Le balisage sémantique : The Next SEO Big Thing ! (1ère partie)

Par Erlé Alberton

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Qui peut aujourd'hui ignorer les concepts de balisage sémantique, données structurées, rich snippets ou autres Schema.org à partir du moment où on s'intéresse au SEO et aux moteurs de recherche (entre autres) ? Ces données sont en effet au coeur du fonctionnement des moteurs et le seront de plus en plus. Il est donc nécessaire de les comprendre pour mieux les appréhender et les intégrer dans nos contenus. Voici une série d'articles qui devrait vous y aider.

Dans cet article en plusieurs parties, nous apporterons un éclairage sur l'apport des optimisations par balisage sémantique sur le plan du Search Engine Marketing, les origines de cette pratique et son héritage à l'heure du HTML5. Par la suite, nous détaillerons la syntaxe utilisée, que les développeurs affectionnent déjà, le Json-LD pour "Linking Data" et son application dans l'écosystème digital moderne (enrichissement des mails et des applications web); Nous mettrons en évidence les impacts sur la mutation, prévisible, des acteurs du "Search" autour de la compréhension des concepts réels manipulés par l'esprit humain. Des concepts qui se retrouvent, de plus en plus, dans les cartouches qui entourent nos requêtes et bientôt dans notre façon d'interagir vocalement avec nos smartphones, ordinateurs, objets connectés, montres, robots, voitures, drones, cafetières, lampes, prises, sites/services web… la liste s'enrichit et s'allonge chaque jour...

Aujourd'hui, plus un jour ne passe dans notre petit univers sans que l'on entende parler de cette "nouvelle" idée qu'est la sémantique appliquée au Web. N'étant pas linguistes, nous ne parlerons ici que de syntaxe et de l'importance qu'a cette technique à l'heure des graphes de données "Knowledge Graph" et "Knowledge Vault". Force est de constater qu'il existe un ardent désir pour les moteurs de recherche d'intégrer les concepts du Web 3.0 dans l'ADN des algorithmes de classement et les interfaces de restitution, toutes si différentes et diaboliquement efficaces.

Quels sont les fondements de ces techniques ? Que se cache-t-il vraiment derrière le balisage sémantique ? Quelles sont les bonnes pratiques qui pourront pousser vos contenus à être mieux classés et analysés que ceux des autres ? Classés et pas forcément positionnés… Quel avenir doit-on envisager grâce aux principes de la sémantique ?

La sémantique est partout !

Les experts, analystes et outils SEO les plus pertinents, nous parlent de plus en plus de corpus, d'ontologies, de nGram, de balisages schema.org, de thématisations, de clusters et cocons sémantiques. Le monde du SEO évolue, contraint par les évolutions des langages, les mises à jour d'algorithmes et les nouvelles capacités des moteurs de recherche. Si vous êtes optimiste, comme nous, vous verrez que c'est une très bonne nouvelle pour la croissance

de notre métier, mais qu'il faut être attentif aux mutations pour continuer à être performant et faire la part des choses entre effet d'annonce, prémonitions et réalité.

Dans un monde en pleine révolution digitale où Google, en chef d'orchestre, organise et affine ses filtres pour mieux classer les "résultats naturels", il est nécessaire de s'adapter et/ou d'évoluer dans la même direction. Les moteurs deviennent artificiellement plus intelligents, les crawlers mutent pour suivre les avancées technologiques, les indexeurs se "sémantisent," les développeurs et SEO doivent suivre et profiter de ces nouvelles possibilités.

Dans sa version 5, la spécification du langage HTML permet d'ouvrir la porte à des balises purement sémantiques comme main, header, footer, nav, aside, section. Elles viennent s'ajouter aux balises h1...n, title, meta, em, strong, ul/li, ol/li, dd/dt, figure, blocquote, time, date, que les développeurs et aujourd'hui les SEO doivent utiliser de façon optimale pour favoriser la qualité du code source. Puisque c'est bien de qualité dont il est question !
Les balises meta et link ne sont plus réservées aux en-têtes des pages mais jouent le rôle de liaison de métadonnées dans des blocs de contenu.

Que représente pour les moteurs le balisage sémantique des données structurées ?

Imaginez notre monde réel avec les yeux d'un moteur de recherche : vous êtes aveugle et probabiliste. Vous lisez à toute vitesse des milliards de balises HTML et vous fonctionnez, tout à coup, avec des équations et des calculs qui sont tout sauf humains. Votre capacité à classer les réponses à une requête repose sur un ensemble de composants logiques : votre algorithme. Il est sûrement extrêmement complexe et évidemment très performant mais vous n'êtes qu'une machine, vous ne manipulez pas de concept et le monde réel vous est fondamentalement étranger.

Vous n'êtes pas humain et pourtant vos utilisateurs le sont, eux ! Comment comprendre leurs questions et leur façon de vous interroger, comment saisir l'essence d'un contenu en ne lisant que le code source, souvent très dense et encore trop souvent peu optimisé syntaxiquement et très complexe. Une page web contient beaucoup de code parasite et c'est un énorme problème dans votre méthode de fonctionnement.

Par exemple : qu'est-ce qu'un point de vente ? Qu'est-ce qu'un smartphone ? Qu'est-ce qu'un avis client ? Quel est le rapport entre les trois? Un langage est nécessaire pour les relier et c'est là que le balisage sémantique intervient.

Ne serait-il pas plus "simple" qu'un indexeur classe des concepts ou des entités nommées, surtout s'il est guidé pour les extraire ? Sans pour autant réduire l'importance de l'approche probabiliste de comparaison des résultats de calcul d'algorithmes de détection de mots, pour lesquels des gens comme Christian Méline ou les frères Peyronnet - entre autres - sont beaucoup plus à même de vous apporter des éclairages.

Simplifier le temps de traitement de l'analyse d'une page, c'est améliorer la capacité d'extraction du contenu important. C'est un gain de performance notable dans l'analyse et depuis toujours les moteurs s'efforcent de découvrir des motifs qui se répètent et d'éliminer le bruit HTML, mais pour aller plus loin, il faudrait permettre aux éditeurs de clairement déclarer ce contenu important dans un langage ordonné et hiérarchi-

sé. En juin 2011, Google, Bing et Yahoo se sont donc unis avec un but : "create and support a common set of schemas for structured data markup on web pages" (source: http://googlewebmastercentral.blogspot.fr/2011/06/introducing-schemaorg-search-engines.html).

Schema.org était né, porté par la possibilité du HTML5 de définir de nouveaux attributs de balise itemprop, itemscope, itemtype.

Ce vocabulaire s'inscrit dans la continuité des travaux initiés dans les années 90 et marquera pourtant une rupture dans le développement et la conception de nos applications.

Si l'on fait un rapide retour en arrière, aux origines du web sémantique existaient les Triplets, issus des premières définitions de RDF (Ressource Description Framework), le modèle de graphe destiné à décrire de façon formelle les ressources web et leurs métadonnées (source Wikipédia https://fr.wikipedia.org/wiki/Resource_Description_Framework).

Longtemps mis à l'écart, ils définissent pourtant une norme très importante à l'heure de la fusion de notre monde avec le numérique. En vingt ans, ils ont muté eux aussi et se sont complexifiés, mais le modèle de base (sujet - prédicat - objet) est aujourd'hui un principe qui alimente des domaines de recherche très différents comme la communication entre machines ou les recherches sur la nature du langage humain appliqué aux machines.

- Le sujet représente la ressource à décrire ;
- Le prédicat représente un type de propriété applicable à cette ressource ;
- L'objet représente une donnée ou une autre ressource : c'est la valeur de la propriété.

Un document RDF correspond à un multigraphe orienté et étiqueté. Chaque Triplet correspond à un arc orienté dont le nœud source est le sujet, le label est le prédicat et le nœud cible est l'objet (source Wikipédia : https://fr.wikipedia.org/wiki/Resource_Description_Framework#Principes_fondamentaux).

La structure RDF est une base très générique qui sert dans la définition d'autres schémas de syntaxe approuvés par le W3C - World Wide Web Consortium : organe qui regroupe les spécifications des langages du web et dont la devise est : "Leading the Web to its full potential". Un de ces langages est l'ontologie informatique, principe qui prend tout son sens à l'heure des Knowledge Graph et Vault.

Les précurseurs

A l'origine du web sémantique, dans les années 90, deux grands hommes : Ted Nelson et Sir Timothy John Berners-Lee. Le premier est à l'origine de l'HyperText de HTTP et le second est considéré comme le père de la mise en réseau des documents structurés - ont collaboré à l'élaboration des normes du World Wide Web qui font autorité aujourd'hui.

