Le Web sémantique

Construit sur la base de technologies existantes (protocole HTTP (*HyperText Transfert Protocol*), identifiant URI (*Uniform Resource Identifier*), qui désigne de manière unique un document sur le Web…), le Web sémantique, par une participation toujours plus active de la communauté des internautes, constitue une nouvelle étape dans la logique de partage qui caractérise le Web actuel.

par Alexandre BERTAILS*, Ivan HERMAN** et Sandro HAWKE***

Le Web tel que nous le connaissons aujourd'hui est encore conforme à la vision qu'en avait Tim Berners-Lee il y a quinze ans : il s'agit d'un Web de documents. Ceux-ci sont écrits en HTML (*Hypertext Markup Language*), identifiés de manière unique par des URLs (*Uniform Resource Locator*) et reliés entre eux par des liens hypertextes. L'utilisateur surfe manuellement de page en page et peut depuis quelques années interagir avec le Web grâce aux technologies du Web 2.0 (Ajax).

Cependant, l'information reste essentiellement textuelle et l'utilisateur ne voit que le sommet de l'iceberg : les données réelles, brutes et structurées, ne lui sont pas accessibles. Elles sont stockées, la plupart du temps, dans des bases de données et l'utilisateur n'en visualise que le rendu.

Or toute la valeur du Web est en réalité dans ces données ! Les exposer facilite la recherche de l'information ainsi que sa compréhension. L'étape suivante pour le Web est donc de pouvoir lier toutes ces données et de les combiner à loisir dans des applications composites (*mashups*). Le Web a besoin d'être équipé des technologies nécessaires à la création d'un Web de données (*Web of Data*).

Les technologies du Web sémantique complètent le Web actuel avec des outils sémantiques. Il ne s'agit donc pas de créer un nouveau Web ou un Web séparé de l'existant : ce Web de données repose entièrement sur les technologies et concepts qui ont fait le succès du Web tel que nous le connaissons aujourd'hui (voir la photo 1).

N.B : Dans la suite de cet article, nous ferons l'amalgame entre les termes URI et URL, bien qu'ils ne désignent pas tout à fait la même chose.

LE WEB DE DOCUMENTS

On trouve des données un peu partout : dans des documents XML, des feuilles de tableur, des fichiers textes plats et surtout dans des bases de données relationnelles. Comment et pourquoi y appliquer les concepts du Web ?

Le Web repose sur trois technologies fondamentales :

- Le langage HTML permet de décrire la structure d'une page Web ;
- Une URI désigne de manière unique un document sur le Web ;
- HTTP est un protocole décrivant les requêtes et réponses échangées entre deux machines (client/serveur).

Depuis sa création, le concept d'URI a été étendu de manière à pouvoir identifier autre chose que des pages Web, comme par exemple des objets ou tout concept abstrait. De même, le besoin de plus de structure dans la notion de document a conduit à la généralisation de HTML en XML *(Extensible Markup Language)*. Tout un ensemble de technologies a alors dû être spécifié pour interagir avec XML : espaces de noms, schémas, requêtage XQuery/XPath, DOM, etc.

