NETCHERCHEUR

N°1 – Avril 2004 www.netchercheur.com
La lettre des professionnels de l'information et de la veille - 100% Recherche 100% Pratique 100% Veille

SOMMAIRE

LE FAIT DU MOIS :

LA FIN DE GOOGLE ?

Qui n'a pas interrogé Google ces dernières 24 heures ? En 67 mois, il s'est imposé comme LE moteur de recherche sur Internet. Né en septembre 98, il revendique aujourd'hui plus de **4 milliards de pages web indexées**, loin devant son outsider, Alltheweb et ses 3 milliards de documents HTML référencés. Côté français, seul Exalead pourrait techniquement rivaliser avec le moteur n°1. Ce moteur, conçu par l'ancienne équipe française d'Altavista, s'apprête à finaliser son premier milliard de pages web. Positionné sur le marché des moteurs pour Intranet, Exalead ne cherche pas à concurrencer Google et ce pour des raisons de coûts. Car, un expert américain a récemment évalué le coût du crawling de **50 millions de pages web à 100 000 $ par mois !**

Le leadership de Google semble fait pour durer. Ce moteur issu de l'université de Standford a pris une telle ampleur - 76% des recherches américaines passent par lui - qu'on peut raisonnablement parler de Googlemania. Pour preuve, un énième livre est paru en mars : *Tout sur Google*. Il fait suite à l'ouvrage d'Olivier Andrieu portant exclusivement à ce moteur. L'an dernier, un prix Pulitzer du New York Times s'interrogeait même : Google était-il Dieu ? Il lui faut reconnaître un certain don d'ubiquité. Alors que d'autres moteurs se fourvoyaient dans la vogue des portails, Google s'est imposé en misant tout sur la technologie.

Après des années de collaboration, Yahoo a rompu avec Google le 10 mars et s'est doté d'une technologie maison provenant d'Inktomi. Avec son interface surchargée, le pionnier des portails Yahoo déconcerte les nouveaux internautes. A l'inverse, fort d'une page d'accueil dépouillée, Google a su séduire les néophytes de l'internet tandis que ses multiples options de recherche avancées réussissent à conquérir le public averti des professionnels de l'information, ces netchercheurs. Résultat : **Google traite chaque jour plus de 200 millions de requêtes et compte en Europe 55 millions d'utilisateurs** .

Fort de ce succès, le moteur américain creuse son écart et expérimente sans cesse de nouvelles fonctionnalités dans son Labs. C'est là que sont d'abord nés les alertes par mail, l'indexation des journaux français, les suggestions orthographiques, la barre de recherche et la presse personnalisée. Demain, des requêtes pourront être lancées à partir d'un téléphone portable. A son zénith, ces efforts de R&D retarderont-ils la chute de Google ? (la suite p 2)

LA FIN DE GOOGLE (suite)

Qui se souvient qu'Altavista était hier considéré comme le meilleur moteur? L'effet de mode est démultiplié sur Internet. Si Google semble aujourd'hui à son paroxysme tant il fait l'unanimité, demain les nouveaux internautes oublieront peut-être jusqu'à ce nom. Ce déclin pourrait provenir d'abord *des tenants des libertés publiques*. Déjà, des voix dénoncent les cookies de Google, valables 34 ans ! D'autres critiquent la base de données de géolocalisation de Google qui permet de tracer l'origine d'une requête. Cette association entre requête, IP et localisation géographique posera aussi problème aux professionnels de l'information, soucieux de discrétion. D'abord séduits par les milliards de pages indexés, ces netchercheurs sont ensuite exaspérés par le nombre de réponses, un volume humainement indigeste même pour le meilleur des analystes. De plus, les enjeux commerciaux du référencement à Google faussent les résultats du moteur, en tout cas pour les premières dizaines de réponses, les plus consultées. Il y a quelques mois, un journal informatique a publié une méthode pour obtenir plus de 100 000 référencements sur Google. Un autre point noir du moteur résulte des 4 milliards de pages crawlés (pages rapatriées vers le serveur de Google). Dans cette marée d'information, la gestion de sa base d'index (où sont stockés toutes les références aux pages visités) pose un problème essentiel.

