# Google veut devenir un moteur syntaxique et sémantique : quel impact pour le référencement ? (1ère partie)

Retour au sommaire de la lettre

| | | |
|---|---|---|
| **Domaine :** | Recherche | **Référencement** |
| **Niveau :** | Pour tous | **Avancé** |

*En février 2012, plusieurs sites web ont repris les mots d'Amit Singhal, grand prêtre de l'algorithme de pertinence de Google, expliquant que le moteur de recherche allait fortement sémantiser ses résultats dès cette année et répondre directement, dès l'affichage de la SERP, aux questions posées par les internautes et comprises par Google. Mais comment Google a-t-il mis en place ces algorihmes sémantiques et quelle sera leur portée en termes de SEO ? Cet article, en deux parties, explique le concept d'un moteur de recherche sémantique et analyse la façon dont Google pourrait bien fonctionner dans les mois qui viennent. Une petite révolution semble bien se préparer dans le domaine du SEO...*

Le 13 février dernier, un article paru sur le site Mashable (http://mashable.com/2012/02/13/google-knowledge-graph-change-search/) a fortement intrigué la communauté des référenceurs anglo-saxons. Cet article évoquait en effet des propos d'Amit Singhal, annonçant dans les mois qui viennent un changement majeur dans le moteur de recherche Google, pouvant impacter jusqu'à 20% des requêtes effectuées par les internautes : l'exploitation par Google du « knowledge graph » (graphe de connaissances).

La portée de ce changement a été ensuite nuancée par Amit Singhal, qui est chargé de la qualité de l'algorithme de classement de Google, notamment suite à un autre article paru le 15 mars dans le Wall Street Journal (http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html) qui commentait le papier de Mashable. Mais cela reste une évolution majeure, bien qu'attendue depuis longtemps, du moteur de recherche leader vers un fonctionnement de type moteur sémantique (au sens du "web sémantique").

Mais quels sont ces changements annoncés par Google ? En quoi est-ce réellement révolutionnaire ? Quelle est la portée réelle de cette évolution ? Nous allons essayer ici de faire le tri parmi les informations lâchées par Google, en expliquant les méthodes et les algorithmes à l'œuvre derrière les moteurs sémantiques.

Pour mieux comprendre cet article, on pourra relire les articles suivants, du même auteur, parus dans la lettre « Recherche et Référencement » d'Abondance :

- *Vers un moteur de recherche conceptuel grâce à la sémantique*. Numéro de janvier 2011 : http://recherche-referencement.abondance.com/2011/01/vers-un-moteur-de-recherche-conceptuel.html

- *La reconnaissance des entités nommées par les moteurs de recherche*. Numéro de mai 2009 : http://recherche-referencement.abondance.com/2009/05/la-reconnaissance-des-entites-nommees.html

## L'objectif : répondre directement aux questions de l'internaute

Le moteur Google que nous connaissons se place dans la droite continuité des premiers moteurs de recherche conçus pour le web. Le principe de ces moteurs a toujours été d'afficher une liste de "liens bleus" pointant vers des pages dont le contenu est estimé « proche » de la requête tapée dans un champ de recherche.

Comme nous l'avons vu dans un précédent article (*les requêtes commerciales dans Google*, février 2012 : http://recherche-referencement.abondance.com/2012/02/les-requetes-qic-

query-is-commercial.html), on classe traditionnellement les requêtes tapées par les internautes dans trois catégories : les requêtes navigationnelles, informationnelles et transactionnelles.

Une requête navigationnelle correspond parfaitement au cas d'usage auquel les moteurs de recherche usuels actuels sont adaptés : il s'agit de trouver une page sur le web, et de « naviguer » vers celle-ci. Proposer une liste de liens pertinents à l'internaute est parfaitement adapté à ce besoin.

Mais ce système s'avère peu efficace dès lors que l'internaute cherche une information (requête informationnelle) ou cherche à accomplir une action sur le web (requête transactionnelle : acheter, télécharger...). La philosophie de Google a toujours été de faire mieux que ses concurrents sur les requêtes informationnelles, qui sont les requêtes les plus fréquentes sur le web.

