Qu'est-ce que le web sémantique?

Jérôme Euzenat
INRIA Rhône-Alpes
Jerome.Euzenat@inrialpes.fr

Origines, problématique

Si le web actuel contient une quantité d'information formidable, il reste difficile à exploiter. Ainsi, la recherche d'un «livre sur Agatha Christie» n'est guère aisée à l'aide des moteurs de recherche: ils commencent par supprimer le mot clé «sur» comme peu discriminant et retournent nombre de pages consacrées aux ouvrages de la romancière. Si l'on désire que les machines nous aident à trouver l'information, il nous faut les aider un peu plus en la leur explicitant.

Le web est constitué par un ensemble de documents, principalement textuels, formatés dans un langage particulier (HTML) permettant d'exprimer des liens entre un objet dans le document source (l'ancre) et un objet du document cible. Ce web est exploité par des dispositifs logiciels (navigateurs ou robots de recherche) qui traversent ces liens lorsqu'ils les rencontrent (ou lorsque l'utilisateur clique sur une ancre). Le travail d'exploitation de ce web est principalement dévolu aux utilisateurs humain qui doivent analyser le contenu des pages pour déterminer sur quel lien cliquer. Des dispositifs logiciels peuvent l'aider en analysant ce contenu, mais comme on l'a vu leur aide, bien que remarquable, reste limitée car le contenu des documents du web s'adresse aux utilisateurs humains.

En première approximation, le but du web sémantique est de développer un web dont le contenu s'adresse, au moins pour partie, aux machines, afin qu'elles puissent aider les utilisateurs humains. Si l'on cherche à préciser, un tel web doit doter ses ressources (documents, service…) d'annotations dont le but n'est pas d'assurer l'affichage des documents mais l'appréhension de son contenu par divers outils logiciels. Le web sémantique doit donc être une infrastructure juxtaposant au web actuel des documents structurés par des langages pour exprimer la connaissance, pour décrire les relations entre connaissance, pour décrire les conditions d'utilisation, pour décrire les garanties et les modes de payement et de dispositifs permettant de trouver les ressources.
Un tel web doit, de plus, hériter des particularités du web: ouvert, interopérable et distribué.

Acteurs

Les idées du web sémantique (un terme dû à Tim Berners-Lee, l'initiateur du web actuel) ont commencé à fleurir avec le web mais ont été expérimentées au milieu des années 1990 à une échelle modérée dans deux initiatives: SHOE à l'université du Maryland et Ontobroker à l'université de Karlsruhe. Les développements actuels ont réellement commencé avec l'initiative DARPA Agent Markup Language (DAML) aux États-Unis d'Amérique. Elle avait pour but de produire un langage d'expression de connaissance successeur de SHOE. Peu après, les chercheurs européens ont créé le réseau thématique Ontoweb afin de fédérer la recherche dans ce domaine. Les deux groupes ont conçu en commun le langage DAML+OIL précurseur des langages actuels. Indépendamment, l'ISO a normalisé le langage des cartes topiques ("Topic maps") qui a su rassembler autour de lui de nombreux développeurs issus du monde documentaire.

Finalement, après un long temps de maturation, le W3C (World wide Web Consortium) a créé son activité sur le sujet à l'automne 2001. Cette activité a eu pour tâche de relancer les développements sur le langage RDF et ses extensions, rassemblant dans un cadre formalisé les chercheurs américains, européens et asiatiques. Les activités pour la construction du web sémantique sont maintenant très actives dans le monde entier. Elles sont centrales aux «technologies de la connaissance» promues par le 6e programme cadre de la recherche et technologie européenne.

L'un des bons côtés du web sémantique est qu'il fait coopérer étroitement des acteurs d'origines très différentes : depuis les protocoles de communication jusqu'aux relations entre ordinateur et sens. On peut noter une double partition dans les développements du web sémantique :

— D'une part entre théoriciens, qui désirent que les langages soient définis correctement avant toute utilisation, et praticiens, qui désirent que les programme fonctionnent rapidement avant de figer l'architecture. L'un des intérêts de l'activité Web sémantique du W3C est qu'elle fait dialoguer les deux camps au profit, on peut l'espérer, de tous.

— D'autre part, entre industriels, qui estiment qu'une telle entreprise ne pourra fonctionner que si elle est rentable pour des applications importantes, et militants, qui rappellent que le web n'a dû son existence qu'à des initiatives, nombreuses, d'amateurs. Le web actuel profite à la fois aux industriels et aux citoyens ; on ne peut qu'espérer qu'il en sera ainsi du web sémantique.

Enjeux, le web sémantique pour quels usages ?

Comme vu plus haut, le premier scénario du web sémantique est la recherche d'information. Il nécessite donc l'expression de «métadonnées» : des données à propos des documents disponibles. Il est donc normal que le premier maillon du web sémantique soit le langage RDF destiné à décrire les ressources. Au-delà de la ressource elle-même, son format, son objet, son auteur, on voudra aussi décrire le coût de son accès et la confiance que l'on voudra y accorder.

Les fonctions d'apprentissage à distance (eLearning) sont aussi très demandeuses de métadonnées décrivant les objets éducationnels. Mais elles ont encore plus besoin d'accéder au contenu de ces objets pour évaluer leur adaptation à l'utilisateur, pour pouvoir proposer une évaluation de l'apprenant aux objets offerts et pour assurer l'interopérabilité entre les ressources disponibles.

Au-delà de la description des ressources du web, le web sémantique veut décrire toutes les ressources possibles. C'est pourquoi il s'applique à des applications comme le commerce électronique où l'on veut décrire les produits au catalogue d'un marchand. À l'instar de l'autobiographie d'Agatha Christie, on voudra retrouver les produits sur des critères précis voire rassembler un ensemble de produits compatibles (un appareil photo et ses objectifs), des produits d'accompagnement (le port et l'assurance) tout en optimisant un critère (le prix ou la disponibilité). L'une des difficultés à surmonter est de faire interopérer des sources d'informations décrites de manières différentes. Les langages du web sémantique devront donc permettre d'exprimer et d'exploiter des relations entre ces descriptions.

En plus des descriptions que l'on veut interopérables, l'un des domaines prometteurs est celui des services web décrits à l'aide de ces techniques. Cela est en effet nécessaire si l'on désire composer des services (c'est-à-dire qu'une sortie de l'un soit l'entrée d'un autre service). Il faut alors disposer d'une description compréhensible des prérequis à l'utilisation d'un service et des fournitures auxquelles le fournisseur s'engage.

Sur un plan plus personnel, on peut imaginer «□sémantiser□ la gestion des informations personnelles□ agenda, carnet d'adresses, mais surtout un ensemble de préférences afin de trouver plus facilement un rendez-vous, de commander le moyen de transport et d'hébergement le plus adapté. Par exemple, le scenario de l'agence de voyage met en œuvre un agent logiciel capable de planifier un voyage complexe impliquant plusieurs lieux, plusieurs moyens de transport, l'instcription à des activités tout en résolvant des contraintes hétérogènes (dormir dans un hôtel convenable, minimiser le coût d'ensemble du voyage, diminuer le temps de correspondance, utiliser les transporteurs favoris et trouver des repas végétariens).

Ce dernier scénario assigne au web sémantique les tâches qu'accomplirait un très bon assistant d'une manière plus systématique et rationnelle (et certainement moins flexible et agréable). Le web sémantique ne doit toutefois pas être confondu avec la thèse de l'intelligence artificielle forte pour trois raisons principales□

— Il n'a aucune visée anthropomorphique□ son but est de compléter l'être humain là où il n'est pas le plus efficace (comme la recherche rapide dans une grande quantité d'information, le travail ininterrompu, etc.)
— Il retient beaucoup de leçons du web quant à sa taille, son manque de cohérence et son étendue□ le web sémantique doit donc passer à l'échelle, être robuste et décentralisé□
— Il doit être adapté au contexte dans lequel il évolue.

Comme on peut le voir, les applications potentielles du web sémantique ne sont limitées que par notre imagination. Les scenarii les plus élaborés n'impliquent pas uniquement la recherche mais l'évaluation, la sélection, la composition de ressources, l'appel à des services, la mise en corespondance de profils et de descriptions de ressources, le réordonancement automatique et la recomposition de services lors de l'arrivée d'un événement. Un scénario du type agence de voyage requiert quatre ingrédients principaux□

— un web de ressources bien annotées□ disponibilité des horaires et tarifs de transports, description des lieux d'hébergement et de restauration…
— des connaissances générales (ou «□ntologies□)□ un bus est un moyen de transport, la Sardaigne fait partie de l'Italie…
— une description des préférences de l'utilisateur□ régime alimentaire, programmes de voyageur fréquent, agenda, besoin d'hébergement…
— des capacités d'assembler ces ressources pour accomplir la tâche spécifiée□ inférence taxonomique, raisonnement temporel, propagation de confiance…

Nous discutons ci-après de ces différentes resources.

Les outils du web sémantique

Comme on a pu le voir avec l'exemple de l'autobiographie d'Agatha Christie, l'annotation des ressources à l'aide de simples mots-clé, voire des catégories, n'est pas suffisant. Il faut être capable d'exprimer l'information relationnelle□ qu'un objet «□ivre□ peut avoir un «□auteur□ qui est une «□personne□ et un «□sujet□, ou qu'une «□utobiographie□ est une «□iographie□ dont l' «□auteur□ est le «□sujet□. Il est donc naturel que le premier langage pour le web sémantique, RDF ("Resource Description Framework") mette l'accent sur les relations.

RDF est un langage, recommandé par le W3C, fondé sur les notions de ressources et de relations entre ressources. Un triplet <*s*, *p*, *o*> exprime une relation *p* entre un sujet *s* et un

objet *o*. Les relations sont identifiées par des URI (“Uniform resource identifiers” dont l’exemple le plus connu est celui des URL, qui constituent les « adresses » des pages du web). Les ressources peuvent être identifiées par des URI ou anonymes et les objets peuvent être des littéraux (comme une chaîne de caractère ou un entier). Un document RDF constitue donc initialement un graphe étiqueté sur ses arêtes et ses sommets. Les ressources peuvent de plus être typées en utilisant la relation “type”. Depuis peu, le langage RDF est doté d’une sémantique en théorie des modèles précisant comment ces graphes doivent être interprétés et donc comment en tirer des conséquences.
Un langage alternatif est celui des cartes topiques proposé par l’ISO. Il est basé sur trois types d’entités : les thèmes (ou “topics”), les associations et les portées. Les thèmes correspondent aux ressources, les associations aux relations et les portées sont des ensembles de thèmes qui permettent de circonscrire le contexte dans lequel une assertion est valide. À ceci on peut ajouter les noms permettant d’identifier les thèmes (l’approche est donc multilingue d’emblée) et les occurrences agissant comme des médiateurs entre les cartes topiques et le monde extérieur (un type de donnée particulier ou un objet identifiable par un URI). Les cartes topiques sont un modèle très versatile et peuvent donc s’adapter à différentes situations. Cependant, l’absence d’une sémantique claire du formalisme rend difficile son appréhension.

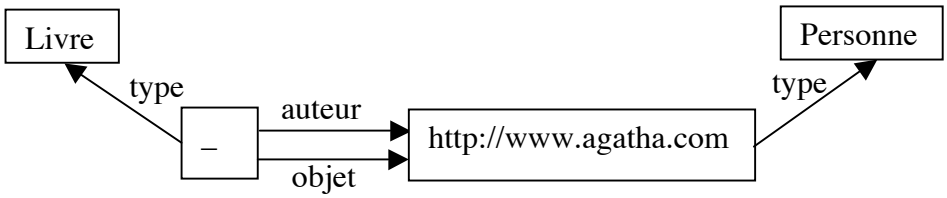

Figure 1 : Ce graphe RDF s’interprète comme « il existe un livre dont l’auteur et l’objet sont la personne identifiée par http://www.agatha.com ».

On l’a vu plus haut, il est nécessaire de décrire le vocabulaire dans lequel les ressources sont décrites. On parle de connaissances générales ou d’« ontologies ». Il existe une tradition de développement de langages d’expression d’ontologies en représentation de connaissance. Ils décrivent en général un ensemble de catégories organisées dans une structure permettant de relier une catégorie à ses catégories plus générales.
À partir de RDF, le langage RDF-Schéma (RDFS) a été développé. Il permet d’enrichir RDF de quelques types de ressources prédéfinies (Resource, Class et Relation) et de quelques relations prédéfinies (type —déjà évoqué—, subClassOf, range et domain). Ces primitives constituent la base de tout langage d’ontologie, permettant de signifier l’appartenance d’un objet à une catégorie, de déclarer la relation de généralisation entre catégories et de typer des objets reliés par une relation. RDFS est cependant un langage difficile à interpréter car réflexif : tout y est ressource ; Relation, Resource et Class sont des Class ; type et subClassOf sont des relations…
Indépendamment, différents langages de description d’ontologie ont été développés pour le web. Le projet européen ontoknowledge avait développé le langage OIL comme une extension d’XML-Schéma (proche de RDFS) offrant des primitives inspirées des logiques de descriptions. DAML a pour sa part proposé le langage DAML-ONT, fondé sur RDF, et plus proche des langages objets. Ces deux langages ont été fusionnés en un langage connu sous le nom de DAML+OIL qui a servi de base à l’élaboration, par le W3C, du langage d’ontologie pour le web OWL. Pour simplifier les choses, OWL est disponible en trois parfums : OWL-Lite, OWL-DL et OWL-Full. En schématisant, OWL-Lite est un langage dans lequel on peut décider efficacement de la relation entre deux classes, OWL-DL est un langage beaucoup plus