La mise à jour de cette base, surnommée la danse de Google, ne s'effectue qu'une fois par mois. L'an dernier, un spécialiste américain des moteurs avait relevé que cinq semaines pouvaient s'écouler entre deux visites d'un même site. En pratique, cela signifie qu'une page web affichée sur Google peut remonter à plus d'un mois. De ce fait, Google devra revoir, tôt ou tard l'architecture distribuée de son moteur, qui date quand même de 1990 !

©Google

Aujourd'hui, elle se partage entre 13 points (tous aux Etats-Unis sauf un en Irlande à Dublin) qui centralisent les données recueillies par une batterie de plus de 10 000 serveurs répartis sur toute la planète. Hier, Google avait détrôné Altavista avec une technologie plus performante générant des réponses plus pertinentes. De nombreux prétendants, dotés de solutions innovantes, espèrent devenir LE moteur du futur. Qui sera demain le nouveau Google ? D'ici là, les netchercheurs exploiteront toutes les alternatives au moteur américain.

Demain, il y aura-t-il encore quelqu'un pour interroger Google ? EC

ETUDE SUR LES REQUETES SUR LES MOTEURS

D'après une étude de Onestat, publiée en février, plus de la moitié des internautes emploient au moins 2 mots clés lors d'une recherche sur les moteurs. Mieux, un quart des connectés (25,61%) formule leurs requêtes avec 3 mots clés. Les netchercheurs se reconnaîtront dans les 9% qui expriment leurs besoins avec plus de 4 mots-clés.

Depuis avril 2003, Onestat constate une augmentation de 3,36% des requêtes de deux mots. Les recherches avec un seul mot clé ont baissé de 5,74%. Ces néophytes représentent aujourd'hui encore près de 20% des connectés. Ces statistiques ont été obtenues à partir d'un échantillon de 2 millions d'internautes, provenant de 100 pays différents.

LE MOT DU MOIS : OPEN WEB

L'Open web représente l'ensemble du web indexé par les moteurs de recherche. Cette expression s'oppose au désormais célèbre web invisible. Depuis quelques mois, un débat oppose aux Etats-Unis les tenants des sources payantes aux partisans du web gratuit, accessible via les moteurs. Une nouvelle génération de professionnels de l'information, certains bardés de diplômes, considère que le web ouvert (ou visible) suffit.

Les restrictions budgétaires d'un hôpital public de Denver sont à l'origine de

ce débat. La résiliation des abonnements à des revues scientifiques a provoqué un tollé dans le milieu de l'information. Des relais des sources payantes mentionnent le cas d'un chercheur de l'université Hopskin qui aurait réussi à éviter le décès d'un patient si le médecin avait consulté les bases spécialisées.

VOILA DU NOUVEAU

Le 15 mars, le site Secret2moteurs a annoncé la nouvelle version de Voila. Le moteur de France Telecom/Wanadoo a complètement revu son interface. Il abandonne son code couleur jaune-orange au profit du bleu-orange, moins criard. La nouvelle page d'accueil est plus aérée grâce à une taille des caractères agrandie. Mais, ce lifting s'avère superficiel. La technologie du moteur n'a pas été modifiée. Aucune nouvelle fonction de recherche n'est apparue.

TEST DU MOIS : FLOW PROTECTOR 3.0 CONTRE LES LOGICIELS ESPIONS

97% des ordinateurs sont contaminés par des spywares, d'après l'éditeur Checkpoint. Ces logiciels espions récupèrent des données personnelles de l'internaute qu'ils transmettent à leur insu à des tiers. Ce nouveau fléau, plus inquiétant que le spam, constitue une véritable menace pour les professionnels soucieux de préserver leur anonymat. On distingue habituellement deux types d'espionciels : le spyware proprement dit, semblable à un cheval de Troie, et les adwares, développés à des fins publicitaires. Les programmes de partage de fichiers (peer to peer) pullulent d'adwares.

