'exemples de ces classes – différents exemplaires des concepts qui tiennent des valeurs spécifiques pour leurs propriétés.
- L'éditeur de «Protege-OWL» permet à des utilisateurs d'établir des ontologies pour le Web sémantique, en particulier dans la spécification du W3C (OWL).

La BBC a mis en application «Protege» pour contrôler une description RDF/OWL des contenus des nouveaux médias, pour éditer les sites Web, la TV interactive, et les plates-formes mobiles, etc. La BBC a créé une ontologie OWL décrivant le métamodèle de l'entreprise. Elle utilise «Protege» (avec quelques connexions faites sur demande) pour créer des schémas d'objet et des conceptions d'interaction en vue d'éditer ces objets et faciliter la saisie sur les systèmes de gestion de contenus.

L'aventure ontologique commence à l'adresse suivante: <protege.stanford.edu>.

Sitographie

Cover Pages, METS: <xml.coverpages.org/mets.html>.

Dublin Core: <dublincore.org/documents>.

Introduction au web sémantique: <www.ics.forth.gr/isl/swprimer>.

Metadata Principles and Practicalities, 2002, "The D-Lib Magazine", April 2002, Vol. 8 N. 4: <www.dlib.org/dlib/april02/weibel/04weibel.html>.

MPEG-7: <mpeg.tilab.comcse.it>.

RDF' by W3C: <www.w3.org/TR/rdf-primer>.

The Metadata Encoding and Transmission Standard, (METS), <www.loc.gov/standards/mets/>.

Un moteur d'ontologie: <protege.stanford.edu>.

Remerciements

Je voudrais remercier toutes les personnes qui ont contribué directement ou indirectement à cette présentation au Congrès de l'Association Italienne de Terminologie:

- SKEMA (UTC Compiègne), pour les contributions aux développements sur les Ontologies et le projet MediaMap
- PROSI* and MEMNON*, in particulier M. Guy Marechal et M. Michel Merten pour les développements d'AXIS
- EBU, en particulier M. Jean-Pierre Evain
TITAN, pour l'organisation des "European Media Wrapper Conference and Round Table"
- SBS SIEMENS*, en particulier M. John Jordan (past manager of the BBC SMEF project)
- "ISO", pour la contribution à la normalisation de l'OAIS