* W3C, bertails@w3.org

** W3C, ivan@w3.org

*** W3C, sandro@w3.org

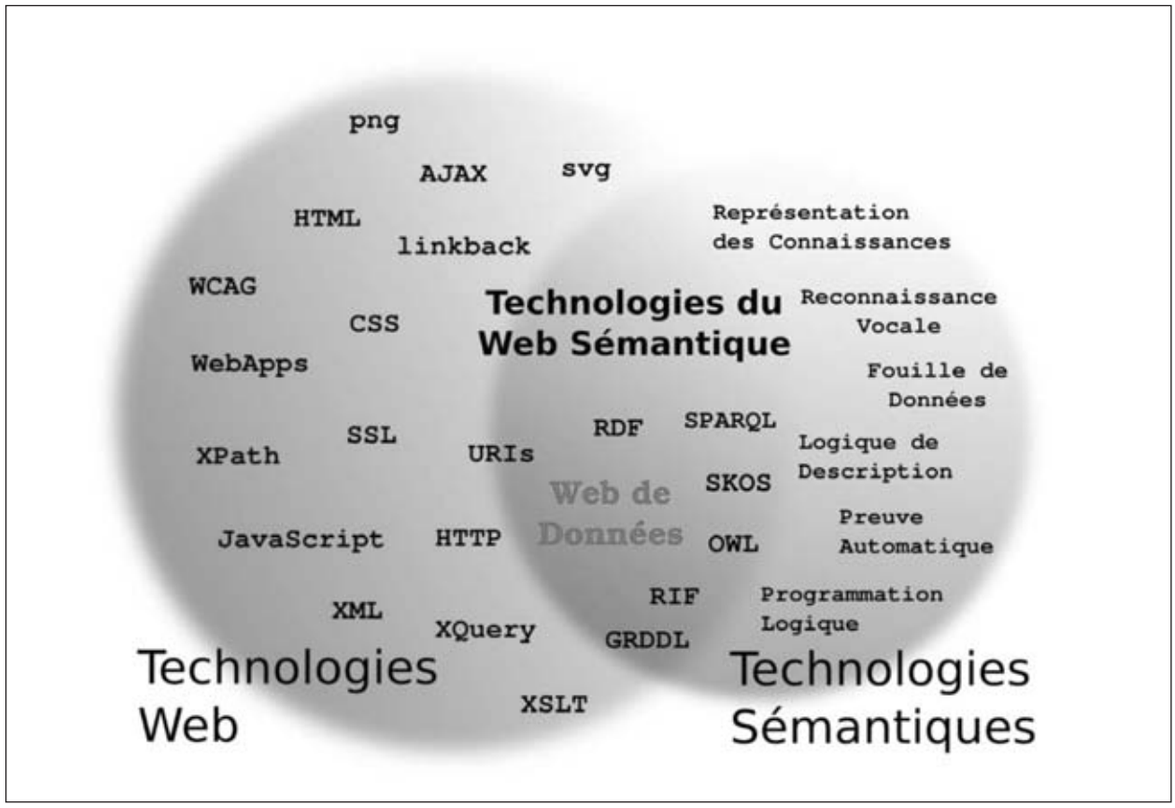


Photo 1 : Technologies du Web sémantique.

DÉCENTRALISER

Le modèle d'architecture centralisée est la réponse la plus simple pour organiser du contenu. Or, le Web est fondamentalement **décentralisé** et c'est ce qui fait son succès. Le Web de données a lui aussi besoin d'être décentralisé afin d'éviter certains problèmes classiques :

• éviter les goulots d'étranglement permet de garantir les **performances** ;

• réduire les points individuels de défaillance (Single Point of Failure – SPOF) réduit la dépendance technique ;

• empêcher une politique de publication centralisée permet de s'affranchir du bon vouloir d'un tiers et est donc une garantie de liberté.

On peut donc étendre l'architecture du Web de documents au Web de données, en utilisant des technologies **déjà existantes**. L'objectif est qu'une machine soit capable de comprendre, parcourir et utiliser ces données. Voyons maintenant comment on peut appliquer les recettes du Web aux données.

IDENTIFIER PAR UNE URI (*UNIFORM RESOURCE IDENTIFIER*)

Les bonnes pratiques de conception des sites Web mettent en avant le choix des URIs pour désigner les sous-parties des sites, les services, etc. Il en est de même avec les données.

Chaque idée/concept/ressource étant identifiée par une URI qui lui est propre, une attention particulière doit être apportée au choix de cette URI. Ainsi, un être humain qui lit une URI donnée doit déjà avoir une bonne idée de ce qui lui est associé. L'URI ne doit donc pas être ambiguë et doit être pensée avec un souci de pérennité : que désignera cette URI, dans dix ans ? Une bonne pratique est de préciser quelle logique de construction des URIs a été suivie.

Des informations transverses peuvent être associées aux données. Un bon exemple est de prendre en compte leur volatilité : lorsqu'un consommateur récupère les données associées à une URI en utilisant le protocole HTTP, il peut choisir de les mettre en cache durant une période de périssabilité récupérée dans la réponse. De même, HTTP supporte la négociation de contenu. On peut aussi demander un format particulier, une langue particulière, etc.