complexe, mais reste décidable alors qu'OWL-Full est la fusion entre OWL-DL et RDFS (c'est-à-dire qu'il abolit la distinction forcée entre classes et ressources). On peut appréhender le langage OWL en observant la définition d'un livre et d'une biographie présentée ici. Disons qu'en plus des primitives de RDFS, OWL-DL permet de contraindre plus précisément la description des classes (en les décrivant comme union, intersection, complémentaire d'autres descriptions ou comme l'ensemble d'un certain nombre d'individus), des domaines de relations (en spécifiant le type de toutes leurs valeurs, ou d'un certain nombre de leurs valeurs) ou des relations (en les déclarant transitives, symétriques ou en spécifiant leur inverse). Par ailleurs, il est possible de déclarer que deux classes ou ressources sont équivalentes ou, au contraire, différentes.
Cette fois encore, OWL est doté d'une sémantique en théorie des modèles permettant de spécifier tout ce qui est déductible d'un ensemble d'assertions de OWL. Le langage devrait être recommandé très bientôt et la phase suivante sera l'implémentation de moteur d'inférence permettant d'offrir les capacités déductives aux applications présentées auparavant.

```
<owl:Class rdf:ID="Biographie">
      <owl:intersectionOf>
            <owl:Class rdf:resource="Livre"/>
            <owl:Restriction>
                  <owl:onProperty rdf:resource="object"/>
                  <owl:allValuesFrom rdf:resource="Personne"/>
            </owl:Restriction>
      </owl:intersectionOf>
</owl:Class>

<owl:Class rdf:ID="Autobiograpie">
      <rdfs:subClassOf rdf:resource="Biographie"/>
</owl:Class>
```

Figure 2 : La première expression OWL s'interprète comme « la classe Biographie est l'intersection de la classe Livre et des objets dont la propriété objet prend ses valeurs dans la classe Personne », la seconde signifie que « les Autobiographies sont des Biographies ».

L'étape suivante conduirait à exprimer des règles dans le web sémantique afin de permettre l'expression des mécanismes de fonctionnement, des contraintes ou d'introduire un aspect opportuniste dans le web sémantique. Il y a suffisamment de volonté sur ce thème pour que l'on anticipe un tel langage mais rien n'est pour l'instant particulièrement formalisé.

```
<r:Rule>
      <r:premisse>
            <Autobiograpie rdf:ID="?x">
                  <auteur rdf:resource="?y"/>
            </Autobiographie>
      </r:premisse>
      <r:conclusion>
            <Autobiograpie rdf:ID="?x">
                  <objet rdf:resource="?y"/>
            </Autobiographie>
      </r:conclusion>
</r:Rule>
```

Figure 3 : On peut facilement imaginer un langage de règles permettant d'écrire que « les Autobiographies ont pour objet leur auteur ».

Enfin, le dernier élément manquant est la possibilité d'exprimer les préférences des utilisateurs et la confiance que l'on peut associer aux ressources. En ce qui concerne les profils utilisateurs, le W3C dispose déjà d'un langage (CC/PP) et de nombreux acteurs estiment que RDF est une base suffisante pour cela. Mais concernant la notion de confiance, le travail ne fait que commencer et aura à faire ses preuves.

Perspectives

En somme, les travaux concernant l'infrastructure permettant de donner le jour à un web sémantique sont relativement bien avancés. On peut imaginer tester dans un futur proche des outils permettant d'extraire, de composer et de déduire les conséquences de ressources du web sémantique.
Ce qui reste la principale difficulté immédiate est la disponibilité de ressources annotées en nombre et qualité suffisante pour pouvoir être exploitées avec profit. C'est le contenu qui a fait la valeur du web, c'est encore lui qui fera celle du web sémantique. Il est donc nécessaire de disposer d'outils permettant de décrire aisément les ressources mises à disposition.
Une fois le contenu disponible, les applications prometteurses, et d'autres non encore imaginées, pourront se déployer et démontrer par l'usage l'intérêt du concept de web sémantique.

Pour en savoir plus

Sur les motivations du web sémantique□Tim Berners-Lee, James Hendler, Ora Lassila, The semantic web, *Scientific american* 284(5):35-43, 2001, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

Sur les technologies disponibles pour le démarrage du web sémantique□Dieter Fensel, James Hendler, Henry Lieberman, Wolfgang Wahlster (eds.), Spinning the semantic web, The MIT press, Cambridge (MA US), 2003

Recherches actuelles□Isabel Cruz, Stefan Decker, Jérôme Euzenat, Deborah McGuinness (eds.), The emerging semantic web: selected papers from the first Semantic web working symposium, IOS press, Amsterdam (NL), 2002, 300pp., http://www.inrialpes.fr/exmo/papers/emerging/

Recherches futures□Jérôme Euzenat (ed.), Research challenges and perspectives of the Semantic web, http://www.ercim.org/EU-NSF/semweb.html

www.w3.org/2001/sw/ L'activité web sémantique au W3C
www.ontoweb.org Le réseau thématique européen OntoWeb
www.daml.org Le programme DAML de la DARPA
www.lalic.paris4.sorbonne.fr/stic/ L'action spécifique sur le web sémantique du CNRS