## EntityCube, un moteur sur les entités nommées, "made by Microsoft"

Retour au sommaire de la lettre

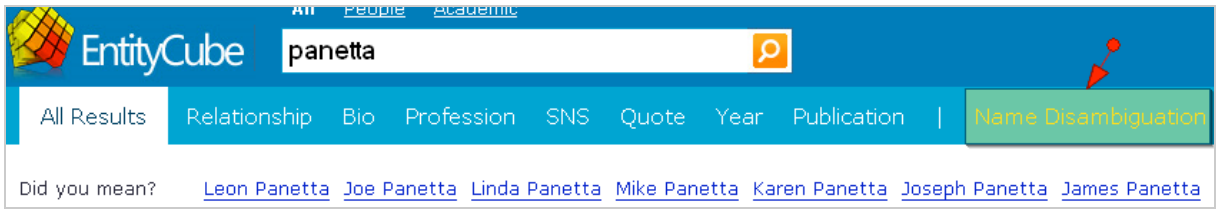
| | | |
|---|---|---|
| **Domaine :** | **Recherche** | Référencement |
| **Niveau :** | **Pour tous** | Avancé |

*EntityCube est un moteur de recherche basé sur les entités nommées et créé par les équipes chinoises de Microsoft. Bien que certaines possibilités ne semblent pas fonctionner encore et que la langue française ne soit que partiellement prise en compte, l'outil est prometteur et permet de mettre en relation de nombreuses informations structurées pour mener à bien une veille de qualité. Un outil à tester absolument...*

Après l'Egypte et Kngine le mois dernier, nous vous proposons de continuer ce tour du monde improvisé des moteurs innovants avec **EntityCube** (http://entitycube.research.microsoft.com), un moteur de recherche et d'exploitation des contenus web développé par les équipes chinoises de Microsoft et issu d'un projet en langue chinoise déjà opérationnel, baptisé **Renlifang** (http://renlifang.msra.cn/).

Comme son nom l'indique, EntityCube est un moteur qui travaille à partir des entités nommées. La version actuelle, qui est une bêta, exploite déjà (d'après ses créateurs) 3 milliards de pages web. Des créateurs qui s'expriment d'ailleurs assez peu. On saura juste grâce à la page "About" qu'"*EntityCube génère des résumés de milliards de pages web publiques contenant des informations relatives à des personnes, des lieux et des organisations et permet d'en explorer les relations*". Sans équivoque.

Pour tester l'outil et afin d'être certains d'avoir des résultats dans l'actualité, nous avons décidé d'utiliser le nom de "Leon Panetta", l'actuel directeur de la CIA.
Première constatation : en ne tapant que le nom propre de celui-ci, la page de résultats d'EntityCube propose de désambiguïser le nom. Normal pour un moteur sémantique mais toujours appréciable.



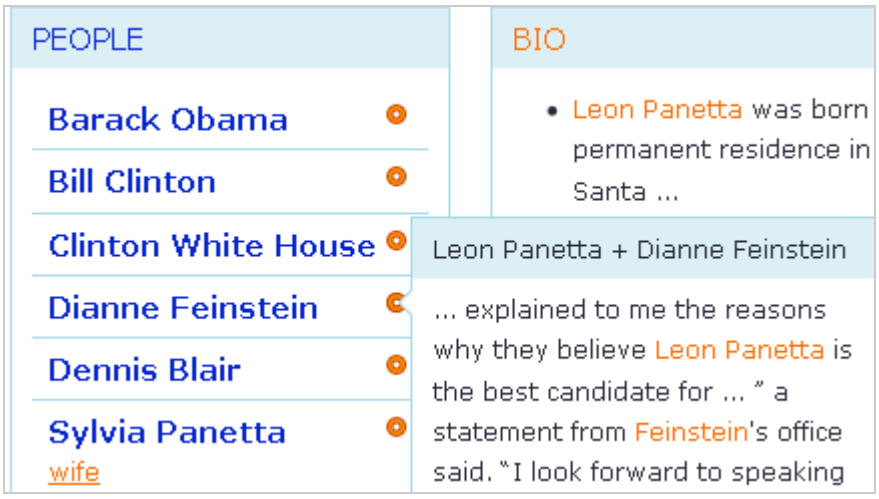

En cliquant sur son nom complet, on arrive alors sur la page dynamique qui lui est consacrée. Si ces éléments son présents sur le web, on y trouvera quelques éléments biographiques ainsi que, sur le côté gauche, les noms des personnes les plus souvent liées à celle recherchée.