Il en existerait plus d'un millier. Aux Etats-Unis, ces logiciels espions ont pris une telle ampleur qu'un projet de loi, Spyblock (blocage d'espion), a été déposé fin février au sénat.

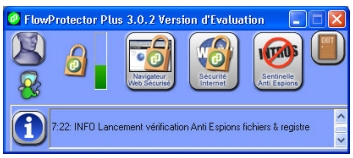
C'est sur ce créneau porteur que se positionne l'éditeur français Checkflow. A la différence des antivirus, les solutions de Checkflow analysent les activités des programmes installés et détectent les scripts java malicieux.

Netcherchercheur a pu tester avec succès une version d'évaluation de 30 jours de Flowprotector 3.0, téléchargeable depuis le site de Checkflow. Ce logiciel intègre trois grandes fonctions. La sentinelle anti-espion scrute en permanence le disque dur et la base de registre à la recherche d'espionciels. Une interface intuitive facilite la configuration de cette sentinelle. Flowprotector peut aussi éliminer automatiquement toutes les traces laissées par l'internaute (cookies et fichiers temp). Il scanne aussi tous les ports ouverts lors d'un connexion. Sa troisième fonction, le navigateur web sécurisé, s'active dès qu'une page sécurisée s'affiche. Ce browser permet aussi d'effacer l'historique des sites visités et accepte la consultation d'un carnet de favoris importés d'un autre navigateur.

Ce navigateur donne également accès, via le Flowclub, à quatre moteurs de recherche (Google, Altavista, Excite et Lycos). Une fonction de traduction de page web est aussi disponible avec Systran.

Globalement, Flowprotector va beaucoup plus loin que Spybot ou Adware qui ne détectent que les logiciels espions recensés.

Fondé en 1999 par des ingénieurs du CNRS et du MIT à Boston, Checkflow a signé un partenariat avec le célèbre éditeur de logiciel anti-virus McAfee. Ils commercialisent I-Munity PC.



FEEDSTER : LE GOOGLE DES WEBLOGS

Depuis la fusion entre Feedster et RSS-Search l'an dernier, Feedster s'impose comme le Google des Weblogs pour reprendre l'expression du Wall Street Journal. Uniquement focalisé sur les weblogs (alias blogs), Feedster se veut plus pertinent que le moteur numéro 1. Il indexe quotidiennement les grands services de blogs ce qui n'est pas le cas de Google.

Ses nombreuses fonctions avancées (fonctions booléennes, recherche par date ou weblog ...) faciliteront l'exploitation de l'univers des weblogs. Tout comme les forums de

discussion et les chats, les blogs véhiculent beaucoup de rumeurs. C'est aussi là où fourmillent des experts, accessibles en quelques clics. www.feedster.com

L'ASTUCIEUX SOOPLE
Soople énumère en une page toutes les fonctions avancées de Google. A partir d'une seule interface, le netchercheur pourra interroger la presse, trouver des images, voir les liens pointant sur un site, définir un terme…. A ce jour, Soople n'existe qu'en anglais, à intégrer dans ses favoris. www.soople.com

YAHOO NEWS SEARCH 2.0
La guerre des moteurs rebondit sur les actualités. Mi mars, Yahoo a lancé une nouvelle version de son moteur sur les articles de presse. Yahoo News Search 2.0 traite désormais chaque jour 7000 sources d'information contre 4500 journaux auparavant. Il dépasse donc largement Google et ses 4000 sources. De plus, le nouveau Yahoo News offre de meilleures fonctions de recherche.news.yahoo.com

DEUX ESCROCS DE GOOGLE ARRETES
Le 16 mars, le FBI a arrêté un jeune programmeur californien. