

François Bourdoncle (Exalead) : "Nous explorons la recherche par sérendipité"

[Retour au sommaire de la lettre](#)

Depuis quelques semaines, la société Exalead a lancé de nombreuses nouveautés sur son moteur de recherche : recherche sur les blogs, les forums et Wikipedia, campagnes de buzz, blog sur la vie de la société, etc. Il met également en place une ambitieuse politique de partenariats en France et à l'étranger. Dans la foulée d'un projet Quaero qui se cherche encore, il nous a semblé intéressant de faire le point sur cet acteur majeur des moteurs de recherche en France avec son PDG, François Bourdoncle...

François Bourdoncle, un des pionniers du marché du logiciel de recherche, est le Président Directeur Général d'[Exalead](#). Il co-fonde Exalead en 2000 pour révolutionner le monde de la recherche en entreprise en proposant une plate-forme technologique qui centralise toutes les données de l'entreprise, et permet aux utilisateurs un accès unifié à toutes ces informations.

Chercheur à l'origine, François Bourdoncle est un vrai passionné de technologie : on lui doit le développement de plusieurs langages de programmation, Jazz, un langage dédié aux circuits électroniques et surtout Exascript, un langage orienté-objet XML basé sur la technologie de langage Java.

C'est au laboratoire de recherche de Digital à Paris puis à Palo Alto, en Californie, où il travaille pour le LiveTopics Project d'Alta Vista, qu'il rencontre Patrice Bertin avec qui il fonde Exalead.

X-Mines, Docteur en Informatique, Titulaire d'un Master en Sciences Informatiques de l'Ecole Normale Supérieure, François Bourdoncle est un ancien professeur associé à l'Ecole Polytechnique et à l'Ecole Normale Supérieure. Il est également l'auteur de plusieurs études techniques et publications, et intervient régulièrement dans des conférences spécialisées.

Il nous a reçu dans ses locaux parisiens pour répondre à nos questions sur les projets en cours du moteur de recherche Exalead et pour explorer plus globalement le marché actuel des moteurs de recherche...



François Bourdoncle, bonjour et merci de bien vouloir répondre à nos questions. Tout d'abord, pouvez-vous nous présenter Exalead en chiffres en 2007 ?

Nous sommes actuellement 75 personnes (60 personnes environ à Paris, 8 aux Etats-Unis et 2 en Grande-Bretagne et en Allemagne dans des bureaux nouvellement créés) avec une perspective de 100 personnes à très court terme. Notre moteur de recherche propose un index de 8 milliards de pages pour lequel nous optimisons chaque jour le crawl. L'objectif est de mieux prendre en compte les "gros" sites pour obtenir une meilleure indexation encore des sites web proposant le contenu le plus important et de meilleure qualité. Nous rafraîchissons de 20 à 40 millions de pages par jour (mais nous en crawlons plus) et l'index complet est entièrement renouvelé en 2 à 3 mois actuellement. Notre moteur reçoit chaque mois 1 million de visiteurs uniques et notre audience croît régulièrement depuis de nombreux mois...

Comment vous positionnez-vous dans le paysage actuel des moteurs de recherche ?

Plusieurs études ont démontré qu'actuellement, Google et Yahoo! étaient très proche en termes de pertinence et constituaient le duo de tête. Derrière eux, des moteurs comme Live Search de Microsoft et Exalead étaient de sérieux challengers. Nous travaillons à l'excellence de nos résultats de recherche et d'énormes efforts quotidiens y sont consacrés. Nous allons d'ailleurs sortir une nouvelle release de notre algorithme de *ranking* d'ici un mois qui devrait faire faire un bond à la qualité des résultats fournis à nos utilisateurs.

Quelle est la stratégie et le modèle économique d'Exalead ?

Notre positionnement est totalement différent de celui d'un Google ou d'un Yahoo!. Nous ne sommes ni une régie publicitaire ni un média mais une plateforme de technologies et de services. Nous fournissons une technologie à des clients qui ont besoin global de recherche d'informations.

Nous sommes persuadés que la problématique du "search" est composée de 3 parties qu'il est difficile de différencier :

- Le Web Search ou recherche sur les pages web du monde entier.
 - La recherche verticale, ou recherche sur une partie du Web, par exemple un pays, une thématique (santé, tourisme, finance, etc.) ou un support (les blogs, les forums, etc.).
 - La recherche dans les bases de données de type Pages Jaunes ou toute autre base structurée.
- Les 3 domaines sont selon moi indissociables et une technologie de recherche se doit d'être d'efficace dans chacune d'entre elles. C'est donc le credo de notre offre "One Search" qui veut unifier le "search" d'entreprise. Pour cela, nous passons des accords de partenariats importants, comme par exemple avec [Ilse Media](#) aux Pays Bas, qui est le premier portail indépendant du pays et génère 30 millions de recherches par mois, ou en France avec des groupes de presse de la PQR ou du monde informatique (les annonces officielles seront prochainement faites), pour proposer une solution globale de technologie de recherche "intra site" et sur le Web. On pourrait également parler de [Yakaz](#), créé par des anciens d'AOL, qui propose un métamoteur de petites annonces basé sur notre technologie. Une double approche B2B / B2C nous semble indispensable dans un proche avenir...

Dans quelle voie la recherche d'information s'oriente-t-elle ?

Le premier "plateau d'excellence" à atteindre pour un moteur de recherche est la pertinence. Google et Yahoo! sont bons à ce niveau et nous n'en sommes pas loin. Le deuxième plateau reste la quantité d'informations à indexer. Aujourd'hui, Google indexe 20 à 25 milliards de pages, Yahoo! 15 à 20 milliards, Live Search 4 à 6 milliards, Exalead 8 milliards et le prochain challenge consiste certainement à indexer 100 milliards de pages. Comment rester très pertinent devant cette masse énorme d'information ? Et surtout comment conserver un coût de production du service raisonnable qui ne menace pas la marge industrielle, ce qui est un problème qui commence à se poser pour certains moteurs de recherche. Nous sommes en ce qui nous concerne très sereins, compte tenu de notre technologie 64 bit très performante. Je crois également beaucoup aux moteurs de recherche verticaux, sur un pays, une thématique ou un support comme les blogs ou autres. La solution est certainement de "morceler" le Web et les moteurs pour mieux traiter et mettre en œuvre de la meilleure façon possible une technologie déjà pertinente pour amener des résultats toujours plus précis...

Pour ma part, je crois beaucoup à la notion de "[sérendipité](#)" (ou "caractéristique d'une démarche qui consiste à trouver quelque chose d'intéressant de façon imprévue, en cherchant autre chose" selon Wikipedia). C'est typiquement une démarche que nous avons tous sur le Web. Et le moteur Exalead propose les outils pour la mettre en œuvre. Nous venons d'ailleurs de voir acceptés deux brevets déposés en 2000 en Europe et aux Etats-Unis à ce sujet. Signe des temps : des responsables de moteurs majeurs commencent à parler de cette notion... Nous ne devons donc pas être totalement dans une voie sans issue... :-)

Que faut-il mettre en œuvre aujourd'hui pour devenir un acteur majeur du monde des moteurs de recherche dans le monde ?

Avant tout il faut du temps... Pour faire un bébé, il faut 9 mois. Désolé, mais le fait de mettre 9 femmes enceintes en parallèle ne crée pas un bébé en 1 mois. C'est aussi simple que cela. Bien sûr, l'investissement financier pour créer un moteur à forte audience mondiale est colossal, mais c'est avant tout une question de temps, de compréhension du domaine, d'acquisition de connaissances, etc. Cela explique pourquoi Microsoft n'a pas sorti, dès le départ, un "Google Killer". Ce n'est pas si simple que cela... Pour notre part, nous commençons aujourd'hui à obtenir de bons résultats en termes de pertinence sur le Web alors que c'est notre métier depuis 9 ans maintenant... Les réflexions notamment en termes de *ranking* et de *despamming*, sont longues à appréhender, complexes à mettre en œuvre et doivent évoluer tout le temps... Et il faut grandir avec le marché, la demande et les modèles économiques qui apparaissent.

Mais un moteur de recherche, c'est avant tout un ensemble de choses. La popularité (PageRank) a eu son heure de gloire mais, aujourd'hui trop spammée, elle est en perte de vitesse... Un moteur, c'est comme en cuisine, un plat concocté par un chef : c'est avant tout une recette qui lie et marie au mieux un ensemble d'ingrédients de bonne qualité. Cela demande un savoir-faire important...

Parlons de Quaero et de la perte d'AOL France...