Le reste de la page est composé des différents éléments que l'on retrouve également (de manière plus développée) dans les onglets et que nous allons maintenant détailler :

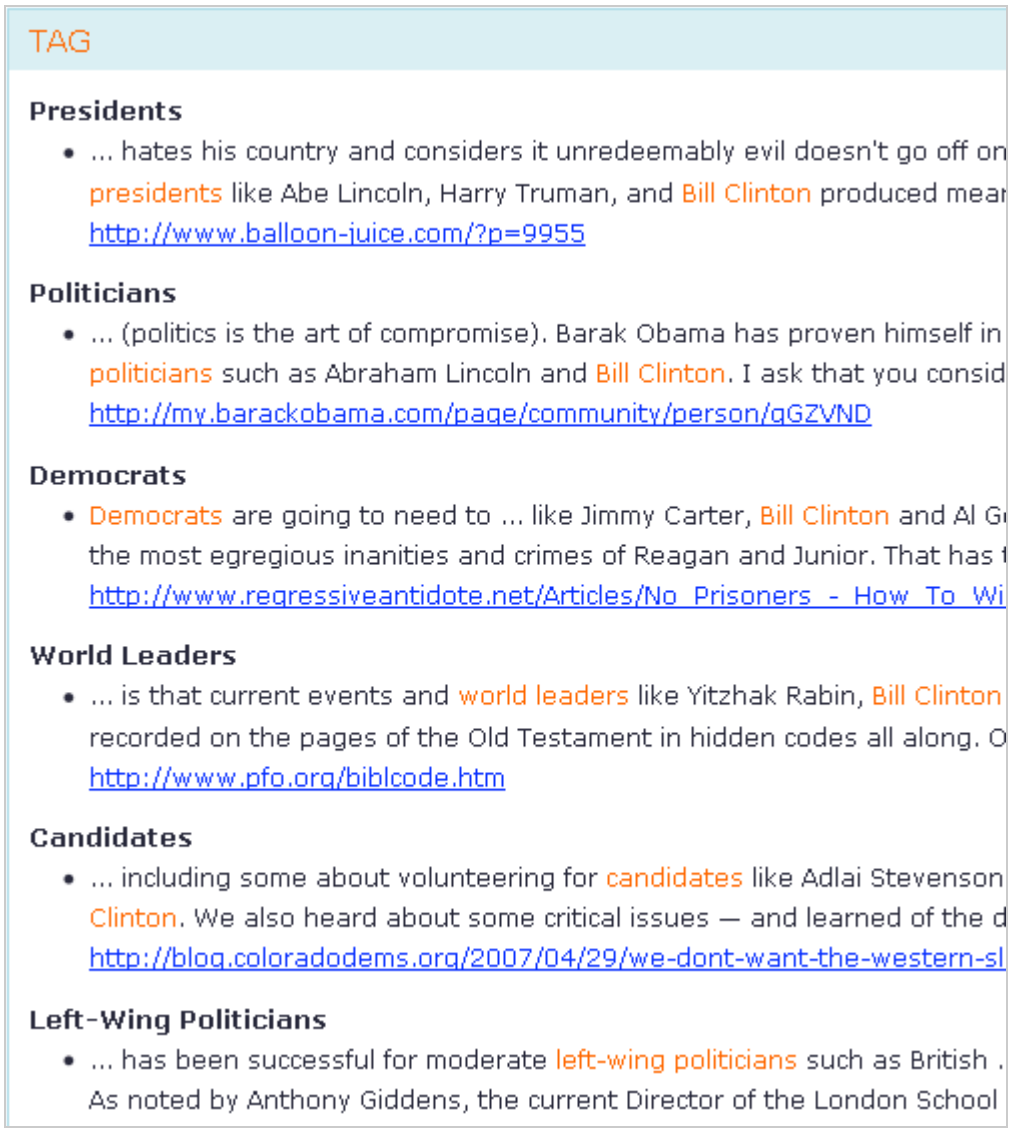
- ***Relationship*** : la page de résultats affiche les extraits de pages web sur lesquels les noms de Leon Panetta et d'une autre personne sont associés. Il suffit de cliquer dans la partie gauche sur "*show details*" sous chaque nom propre pour faire apparaître de nouveaux résultats.


De la même manière on peut faire apparaître les lieux ou les organisations les plus cités avec "Leon Panetta" en choisissant "*LOC*" ou "*ORG*" dans les onglets.

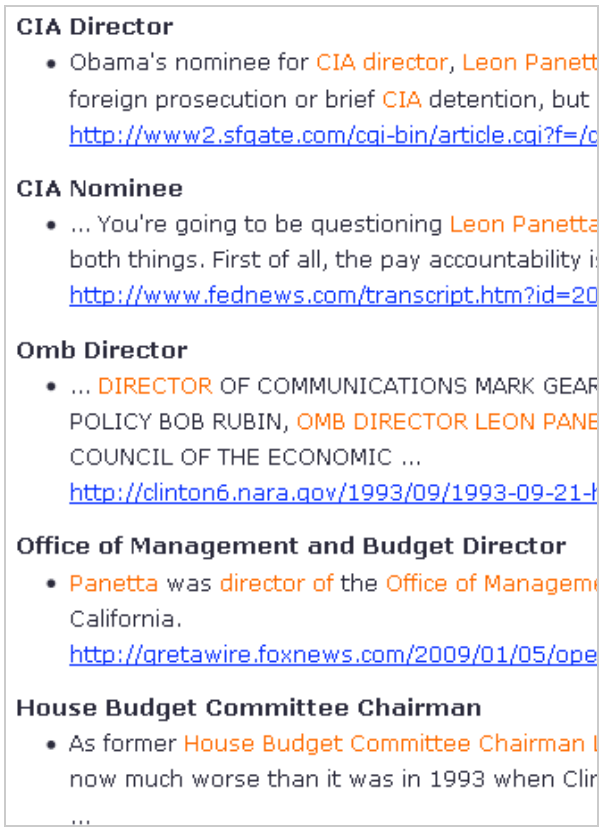

- ***Bio*** : Comme on s'en doute cet onglet regroupe les pages sur lesquelles des éléments biographiques sur la personne recherchée apparaissent. On retrouve ici notre onglet *People* qui va nous permettre :
- De lancer une recherche sur une personne en cliquant sur son nom.
- De lancer une recherche croisée "*Leon Panetta + untel*" dans la base en cliquant sur le rond orange.



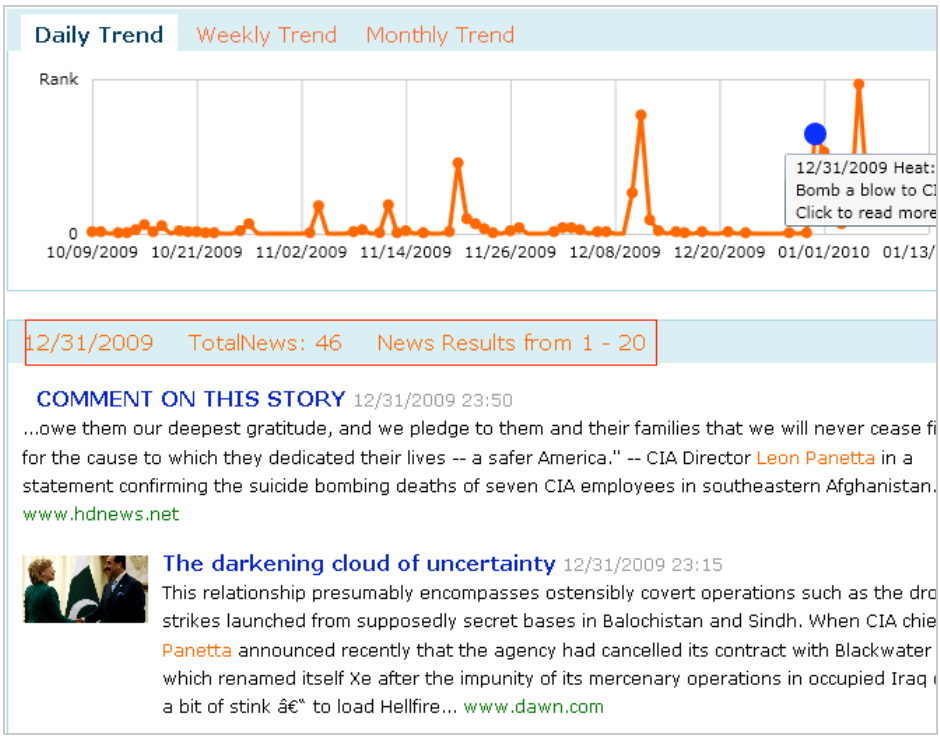

- ***Tag*** : cet onglet identifie les mots-clés (adjectifs) qui sont associés à une personne et qui ne sont pas des entités nommées. Nous utilisons ici le nom "Bill Clinton" qui comporte plus d'entrée et est donc plus démonstratif que celui de Panetta.




- ***Profession*** : cette page de résultats fonctionne comme la précédente mais se concentre sur les intitulés de poste remplis par les "personnes cibles".

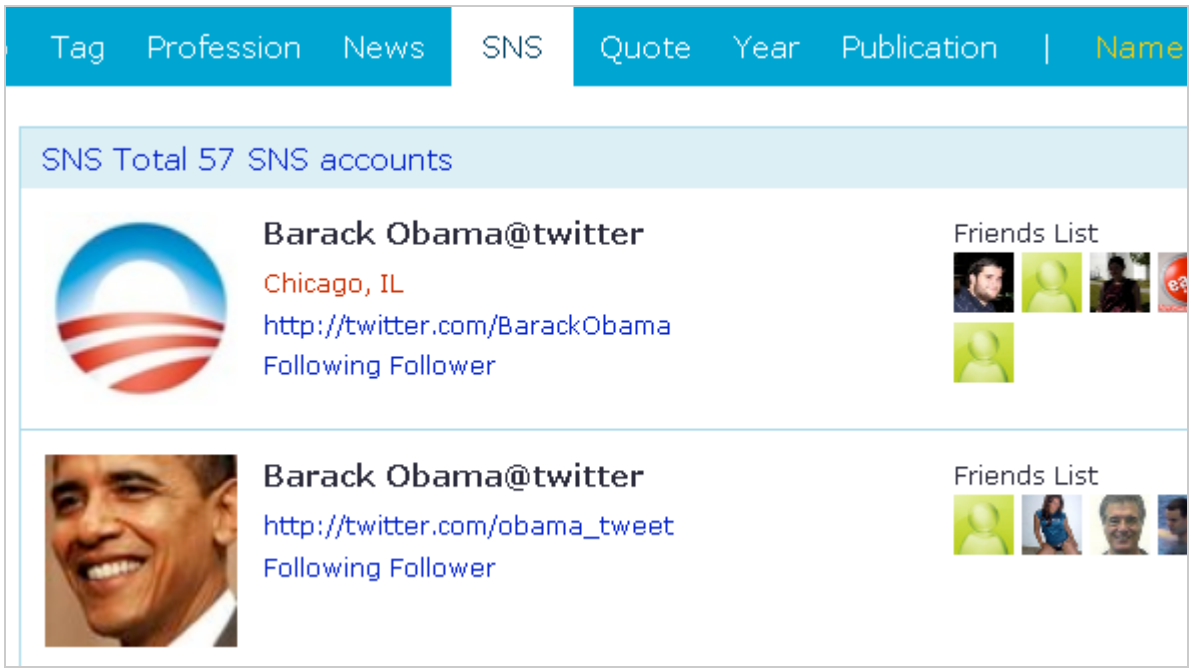



- ***News*** : c'est l'une des forces de l'outil puisqu'EntityCube va à la fois présenter les articles relatifs à votre mot-clé mais également les dédoublonner et en tirer un histogramme de fréquence sur les dernières 24 heures, la semaine ou le mois. Si ces périodes ne vous suffisent pas, vous pouvez donner vos propres dates butoirs dans la fenêtre intitulée "*News Trends*", en haut à gauche de la page de résultats. A noter que les courbes de l'histogramme sont cliquables et vous permettent d'accéder aux articles qui ont "créé" la tendance, ici la journée où 7 agents de la CIA ont été tués par un attentat :



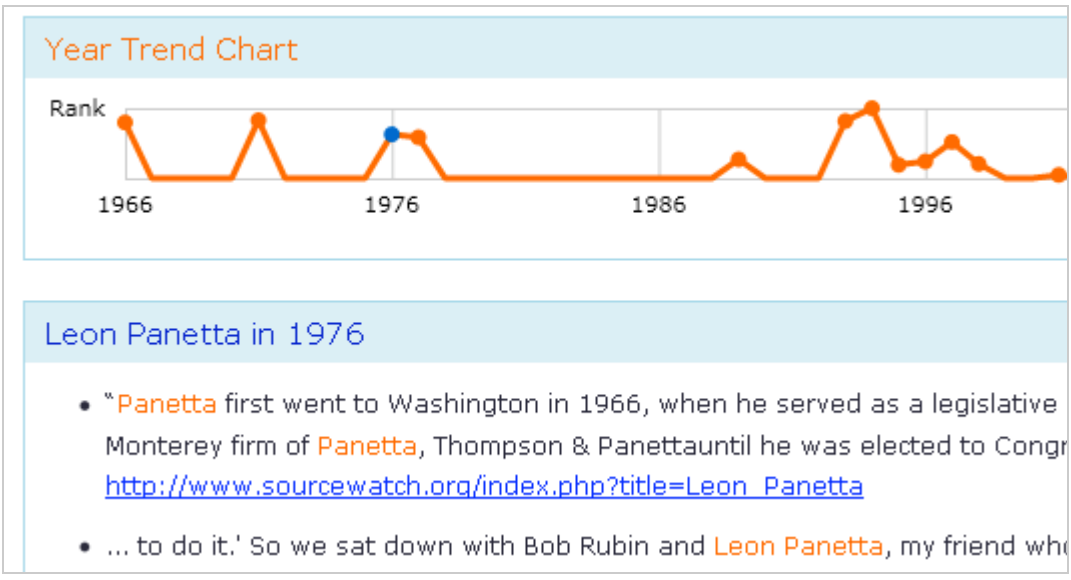
Enfin, il vous est possible de récupérer le flux RSS de la page d'actualité dynamique.

- ***SNS*** : cet onglet permet d'identifier les comptes de services de *social networking* de la personne (ou de pseudos car la désambigüisation ne va pas jusque là). Un exemple avec le compte Twitter de Barack Obama et les 56 faux comptes qui l'accompagnent (le premier est le bon) :




- ***Quote*** : cet onglet permet d'identifier des phrases prononcées par votre "personne cible" et d'accéder aux pages web correspondantes.

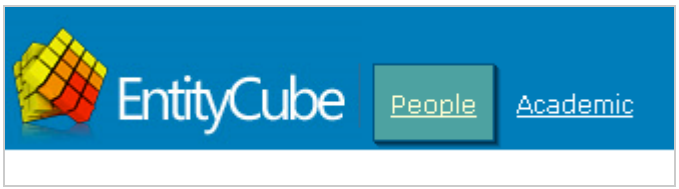
- ***Year*** : replace sur un graphique de tendance les grandes dates qui concernent la "personne cible" et vous permet évidemment d'accéder aux pages relatives d'un clic :




Vous retrouvez une partie de ces résultats présentés par blocs sur la page principale (onglet *All results*) ainsi que les résultats de recherche de Bing. Cette même page propose à gauche des résultats une fenêtre intitulée "*See larger Guanxi Map*". En cliquant dessus vous n'obtenez... rien. Idem si vous cliquez en haut à droite de la page :


Grâce à cette fonctionnalité, EntityCube propose normalement une représentation cartographique de type "réseau social" des entités nommées liées à une personne. Cela peut aider à l'analyse et permet également de naviguer dans les résultats comme on le ferait dans Kartoo (www.kartoo.com) ou, plus encore, dans le toujours excellent Silobreaker (www.silobreaker.com). Malheureusement, elle semblait être désactivée sur la version anglaise de l'outil au moment où nous avons fait nos tests...

Il est toutefois possible d'accéder à un type de représentation de l'actualité assez proche à partir de la page d'accueil d'EntityCube, en choisissant le lien *People* :

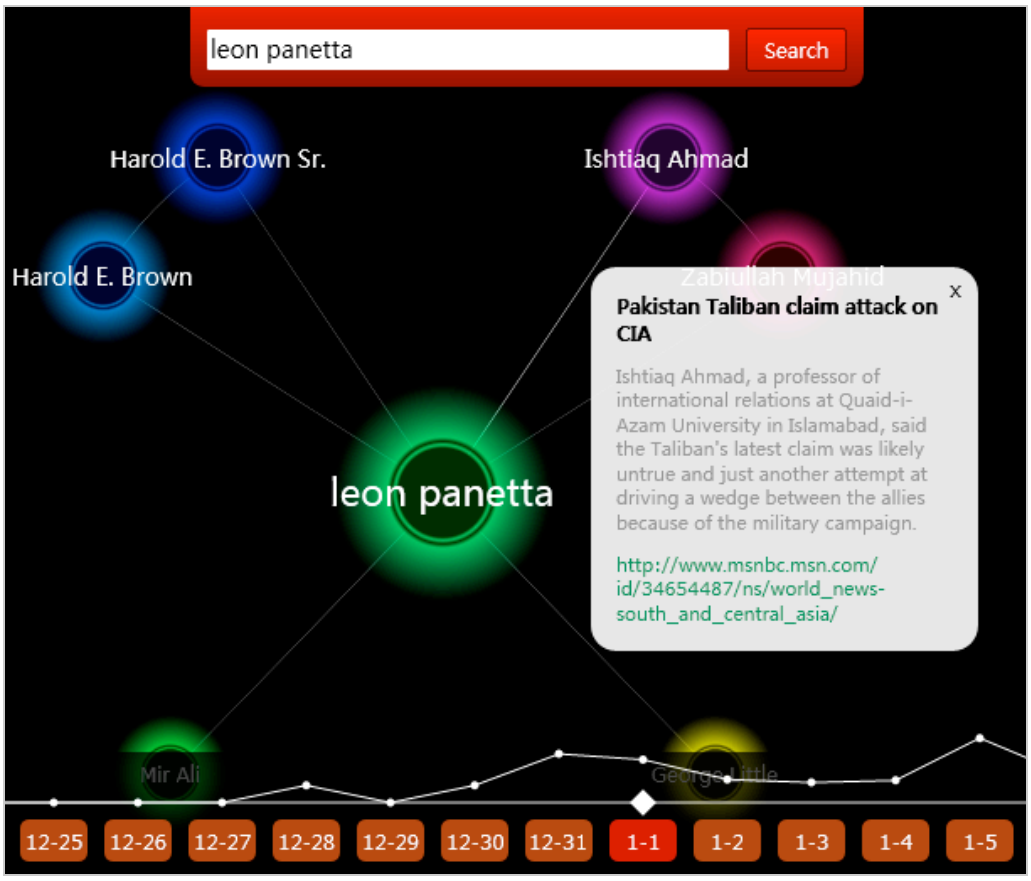



On arrive alors sur une nouvelle page d'accueil proposant trois fonctionnalités :

- ***Hot list*** : classement des "sujets chauds" pour différents thèmes et différentes périodes :


- ***Six-degrees*** : ne fonctionne pas mais permet, d'après Altsearchengines.com, de voir quels sont les intermédiaires qui relient 2 personnes.