EXPOSER AVEC RDF (*RESOURCE DESCRIPTION FRAMEWORK*)

Le Web de données a besoin d'un modèle commun de représentation de l'information. C'est le rôle de la technologie principale du Web sémantique : RDF. Il s'agit

d'un modèle de données extrêmement simple et souple créé il y a environ une dizaine d'années.
Pour un dépôt de données particulier, commencez par identifier tous les concepts qui vous intéressent et associez-leur une URI. Chacun de ces concepts pourra être le sujet d'une question qui pourra lui être associée. Cette question est aussi appelée prédicat. La réponse à cette question est appelée objet et peut être associée soit à un autre concept (par exemple à une URI), soit à une valeur simple. Toute l'information est donc contenue dans un triplet « sujet – prédicat – objet », ou encore triplet RDF. L'exemple suivant utilise le format de sérialisation N3 pour représenter des informations concernant les concepts « France » et « Paris » :

<France> <population> 65447374.
<France> <capitale> <Paris>.
<Paris> <population> 2203817.
<Paris> <maire> « Bertrand Delanoë ».

Ce formalisme est issu de la Logique de Description du premier ordre. Le modèle sous-jacent est un graphe (1) où le sujet et l'objet sont deux nœuds reliés par une arrête étiquetée par un prédicat (2). Voici donc une représentation visuelle de l'exemple précédent sous la forme d'un graphe (voir le graphique 1).
Pour requêter un graphe de données, on peut utiliser le langage de requête prévu à cet effet : SPARQL (*Query Language for RDF*). On peut en réalité faire plus simple, juste en rendant les URIs déréférençables : étant donnée une ressource et l'URI qui lui est associée, une requête HTTP 'GET' sur cette URI doit permettre de récupérer un ensemble de triplets, par exemple ceux où la ressource apparaît.
Le choix des URIs dans l'exemple précédent n'est pas satisfaisant, car on veut pouvoir identifier des concepts très différents : personnes, lieux, gouvernements, entreprises, produits, musiques, musiciens, écoles, plantes, espèces, etc. Et surtout, on veut pouvoir partager ces concepts sur le Web. Nous avons vu que nous pouvons utiliser des URIs pour cela. Cependant, tout concept n'est pas une page Web : on a besoin de pouvoir séparer ces deux entités. La réponse à ce problème a déjà été introduite précédemment : on peut utiliser les propriétés du protocole HTTP pour négocier avec le serveur un contenu particulier, au choix, les données ou une description Web au format HTML.

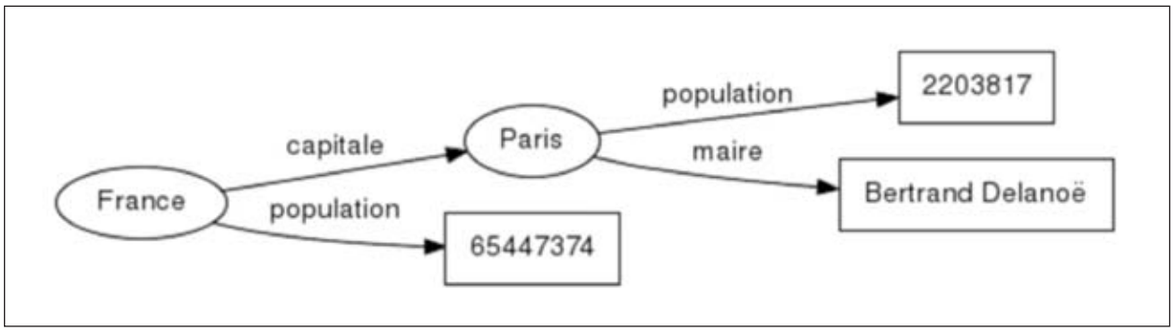
CONSTRUIRE ET DÉCONSTRUIRE UNE URI

Le concept « Paris » n'est pas une page Web : Paris existait par exemple bien avant la création de la page Web http://www.paris.fr. Cette page semble pourtant être un bon candidat pour désigner ce concept. Il existe différentes stratégies pour construire et déconstruire des URIs. Nous présentons ici deux stratégies avec leurs implications.
L'interprétation d'un *fragment* dans une URI (introduit par l'utilisation du caractère '#' et appelé *hash URI*) dépend du contexte d'utilisation. Dans une page HTML, il désigne un élément particulier du document. Dans RDF, il désigne une sous-partie du concept. HTTP 'GET' ignore simplement le fragment et récupère le document entier. Utiliser une *hash URI* permet donc de récupérer un contenu entier via HTTP 'GET' tout en désignant une sous-partie. Voici un exemple d'une telle URI : http://www.paris.fr/arrondissements#5eme.
Une stratégie beaucoup plus populaire dans la communauté est l'utilisation d'une *slash URI*. C'est par exemple la solution retenue par DBpedia (3). Par exemple, la ressource désignant Bertrand Delanoë sur DBpedia est http://dbpedia.org/resource/Bertrand_Delanoë. Il est intéressant de noter que DBpedia introduit une redirection HTTP 303 SEE OTHER lorsque la page Web correspondant à cette URI est demandée. Le navigateur Web est alors redirigé vers l'URI http://dbpedia.org/page/Bertrand_Delanoë.
Utiliser des URIs échangeables sur le Web permet alors de référencer des concepts venant d'autres sources de données : c'est l'essence même d'un Web de données ! Le nom de domaine désigne alors qui est responsable des données associées à l'URI. Voici ce que peut donner l'exemple précédent si on lui applique ce principe

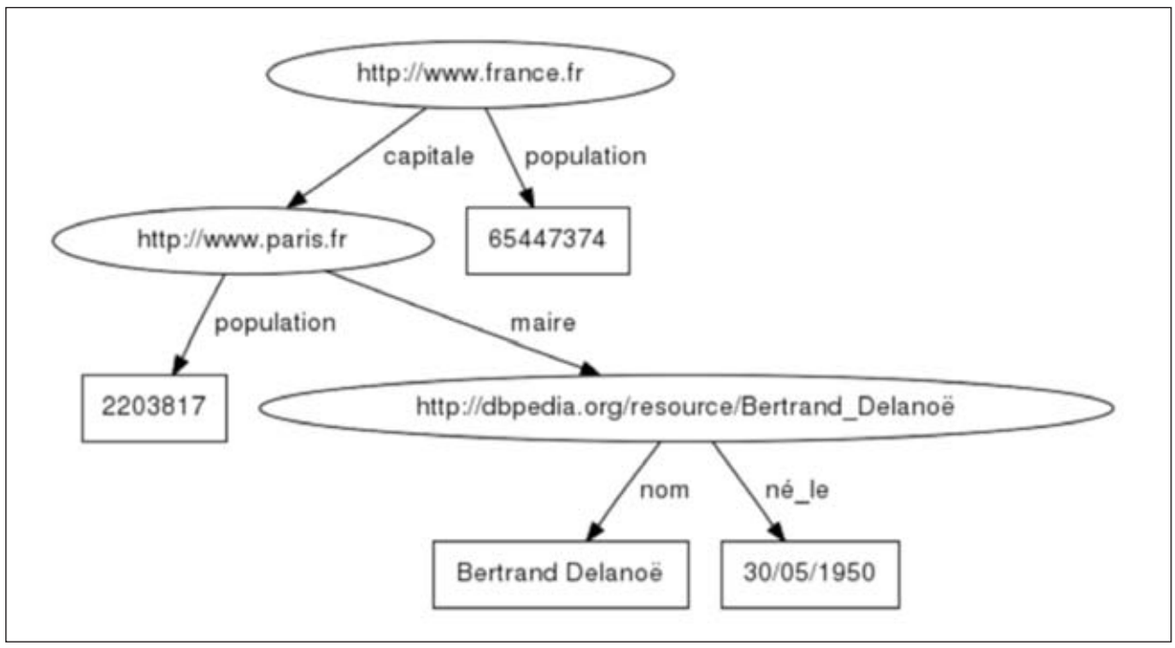
(1) par comparaison, le modèle sous-jacent de XML est un arbre.

(2) en réalité, c'est un peu plus qu'un graphe puisque les prédicats peuvent eux-mêmes être sujets ou objets d'un triplet.

(3) DBpedia est au Web sémantique ce que Wikipedia est au Web de documents : il s'agit d'une extraction automatique, au format RDF, de Wikipedia.



Graphique 1.



Graphique 2.

Cet exemple introduit la problématique des vocabulaires. Comme nous l'avons vu, construire une URI pour une ressource n'est pas difficile. En réalité, RDF spécifie que les prédicats sont aussi des ressources, et donc de véritables URIs. Ils peuvent eux-mêmes être sujets ou objets d'autres triplets, permettant ainsi de les décrire (traductions en diverses langues, propriétés, etc.). Il suffit ensuite de puiser dans les vocabulaires existants pour décrire ces données. Par exemple, DBpedia applique ce principe à la perfection en réutilisant massivement des termes issus d'autres vocabulaires que le sien. On peut maintenant réécrire l'exemple précédent en réutilisant des vocabulaires déjà existants (4) (voir le graphique 3).

Voici d'autres exemples de vocabulaires communément utilisés :

• FOAF (*Friend-of-a-Friend*) permet de décrire des individus. C'est le vocabulaire idéal pour modéliser les réseaux sociaux ;

• DublinCore est un vocabulaire spécialisé dans la description de métadonnées ;

• GeoInfo est spécialisé dans les coordonnées géographiques.

LES PRINCIPALES TECHNOLOGIES

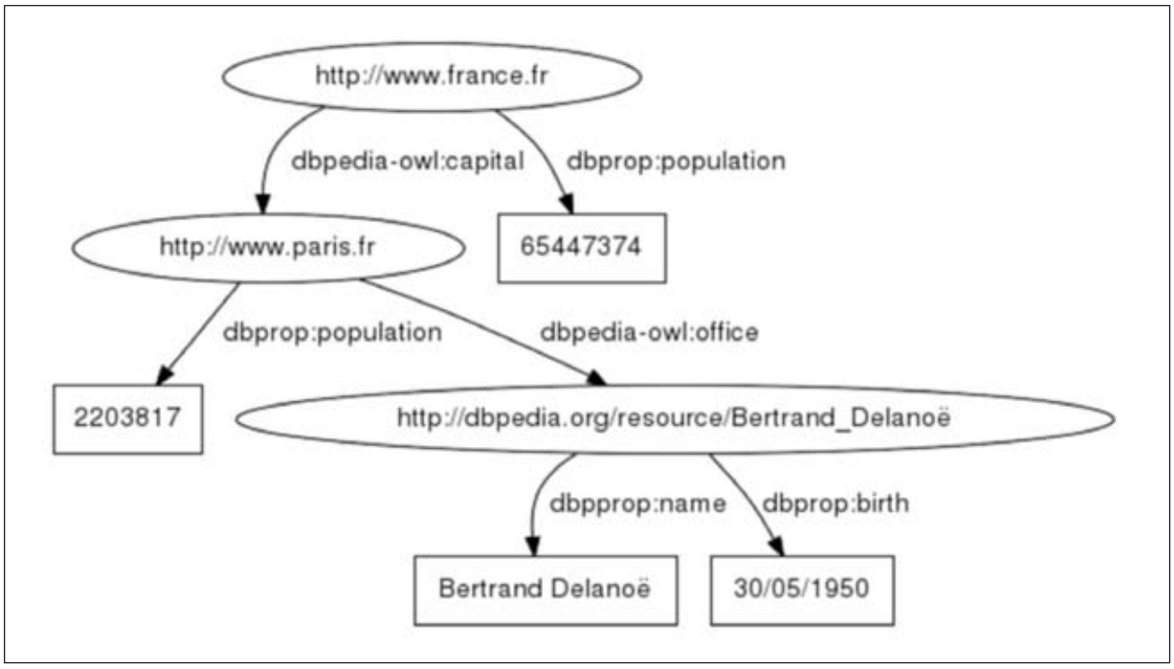
Le W3C (*World Wide Web Consortium)* héberge plusieurs groupes de travail chargés de développer et maintenir les technologies associées au Web. Le W3C et ses Membres ont élaboré et spécifié un ensemble de standards (appelés '*Recommendations*') constituant les technologies du Web sémantique.

Comme nous l'avons vu, RDF désigne le modèle de données du Web sémantique. Plusieurs formats de sérialisation sont possibles, tels que Turtle ou RDF/XML. RDFa permet, quant à lui, d'embarquer du RDF directement dans HTML. RDFS (*RDF Schema*) et OWL (*Web Ontology Language*) permettent de décrire des ensembles de données, de la même manière qu'une grammaire peut définir les bonnes constructions d'une langue. SKOS (*Simple Knowledge Organization System*) permet une représentation standard de tout type de vocabulaire contrôlé et structuré sur le Web. RIF est un format de représentation de règles à destination des moteurs de règles.

SPARQL désigne à la fois le langage de requête pour RDF et le service Web qui permet de soumettre une requête. Ce langage très simple fonctionne essentiellement par filtrage de motifs sur des graphes et s'inspire de la syntaxe de SQL et de N3. Par exemple, voici une requête valide sur DBpedia permettant de retrouver les Grandes Écoles parisiennes et leur nombre d'élèves (ces informations sont disponibles dans Wikipedia donc dans DBpedia) :

```
SELECT DISTINCT ?ecole ?nombreeleves
WHERE {?ecole <http://www.w3.org/2004/02/
skos/core#subject>
<http://dbpedia.org/resource/Category:Grandes_
Écoles>.
?ecole <http://dbpedia.org/ontology/city>
<http:// dbpedia.org/resource/Paris>.
?ecole <http://dbpedia.org/ontology/numberOf
Students> ?nombreeleves}
```

(4) RDF permet de raccourcir les URIs en définissant des préfixes. Par exemple, « dbpprop:population » est équivalent à « http://dbpedia.org/property/population ».



Graphique 3.

OPPORTUNITÉS

Il y a quinze ans, Tim Berners-Lee inventait le Web et demandait aux entreprises, aux gouvernements – bref, à tout le monde – de mettre leurs documents sur le Web et de les lier entre eux. L'idée paraissait un peu folle, mais force est de constater qu'elle a fait son chemin. Aujourd'hui, aucune entreprise ne songerait à ne pas être présente sur le Web avec un site respectant les standards. Le Web est devenu un médium fondamental dans la vie de tous les jours et prend une part toujours plus importante dans l'économie. Maintenant que ce même Tim Berners-Lee demande aux mêmes personnes de mettre leurs données sur le Web, on peut s'interroger sur l'opportunité de le faire et éventuellement se demander quel est l'état actuel du Web de données.

2009 – et 2010 dans la continuité – restera l'année de l'envol du Web de données, non pas pour les technologies arrivant à maturité (elles le sont pour la plupart, depuis quelques années), mais pour l'adoption des technologies du Web sémantique. En effet, le point d'inflexion de la courbe d'adoption a été atteint et diverses initiatives ont vu le jour, plus excitantes les unes que les autres. La plupart de ces contributions au Web de données font partie d'une initiative appelée le *Linked Open Data* (Web de données ouvert).

En octobre 2009, le New York Times a ouvert une partie de son index. Celui-ci accumule des millions de termes (datant, pour les plus anciens, de 1851) répartis selon cinq vocabulaires : sujets, personnes, organisations, lieux géographiques et ouvrages (livres, films, etc.). Un effort particulier a été réalisé dans la mise en relation avec des sources de données externes, telles que DBpedia ou Freebase. La qualité de ces données et le choix d'une licence Creative Commons permettent à tout un chacun d'accéder à ces données, mais surtout de les maintenir et les enrichir, et donc de participer à augmenter la valeur du journal.

Les grands acteurs du Web ne sont pas en reste. En mai 2009, Google a annoncé l'introduction de RDFa dans son moteur de recherche (5). Cette annonce a fait grand bruit, car le poids de Google dans la recherche en ligne pouvait inciter toujours plus de sites à exposer des données en RDFa, et c'est ce qu'il s'est passé. Parmi de nombreux exemples, on peut citer Best Buy, qui annonçait en décembre 2009 l'apparition de RDFa dans la description de ses produits ou, plus récemment, Facebook, qui expose désormais les données de son réseau social dans ce même format.