En ce qui concerne AOL France, leur passage à la technologie de Google était prévu de longue date et nous les avons même aidé à effectuer cette transition. AOL France a donc suivi une directive venant directement des Etats-Unis. C'est ainsi, on n'a pas vraiment eu le choix, et Benjamin Faes, le Directeur General d'AOL France a récemment dit tout le bien qu'il pensait de notre technologie. On peut noter en effet que les statistiques montraient que le passage d'AOL à la technologie Exalead avait amené, à l'époque, un énorme trafic venant de non-utilisateurs du portail et du FAI AOL. Soit des gens qui avaient AOL à la maison et qui continuaient à utiliser le moteur au bureau, soit des personnes qui aimaient bien, tout simplement, la technologie et l'interface. On verra ce que l'avenir dira sur la décision d'AOL monde. Ceci dit, nous nous sommes quittés en très bon termes avec AOL et continuons à travailler sur de futurs projets avec eux...

Pour ce qui est de Quaero, on ne peut en fait que déplorer l'absence de communication qui a accompagné le projet depuis son début. Certains ont même voulu le faire passer pour un "Google Killer", ce qu'il n'était pas du tout. Mais peut-être y avait-il une volonté politique derrière. C'est pourtant un très beau projet sur lequel nous avons travaillé et investi près de 300k€. Cela nous a valu une exposition médiatique importante, mais pas toujours très positive également. Pourtant, il me semble crucial et nécessaire que ce programme continue après les élections présidentielles car de nombreuses PME y ont investi du temps et ont basé certaines stratégies dessus. Il serait dommage que tout s'arrête maintenant car Quaero n'est pas, comme certains pourraient le penser devant l'absence d'infos communiquée, une coquille vide... Maintenant, toutes les décisions ne sont pas entre nos mains... Quaero nous aura en tout cas permis de rencontrer plusieurs PME innovantes avec qui la discussion a été passionnante et pourrait déboucher sur des projets et produits dans l'avenir. Cela n'est d'ores et déjà pas le moindre des mérites de ce projet...

Quelles sont les nouveautés proposées par Exalead dernièrement ?

A part notre blog "corporate", que nous désirions réaliser depuis longtemps et qui est enfin en ligne, nous avons lancé il y a peu un affinage sur les blogs et les forums de discussions web. Lors d'une recherche, lorsque l'algorithme décide que c'est intéressant, un message propose un affinage sur l'un ou l'autre domaine (mais les deux sont également disponibles dans la "zapette" affichée à droite de la page de résultats).



The screenshot shows the Exalead search engine interface. At the top, there's a search bar with 'abondance' entered. Below the search bar, there's a navigation menu with 'Web', 'Images', 'Wikipédia BETA', and 'Plus >'. A search button labeled 'Recherche Web' is visible. The search results show 'Résultats 1-10 sur environ 526 344 pour abondance'. A prominent message asks: 'Voulez vous chercher dans les Blogs qui parlent de abondance?'. To the right, there's a sidebar titled 'Préciser la recherche' with sections for 'Termes associés' (listing 'Chapelle d'Abondance', 'Portes du Soleil', 'Lac Léman', 'Corme d'abondance', 'Haute-Savoie'), 'Type de site' (listing 'Blog', 'Forum'), and 'Multimedia' (listing 'Audio', 'Vidéo', 'RSS').

Les blogs et les forums Web sont automatiquement reconnus grâce à nos algorithmes de détection et d'apprentissage basés sur la reconnaissance du contenu. Nous croyons plus à cette méthodologie d'affinage des requêtes (proche de la "sérendipité" dont je parlais tout à l'heure) qu'à des approches comme celle de Google qui consiste à mettre en ligne un moteur strictement basé sur les blogs, par exemple. Nous proposons dans la zapette plusieurs types d'affinages :

- Par mots clés connexes
- Par support : Web, forums, blogs
- Par contenu : les pages qui contiennent une vidéo, une image, etc.
- Par format : PDF, Word, RTF, etc.
- Etc.

Cela permet de proposer à l'internaute un véritable choix pour aller plus loin dans ses requêtes. Chaque support correspond à une attente : les blogs pour l'actualité "chaude" et le temps réel, les forums pour l'avis des internautes et des consommateurs, des communautés, etc. Nous essayons de proposer une vision la plus large possible et les meilleurs outils pour l'affiner au mieux... Il s'agit

finalement d'un usage assez nouveau et non compartimenté de la recherche d'information sur le Web.

Un moteur sur Wikipedia est également disponible...

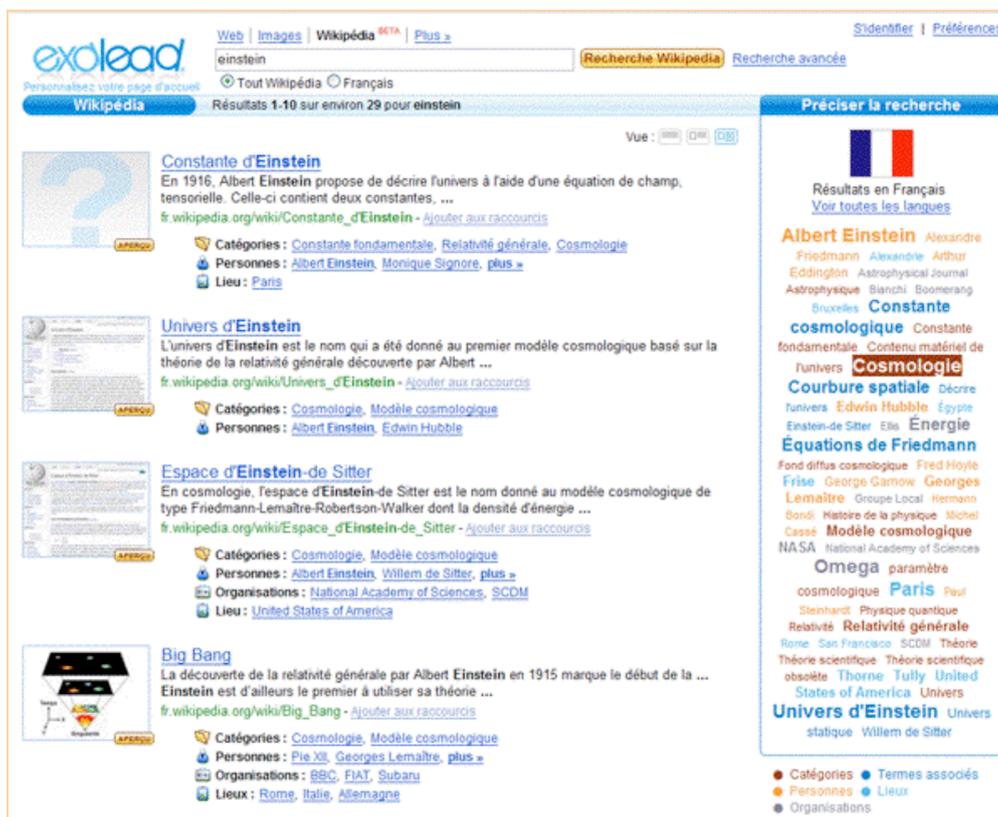
Oui, nous avons mis en place depuis quelques jours un moteur de recherche spécifique sur l'encyclopédie libre Wikipedia (disponible grâce à l'onglet dédié au-dessus du formulaire de saisie des requêtes du moteur). Nous avons beaucoup interagi avec les équipes de la fondation Wikimedia en France pour la mise au point de cet outil que je trouve personnellement vraiment fantastique. Pour le lancement, le moteur vertical sort simultanément en version allemande, française, espagnole, italienne et néerlandaise. L'utilisateur a la possibilité de rechercher au sein de toutes les versions linguistiques de Wikipedia simultanément, mais il peut également restreindre sa recherche à une localisation géographique (une langue) en particulier.

A chaque résultat de recherche listé correspond un article de l'encyclopédie. La zapette habituelle d'Exalead a été entièrement relookée pour l'occasion. Elle est maintenant présentée sous forme de nuage de tags dynamiques avec un code couleur selon qu'il s'agit :

- D'emplacements géographiques
- De personnes célèbres
- De noms d'organisations
- De groupes de mots reliés à la requête
- De catégories de Wikipédia.

Un clic sur ces termes modifie instantanément les résultats pour tenir compte de la précision apportée. Le nuage de tags se modifie à son tour. On peut cliquer sur plusieurs termes à la suite. Chaque résultat est accompagné d'une image extraite de l'article concerné. Chaque résultat présente un extrait textuel représentatif de l'article correspondant. Chaque résultat comporte également directement sous l'extrait de l'article des informations sur les catégories wikipédia, personnes célèbres, organisations et lieux reliés à l'article. Un clic sur ces termes étend la recherche au terme cliqué, amenant ainsi l'utilisateur à naviguer librement à travers l'encyclopédie.