Force est de constater que sur ce type de requêtes, les liens bleus et l'utilisation d'une mesure de similarité sémantique est assez frustrante. Si on cherche la monnaie en vigueur au Burkina Faso, le moteur ne vous donnera pas la réponse directement, mais une liste de pages dont le contenu est proche de la requête. Ce sera à vous ensuite d'ouvrir différentes pages en espérant que la réponse figure aussi dans la page. Et dans de nombreux cas, vous tomberez sur des pages qui ne contiennent pas la réponse, ou plus ridiculement, qui contiennent aussi la question. C'est d'ailleurs cette étrange propriété du Google actuel qui explique pourquoi les clones de Yahoo Answers ou de CommentCaMarche sont aussi bien référencés : Google classe en premier des pages de questions et non... des pages de réponse.

Ce défaut majeur des moteurs de recherche traditionnel a été identifié il y a déjà une quinzaine d'années. La solution est connue aussi depuis très longtemps : elle consiste évidemment à fournir la réponse directement à l'internaute.

Mais fournir la réponse à la question posée s'avère redoutablement complexe. Tout d'abord il faut comprendre le sens de la question posée. Or, les requêtes des internautes ne sont pas tapées en langage naturel, mais sous forme de mots clés. Ces mots clés sont souvent ambigus, et mal orthographiés. Bref, soit on ne dispose pas d'assez d'informations pour comprendre la question, soit la question est ambigüe et approximative, soit les deux. Dans ces conditions, on a toutes les chances de donner une mauvaise réponse ou d'être incapable d'en donner une.

Ensuite, il faut disposer de cette réponse dans une base de données. Nous verrons plus loin que ce sont les progrès que Google a fait dans ce domaine (l'indexation des faits et des données) qui permettent à la firme de Mountain View de lancer cette fonctionnalité en 2012.

## *Les limites de cette approche*

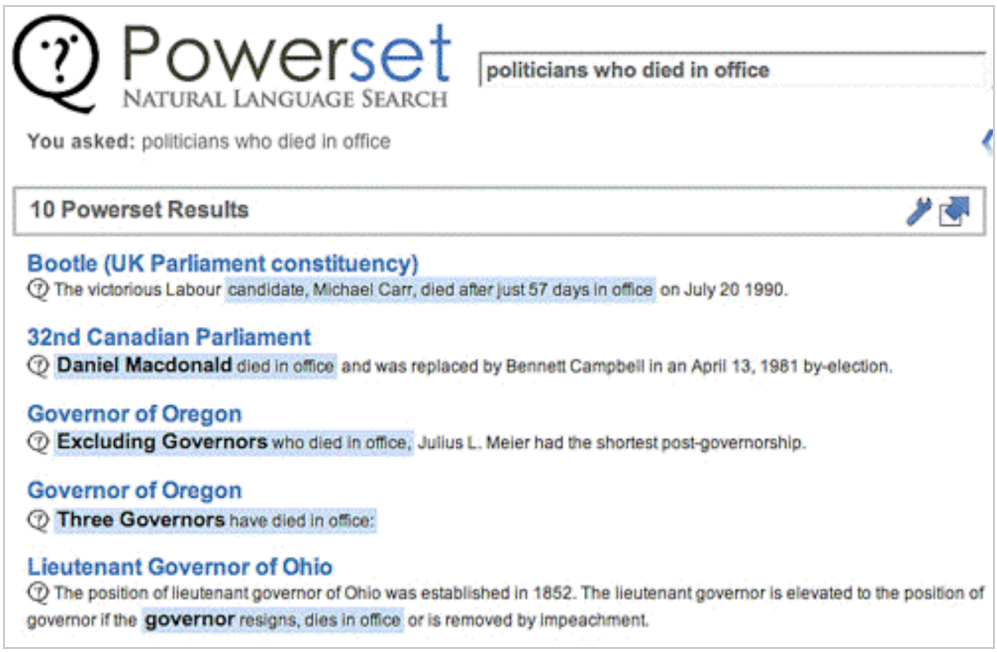
Dans le même temps, les requêtes informationnelles ne sont pas les seules que l'internaute cherche à effectuer sur un moteur de recherche : si quelqu'un cherche une page web, lui montrer des "liens bleus" reste pertinent. C'est vrai aussi pour la plupart des requêtes transactionnelles.

Qui plus est, certaines requêtes informationnelles n'appellent pas de simples informations : si renvoyer « 1515 » suffit à répondre à la requête « Date bataille Marignan », la question « réchauffement climatique » appelle potentiellement des développements un peu plus importants.