En mai 2009, les États-Unis lançaient 'Data.gov', dont le but est de faciliter au public l'accès aux données collectées par l'administration publique. L'initiative américaine a été suivie par le projet anglais 'Data.gov.uk '– lancé en septembre 2009 – et par d'autres initiatives semblables en Autriche, en Australie, etc. Toutes ces initiatives font usage, bien qu'à des niveaux différents, des technologies du Web sémantique. Pour tous ces pays, la question n'est plus de savoir s'il est opportun de participer au Web de

(5) Quelque temps auparavant, Yahoo faisait de même avec la technologie SearchMonkey, mais l'annonce faite par Google a connu un retentissement plus fort.

données, mais de déterminer comment le faire au mieux. Il y a maintenant une vraie compétition tant les enjeux sont réels. En conséquence, la quantité de données disponible augmente très rapidement, concernant aussi bien la qualité de l'air que la situation des entreprises ou le marché immobilier. Par ailleurs, les technologies du Web sémantique ouvrent la porte à des croisements d'informations difficiles à réaliser auparavant, créant ainsi des opportunités complètement nouvelles. Par exemple, un avocat de Zanesville (Ohio) a pu croiser les données de raccordement des habitants au réseau d'eau avec les origines ethniques des propriétaires (certaines demandes étant refusées). Il a pu ainsi démontrer clairement l'existence d'une discrimination.

Outre la création espérée d'opportunités ou des soucis légitimes de transparence, ces projets cherchent à attirer la communauté pour la faire participer. C'est un moyen efficace d'enrichir et de maintenir toutes ces données, en agrégeant toutes les initiatives personnelles. Le pouvoir de la communauté à enrichir des données a été illustré lors du tremblement de terre en Haïti, en janvier 2010. En réponse à la catastrophe, la communauté OpenStreetMap (un projet communautaire et ouvert, concurrent de Google Maps) a enregistré des centaines d'éditions des informations géographiques concernant Port-au-Prince, qui ont pu être directement utilisées par les équipes de secours sur place.

Le développement du Web de données est une formidable opportunité pour le monde de l'entreprise, surtout lorsque les gouvernements y participent et décident d'en faire une arme stratégique : à quand une initiative du gouvernement français ? La compréhension des enjeux, la maturité des technologies du Web sémantique et l'adoption massive par toujours plus d'acteurs permettent d'envisager de beaux jours pour le Web de données.

BIBLIOGRAPHIE

MILLER (E.) & MANOLA (F.), (Eds.), RDF Primer, W3C Recommendation, http://www.w3.org/TR/rdf-primer/, 2004.

HAYES (P.), (Ed.), RDF Semantics, W3C Recommendation, http://www.w3.org/TR/rdf-mt/, 2004.

W3C OWL Working Group, (Eds.), OWL 2 Web Ontology Language, Document Overview, W3C Recommendation, http://www.w3.org/TR/owl2-overview/, 2009.

MILES (A.) & BECHHOFER (S.), (Eds.), SKOS Simple Knowledge Organization System Reference, W3C Recommendation, http://www.w3.org/TR/skos-reference, 2009.

SEABORNE (A.) & PRUD'HOMMEAUX (A.), Eds., SPARQL Query Language for RDF, W3C Recommendation, http://www.w3.org/TR/rdf-sparql-query/, 2009.

Semantic Web Tools, W3C, http://esw.w3.org/topic/SemanticWebTools.

HERMAN (I.), (Ed.), Semantic Web Case Studies and Use Cases, W3C, http://www.w3.org/2001/sw/sweo/public/UseCases/

POLLOCK (J.), Semantic Web for Dummies, John Wiley & Sons Inc., Chichester, West Sussex, Hoboken, NJ, 2009.

ALLEMANG (D.) & HENDLER (J.), Semantic Web for the Working Ontologist, Morgan Kaufmann Publishers, San Francisco, CA, 2008.

ANTONIOU (G.) & VAN HARMELEN (F.), A Semantic Web Primer, 2nd Edition, The MIT Press, 2008.

HITZLER (P.); SEBASTIAN (R.) & KRÖTZSCH (M.), Foundations of Semantic Web Technologies, Chapman & Hall/CRC, London, 2009.