Sir Tim Berners-Lee déclarera : « Je n'ai fait que prendre le principe d'hypertexte et le relier au principe du TCP et du DNS et alors – boum ! – ce fut le World Wide Web ! »

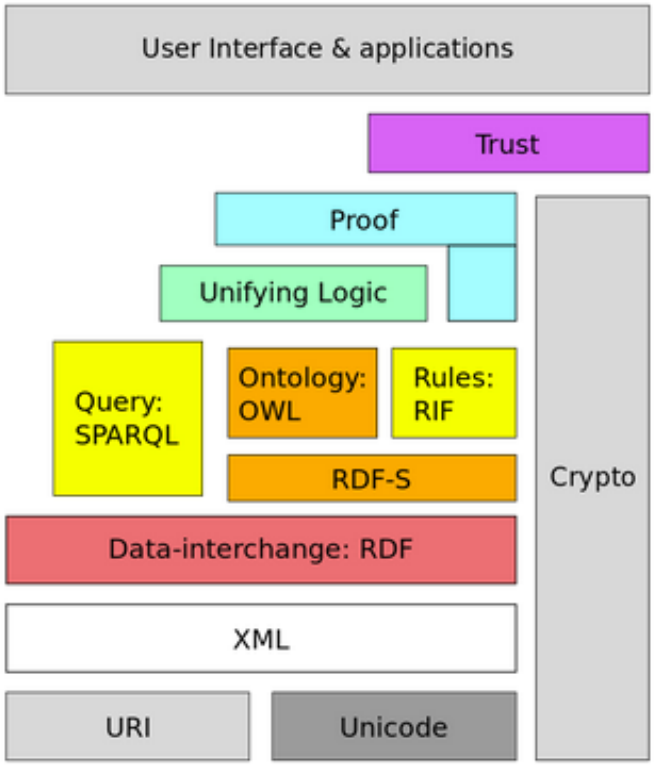

Fig.1. Le layer cake de Tim Berners-Lee en 2006

Ces normes sont regroupées et formalisées par le W3C - World Wide Web Consortium - dont Sir Tim Berners-Lee est l'actuel président, il lui a même été décerné le prix Quadriga pour son engagement dans l'innovation, le renouvellement et l'esprit de pionnier, c'est dire ! Nous vous invitons à lire sa page Wikipédia pour en apprendre plus sur les débuts du Web et de la sémantique : https://fr.wikipedia.org/wiki/Tim_Berners-Lee.

A eux deux et grâce à la somme d'anonymes qui les ont aidés, nous avons vu le Web littéralement exploser en l'espace de vingt ans. Internet participe à l'essor de la culture et du savoir, et favorise l'émergence des langages de communication entre machines pour interagir ensemble, stocker et classifier des données.

Mais cela n'aurait pu être possible sans certains grands fondements : le protocole HTTP, l'identification par URL et le langage HTML, et plus important pour nous ici, l'ontologie et les métadonnées.

Qu'est-ce qu'une ontologie ?

Philosophiquement, d'après Aristote et Wikipédia, c'est l'étude de l'être en tant qu'être c'est-à-dire l'étude des propriétés générales de tout ce qui existe.

Du point de vue informatique, toujours d'après Wikipédia, c'est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'information, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

Dans la première définition, le terme "être" équivaut à tout ce qui "existe", c'est l'étude de la réalité. Dans la deuxième définition, on parle de la modélisation de concepts aux propriétés structurées via des metadonnées dans le but de décrire exhaustivement le savoir ou la connaissance.

Nos pages web font partie de ce savoir hétéroclite que les moteurs se doivent, aujourd'hui, d'organiser le plus justement possible.

Comment coder une ontologie en 2015 et quelle valeur ajoutée pour votre contenu ?

Property	Expected Type	Description
sameAs	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Freebase page, **or official website**.

Fig.2. Définition de l'ontologie via SameAs

Une ontologie est donc la représentation d'un concept traduit dans une syntaxe comprise par un robot. Le site http://schema.org en décrit trois différentes : microdata, RDFa et Json-LD. Cette dernière étant en passe de devenir la plus efficace dans les échanges entre machines. Le format Json est très léger, nous verrons, dans la suite de cet article, en quoi cela peut être important pour l'internet des objets mais si vous maîtrisez les API, vous aurez déjà saisi notre pensée.

Donc le concept "SEO" dans un article peut être codé comme suit :

```
<span itemscope
itemtype="http://schema.org/Thing">
  <em itemprop="name">SEO</em>
  <link itemprop="sameAs"

href="https://fr.wikipedia.org/wiki/Optimisation_pour_les_moteurs_de_recherche" />
</span>
```

La syntaxe microdata est la plus adaptée pour enrichir un texte précisément.