## *Les "moteurs sémantiques" existent déjà*

En fait, les moteurs basés sur cette technologie existent déjà. Leurs applications sont souvent limitées à des applications spécifiques, mais quelques tentatives « grand public » sont apparues ces dernières années. Parmi ces moteurs, on peut citer Hakia, Cuill, Powerset, Cluuz, ou encore Wolfram Alpha.

Aucun de ces moteurs n'a réellement rencontré de succès : Hakia s'est recentré sur des activités sur abonnement, Cuill a disparu (ses fondateurs ont réintégré les équipes de Google), et Powerset n'est plus en ligne depuis son rachat par Microsoft. Wolfram Alpha est encore accessible, et fournit d'ailleurs des résultats de type « moteur sémantique » à la version US de Bing.




*Powerset en action sur la requête « politicians who died in office ».*
*Ce site n'est plus accessible.*

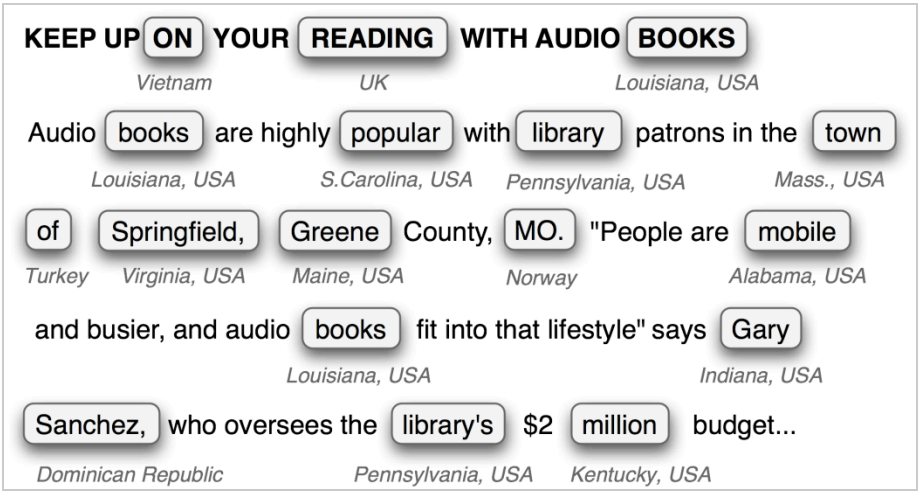
*Réponse de Wolfram Alpha sur la requête « people at the signing of the Declaration of Independance »*

En fait, comme nous l'avions expliqué dans un article de la lettre "Recherche & Référencement » daté de janvier 2011, (« Vers un moteur de recherche conceptuel grâce à la sémantique), il semble que les moteurs sémantiques ou les moteurs conceptuels aient plus d'avenir en tant que fonctionnalité complémentaire d'un moteur grand public classique que comme fonctionnalité « stand alone » (sur un moteur généraliste). Et il semble que Google prépare bien une implémentation de ces fonctionnalités sous forme de complément à son système existant, ce qui semble assez cohérent.

## Comment fonctionne un "moteur sémantique" ?

La première étape du foncitonnement d'un "moteur sémantique" consiste à reconnaître dans les documents ce que l'on appelle les « entités nommées ». Ce concept est en quelque sorte une extension de la notion usuelle de « noms propres » à tout terme qui correspond au nom d'une chose réelle, ou à un identifiant de cette chose.

Voici quelques exemple d'entités nommées : raisons sociales d'entreprises, marques, modèles, noms de lieux, noms de personnes, identifiants de personnes (le Président de la République) n° de sécurité sociale, codes postaux, n° de téléphones, adresses mail, etc.

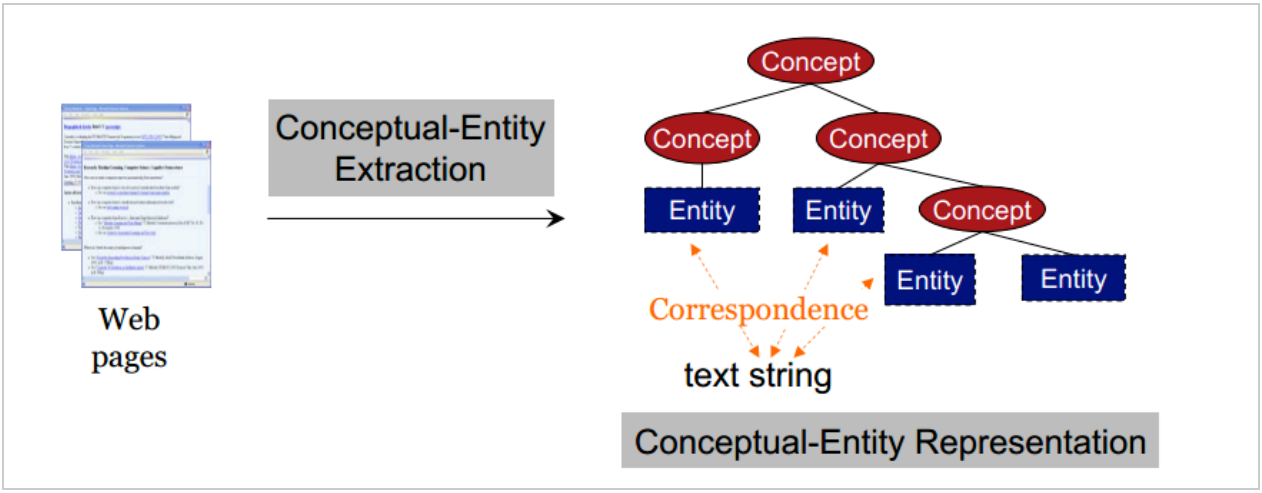



*Une illustration des difficultés rencontrées lors d'une tentative naïve de reconnaissance d'entités nommées. Une simple approche par correspondance est vouée à l'échec, car le même terme peut correspondre à un mot de la langue (ex : « reading » et « Reading » qui est une ville d'Angleterre). Le même nom « Gary Sanchez » peut également correspondre à plusieurs individus. Une analyse syntaxique poussée permet de limiter le premier type d'erreur. L'étude approfondie du contexte permet de résoudre le second. Mais aucun processus de REN (Reconnaissance d'Entité Nommée) n'est exempt d'erreurs.*

La seconde étape consiste à essayer de reconnaître les entités nommées identifiées sur chacune des pages qui en parlent. Ce processus de balisage automatique transforme des pages web contenant une information non structurées en des documents qui eux, sont en partie structurés. L'une des étapes les plus difficiles est de « désambiguïser » chaque cas de correspondance entre une graphie donnée et une entité.

Exemple : si une page parle de Michael Jackson, de quel Michael Jackson parle-t-on ? Le basketteur ? Le chanteur ? Le plombier du Bronx ?

Le problème suivant que les moteurs ont dû résoudre, est de pouvoir travailler sur TOUTES les entités nommées, et de créer un algorithme universel et générique. La plupart des approches se sont contentées jusqu'ici de travailler sur un nombre limité de types d'entités nommées. Or une grande partie des entités ne trouve pas sa place dans les classifications traditionnelles.
Il faut ensuite identifier les attributs associables à une entité nommée (pour une personne, sa taille, son age, son adresse mail, son numéro de téléphone), et extraire les valeurs correspondantes des pages web.




*Dans l'approche de Tom Mitchell de Google, les paires attributs-valeurs liées à une entité nommée sont remplacées par une notion plus générique d' « entité conceptuelle ». L'arbre d'entité conceptuelle décrit de manière plus universelle toutes les relations qui unissent une entité avec d'autres entités. L'un des enjeux consiste aussi à savoir construire une grille de reconnaissance des entités liées à une zone de l'arbre à partir d'une chaine de caractères.*

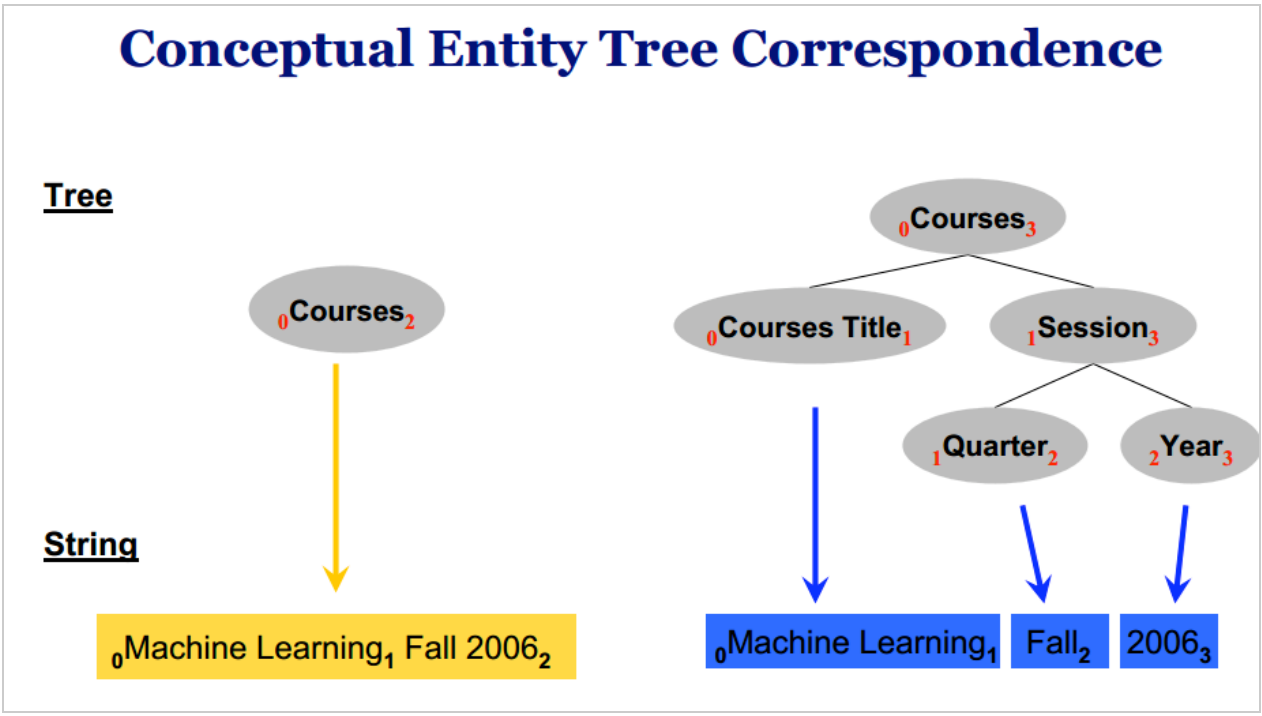
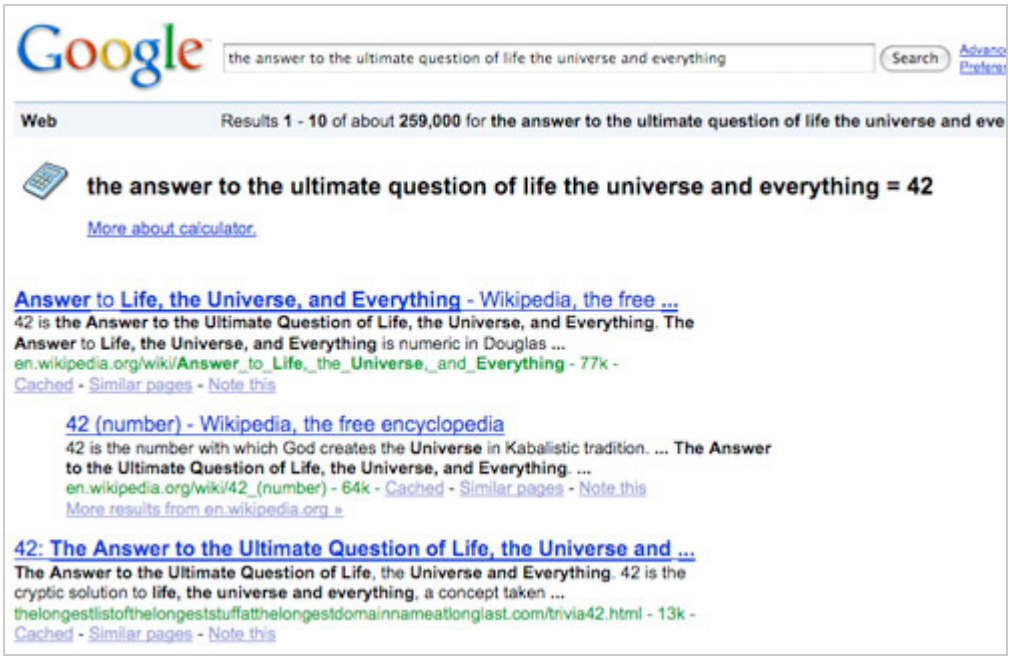



*Illustration issue d'une présentation de Tom Mitchell, chercheur chez Google, présentant les bases de son approche baptisée « arbres d'entités conceptuelles ». A gauche, l'approche traditionnelle, à droite sa méthode, qui autorise l'extraction d'une information plus riche et plus structurée.*

Et pour finir, il faut créer un index spécial, optimisé pour répondre rapidement aux questions des internautes et faire remonter soit les faits ou les données indexées, soit les pages « taggées » et associées avec la ou les entités nommées reconnues dans la requête. Notons au passage que la reconnaissance des entités nommées dans la requête n'est pas un problème trivial, car le contexte très réduit ne permet pas de « désambiguïser » facilement les mots clés tapés.

## *De timides tentatives dans le passé*

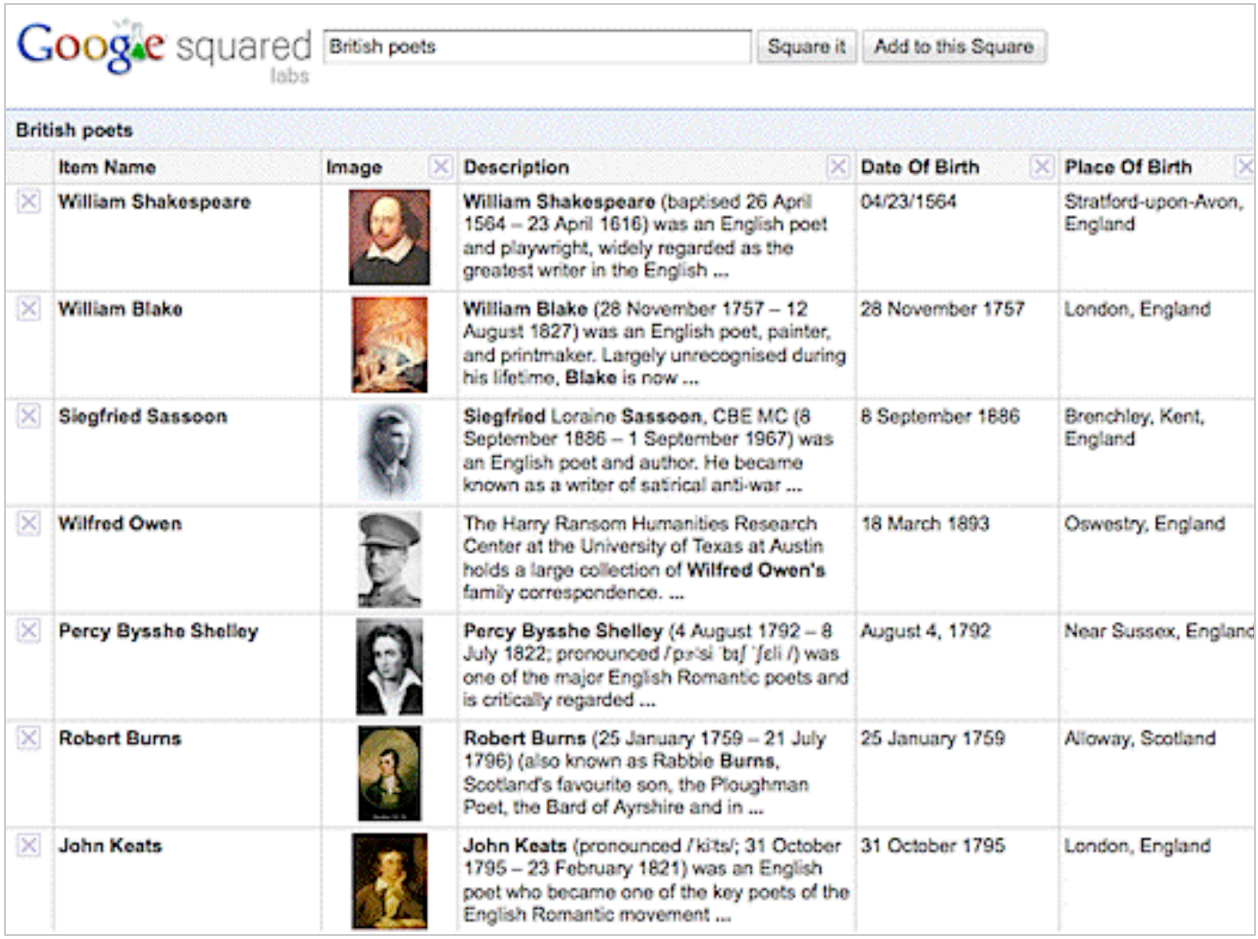
Il est difficile de savoir depuis quand exactement Google s'intéresse aux entités nommées et à une approche de type sémantique. Mais probablement depuis de nombreuses années. Les premières applications visibles de la reconnaissance des entités nommées sont apparues lors de la mise en place de la recherche universelle en 2007, avec des « oneboxes » affichant des faits ou des informations, ou des cartes, ou des vidéos lorsque certaines entités nommées étaient reconnues. Ce mécanisme, très fruste, s'est peu à peu sophistiqué.