The screenshot shows the Exalead search engine interface. At the top, there's a search bar with 'einstein' entered and a 'Recherche Wikipedia' button. Below the search bar, there are tabs for 'Web', 'Images', 'Wikipédia', and 'Plus'. The main content area displays search results for 'einstein' on Wikipedia. The first result is 'Albert Einstein', with a thumbnail image and a brief description. Below the description, there are tags and categories: 'RelativitéQuantique', 'Physicien allemand', '1955 à Princeton, New Jersey, États-Unis', 'Allemagne', and 'États-Unis'. The second result is 'L'Intersection Einstein', with a thumbnail image and a brief description. Below the description, there are tags and categories: 'Roman de science-fiction', 'Roman américain', 'Roman paru en 1967', 'Samuel Delany', and 'États-Unis'. The third result is 'Convention de sommation d'Einstein', with a thumbnail image and a brief description. Below the description, there are tags and categories: 'tenseur', 'Albert Einstein', 'Relativité générale', and 'Albert Einstein'. The fourth result is 'Condensat de Bose-Einstein', with a thumbnail image and a brief description. Below the description, there are tags and categories: 'Albert Einstein', 'Phase', 'Universit  de Leyde', 'Boulder, Colorado', and 'États-Unis'. On the right side of the interface, there is a 'Préciser la recherche' section with a language selection menu and a list of related terms and categories, such as 'Albert Einstein', 'Berlin', 'Isaac Newton', 'Maxwell', 'New York', 'Niels Bohr', 'Paris', 'Schloss Einstein', 'Stephen Hawking', and 'USA'. A legend at the bottom right explains the color coding for the tags: red for categories, blue for associated terms, orange for people, green for locations, and grey for organizations.



D'un point de vue "technologique", voici ce qu'il y a derrière :

- Les différentes versions linguistiques de Wikipédia (qui sont toutes codées légèrement différemment) ont été spécialement « parsées » pour sortir un format homogène (XML).
- Les images de chaque article ont été extraites.
- Les tags présentés pour chaque requête sont fabriqués grâce à trois techniques différentes :
 - * Les entités nommées de lieux, personnes, et organisations sont extraites à partir de dictionnaires.
 - * Les groupes de mots liés à une requête sont générés par des algorithmes statistiques.
 - * Les catégories Wikipédia sont extraites directement à partir de chaque article.

Enfin, un ranking spécial a été mis au point et affiné pour tenir compte des particularités de l'encyclopédie. D'autre part, le « snippet » de texte présent dans chaque résultat est généré d'une manière spécifique à l'encyclopédie afin de montrer le fragment de texte le plus représentatif possible. L'image de droite propose enfin soit la copie d'écran globale de la page, soit l'image qui nous semble la plus représentative de son contenu.

Nous espérons que ce nouvel outil proposera une nouvelle manière réellement innovante pour effectuer ses recherches dans cette énorme encyclopédie en privilégiant une navigation de type "convergence-divergence" entre l'information structurée et non structurée comme peut le proposer le Web.

Comment est constituée l'audience d'Exalead aujourd'hui ?

Nous avons deux cibles majeures assez bien déterminées : les baby-boomers de 40-50 ans qui désirent une certaine structuration de l'information, notamment dans leurs recherches et les jeunes de 18-35 ans qui aiment le côté "fun" de notre interface, la navigation "ipod friendly", la zapette. C'est notamment pour eux que nous avons créé les nuages de mots clés sur l'outil spécifique Wikipedia et mis en place le mouvement "LeMarr" (<http://www.lemarr.fr/>) ou "Mouvement Anti Recherche Relou". Le site propose des vidéos virales, des buzz, des "trucs" marrants. Notre première tentative pour "rajeunir" notre image, qui était un peu austère jusqu'à maintenant, est passée par le logo qui a subi un coup de jeune. On propose également pas mal de tutoriaux en Flash (<http://www.exalead.fr/account/demoSearch>) qui sont très appréciés. Bref, tout cela contribue à donner une image plus "jeune" de nos produits car cette cible est également importante pour nos partenaires médias...

Quels sont vos projets ?

Il y a beaucoup de choses dans nos cartons pour les mois qui viennent :

- De nouveaux affinages verticaux similaires à ce que nous proposons sur les blogs et forums vont voir le jour prochainement, de façon échelonnée dans le temps.
 - Une nouvelle recherche d'images sur un milliard de documents, donc un index similaire à celui de Google, sera bientôt en ligne également, avec quelques belles surprises sur les façons de rechercher l'info, mais je ne peux encore rien dire à ce sujet...
 - Une nouvelle version du moteur très "Web 2.0", toujours basée sur un nouveau concept, est également en développement... Il est un peu tôt pour en parler, mais nous sommes très confiants...
 - Et d'autres moteurs verticaux dont je ne peux là encore pas vous parler dans l'immédiat !
- Tous ces projets sont bien entendu prévus à court terme, en 2007 voire même très prochainement...

Merci, François Bourdoncle, pour vos réponses...

Interview menée dans les locaux d'Exalead à Paris le 12/04/2007