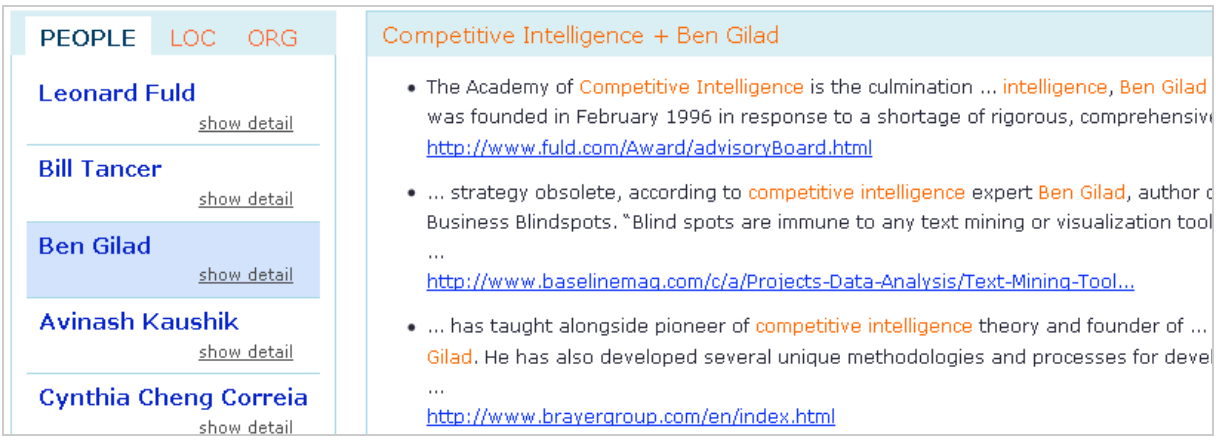
- ***Guanxi timeline*** : (celle qui nous intéresse ici) cartographie l'actualité en faisant émerger les entités qui y sont citées heure par heure. Il est possible, en utilisant le curseur, de remonter dans le temps. Par ailleurs, en cliquant sur le lien unissant deux personnes, on fait apparaître l'article dans lequel ils sont cités et on peut bien sûr cliquer dessus pour accéder à l'article initial. Voici par exemple la cartographie générée pour Leon Panetta le 1er janvier dernier, au lendemain de l'attaque contre la base de la CIA en Afghanistan :




Autant que nous avons pu en juger, la fonctionnalité Guanxi Timeline se différencie de Guanxi en ce qu'elle ne travaille - comme son nom l'indique - que sur l'actualité.

On peut bien sûr entrer dans ce moteur des mots-clés qui ne seraient pas des entités nommées. Ainsi, avec le terme "competitive intelligence", on obtient des résultats issus de Bing mais traités par EntityCube, c'est-à-dire enrichis d'entités nommées. Nous obtenons par exemple dans ce cas des noms de personnes liées à la "*competitive intelligence*" qui sont présentés dans deux blocs à gauche des résultats :

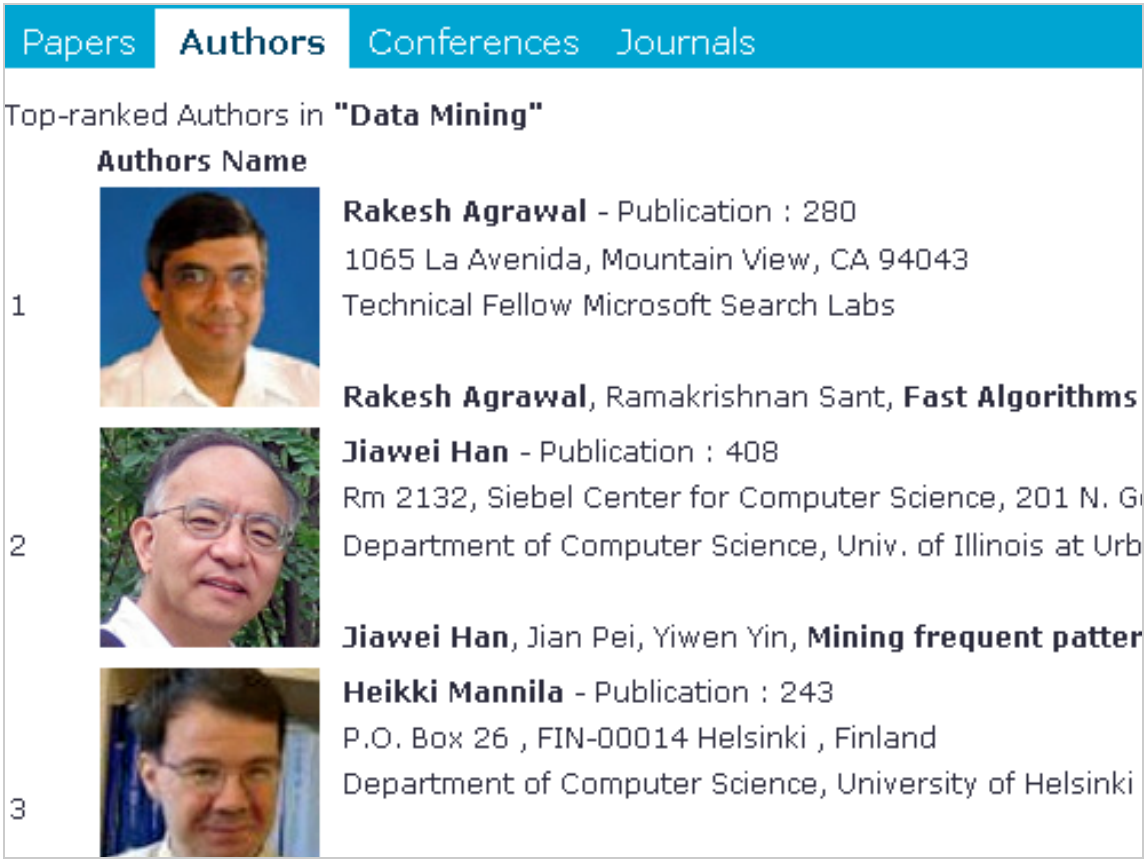
- ***People*** : toutes les personnes liées au terme "competitive intelligence". Pour exploiter ces résultats, il est en fait plus simple de cliquer sur l'onglet *Relationship* dans la page de résultats. On peut alors croiser les résultats et accéder aux pages concernant à la fois le thème choisi et la personne identifiée. Exemple ici avec Ben Gilad, fondateur de l'*Academy of Competitive Intelligence* et auteur de nombreux ouvrages :



A partir de ce même bloc on pourra également identifier les pages web mettant en relation le thème choisi avec des lieux géographiques (onglet *LOC*) ou des organisations (onglet *ORG*).

- ***Author*** : il s'agit des chercheurs identifiés par l'outil comme ayant produit des articles sur le thème de la "competitive intelligence". On peut également répérer à partir de ce bloc les conférences ayant trait au sujet en cliquant sur l'onglet "*Conf*" ainsi que les revues scientifiques dans lesquelles ces contributions sont parues en cliquant sur l'onglet "*Journal*". Le plus simple pour travailler avec ce bloc est de le lancer indépendamment en choisissant l'onglet principal "*Publication*" sur la page de résultats.

Dernier aspect et pas le moindre, EntityCube peut aussi être utilisé comme portail de recherche d'articles scientifiques. Il suffit pour cela de cliquer sur l'onglet *Academic* lorsqu'on est sur la page d'accueil. On arrive alors sur une page spécifique qui permet d'accéder aux articles ou auteurs les plus cités par domaines :




On peut évidemment lancer des recherches par mots-clés et cela nous ramène alors à l'interface *Author* déjà décrite.

Pour ce type de recherche spécifique, les créateurs d'EntityCube utilisent en fait la base de données de Microsoft Academic Search (http://academic.research.microsoft.com/), l'excellent moteur de recherche d'articles scientifiques à base d'entités nommées qu'ils ont également créé.

En conclusion, EntityCube est un excellent moteur dans son genre. Il permet d'accéder à des informations structurées mais aussi, et c'est là l'essentiel, de les exploiter grâce aux différents liens entre entités nommées qu'il detecte et permet de mettre en oeuvre. On est bien sûr un peu frustré de ne pas pouvoir utiliser les Guanxi Maps, mais il est probable que l'interruption de ce servcie ne soit que passagère. Après tout, il s'agit d'une version bêta...

Autre regret, le fait que la langue française ne soit pas traitée, ou plus exactement mal traitée. Ainsi EntityCube ne prend pas en charge les mots accentués. Quoiqu'il en soit, Microsoft surprend très agréablement avec ce moteur de grande qualité.

***Christophe Deschamps***
*Consultant et formateur en gestion de l'information.*
*Responsable du blog Outils Froids (http://www.outilsfroids.net/)*

**Réagissez à cet article sur le blog des abonnés d'Abondance :**
**http://abonnes.abondance.com/blogpro/2010/01/entitycube-un-moteur-sur-les-entites.html**