*Une « onebox » de type Easter Egg dans la recherche universelle de Google. Tapez "answer to the ultimate question of life, the universe and everything" , et vous obtiendrez comme réponse "42", soit la célèbre réponse révélée dans le roman satirique « Hitch Hiker's Guide to the Galaxy ».*

## *Le projet Google Squared*

Une autre tentative plus aboutie est l'application Google Squared (« Google au carré ») qui a été présente un moment dans les *labs* de Google et qui permettait d'obtenir des listes d'entités et de couples « attributs + valeurs », et de jouer avec l'index de faits collecté par Google. Mais la base de faits de Google était particulièrement limitée.




| Item Name | Image | Description | Date Of Birth | Place Of Birth |
|---|---|---|---|---|
| William Shakespeare | | **William Shakespeare** (baptised 26 April 1564 – 23 April 1616) was an English poet and playwright, widely regarded as the greatest writer in the English ... | 04/23/1564 | Stratford-upon-Avon, England |
| William Blake | | **William Blake** (28 November 1757 – 12 August 1827) was an English poet, painter, and printmaker. Largely unrecognised during his lifetime, **Blake** is now ... | 28 November 1757 | London, England |
| Siegfried Sassoon | | **Siegfried** Loraine **Sassoon**, CBE MC (8 September 1886 – 1 September 1967) was an English poet and author. He became known as a writer of satirical anti-war ... | 8 September 1886 | Brenchley, Kent, England |
| Wilfred Owen | | The Harry Ransom Humanities Research Center at the University of Texas at Austin holds a large collection of **Wilfred Owen's** family correspondence. ... | 18 March 1893 | Oswestry, England |
| Percy Bysshe Shelley | | **Percy Bysshe Shelley** (4 August 1792 – 8 July 1822; pronounced /pɜ:si ˈbɪʃ ˈʃeli /) was one of the major English Romantic poets and is critically regarded ... | August 4, 1792 | Near Sussex, Englanc |
| Robert Burns | | **Robert Burns** (25 January 1759 – 21 July 1796) (also known as Rabbie **Burns**, Scotland's favourite son, the Ploughman Poet, the Bard of Ayrshire and in ... | 25 January 1759 | Alloway, Scotland |
| John Keats | | **John Keats** (pronounced /kiːts/; 31 October 1795 – 23 February 1821) was an English poet who became one of the key poets of the English Romantic movement ... | 31 October 1795 | London, England |

*Une liste de poètes britanniques donnée par Google Squared. Outre une liste d'entités (les poètes), Squared indique aussi des informations structurées comme leur nom, leur portrait, leur description, leur date et lieu de naissance etc... Google Squared a été une fenêtre ouverte sur un index de type « graphe de connaissances » simplifié. Il n'est hélas plus accessible dans les labs de Google.*

Mais le rachat de la société Metaweb par Google en juillet 2010 a clairement accéléré le rythme avec lequel Google a pu préparer son évolution (même partielle) vers un moteur sémantique. Ce ce dont nous parlerons le mois prochain : comment Google pense-t-il intégrer les technologies de Metaweb - et éventuellement d'autres - pour mieux "comprendre" et analyser les demandes des internautes pour y répondre directement dans ses SERP.

Encore quelques semaines de patience, donc, pour en savoir plus sur cette "sémantisation" que Google nous promet pour cette année...

***Philippe YONNET** , Directeur SEO international, Twenga.*

**Réagissez à cet article sur le blog des abonnés d'Abondance :**
**http://blog-abonnes.abondance.com/2012/04/google-veut-devenir-un-moteur.html**