Nous déclarons une nouvelle entité grâce à l'attribut HTML5 itemscope, son itemtype est le type le plus bas dans l'arbre de schema.org Thing - identifié par l'URI de la page de définition http://schema.org/Thing dans l'exemple - ensuite vient se placer la propriété name de l'objet et enfin le lien <link> vers la ressource décrivant le sens du concept.

Attention ce n'est pas un lien au travers d'une ancre mais plutôt une liaison de ressource vers une autre.

Des outils en ligne comme semanticmarker.com pourront vous aider à intégrer ces principes à vos services grâce à des API.

NB : L'expression à conceptualiser peut être entourée d'une balise nativement sémantique, , balise de mise en emphase car, in fine, elle doit être importante pour le sens du texte, elle sera mise en valeur.
Nous en profitons pour partager la v.0 d'un plugin WordPress que nous avons écrit pour créer une ontologie grâce à un simple shortcode.
Testez l'ontologie sans effort : https://github.com/omkom-web/wp-ontology

En allant plus loin dans la description de l'ontologie, on remarque que ses domaines d'application vont du web sémantique à l'intelligence artificielle… Enfin, le grand mot est lancé, l'intelligence artificielle : la capacité d'une machine à comprendre et interagir avec le monde qui l'a créé.

Nous sommes en 2015, au matin de la recherche conversationnelle, de la contextualisation des réponses, du knowledgeGraph, des machines à l'apprentissage intelligent et du BigData. La compréhension des moteurs suit cette évolution logique et prévisible qui n'en est qu'à ses débuts.

Grace aux ontologies votre contenu est maintenant un texte qui contient des concepts interprétés par le moteur, imaginez le classement qui peut en découler. Vous avez compris l'impact possible de cette nouvelle manière de créer de la qualité de code source et le potentiel SEO que cela peut apporter.
Il faut ouvrir la page de description de Thing http://schema.org/Thing pour trouver la définition de l'ontologie via sameAs (voir figure 2).

"or official website" pour rappel :

```
<script type="application/ld+json">
{
  "@context" : "http://schema.org",
  "@type" : "Organization",
  "name" : "Ma compagnie",
  "url" : "http://ma-compagnie.fr",
  "sameAs" : [
    "http://www.facebook.com/ma-
compagnie",
    "http://www.twitter.com/ma-
compagnie",
    "http://plus.google.com/ma-compagnie"
  ]
}
</script>
```

La propriété sameAs est utilisée dans la définition des profils des réseaux sociaux d'un site web, c'est donc une ontologie.

Une remarque sur la syntaxe Json : pas de virgule à la fin de la dernière ligne d'un groupe, Google respecte cette notation stricte, vous devez en faire autant.

La liaison des pages entre concepts couplée au maillage existant et aux ancres de liens, peuvent vous aider à valoriser des termes importants de votre contenu. On peut supposer que d'un point de vue sémantique le texte qui contient ce type de liaisons fortes et formelles d'un concept vers sa page de définition, peut aiguiller un robot à classer votre contenu de manière plus pertinente.

Comme toute optimisation, l'ontologie doit être utilisée pour aider le moteur à donner du sens à votre contenu mais doit servir avant tout le décryptage du texte et il serait mal venu d'ajouter de ontologies sur tous les mots de façon brutale et automatique. Souvenez vous qu'en SEO, une optimisation doit se diluer naturellement dans votre site.

Pensez plutôt à mettre l'accent sur des termes clés qui appuieront le sujet principal de votre page de manière ciblée et contextuelle. Faites le même travail SEO qu'avec vos liens et vos ancres, mais dans une nouvelle syntaxe.

Le mois prochain, nous explorerons en profondeur le vocabulaire schema.org pour en déterminer des bonnes pratiques afin d'enrichir de données structurées des sites web, applications, mails et blog. Et nous terminerons sur une analyse des perspective de ces nouveaux principes sémantiques pour améliorer la qualité des pages et peut être devenir une norme à l'heure de l'Internet des Objets sûrement grâce au Json-ld.

Vous coderez ensuite le contenu d'une manière si particulière, qu'il en deviendra un enchevêtrement de concepts complexes, extrêmement détaillé, interprété et classé par une nouvelle génération d'indexeurs sémantiques. Une sacrée perspective ! A bientôt !

Erlé Alberton, *Responsable SEO, Direction Digital Commerce, Orange France (https://twitter.com/cubilizer).*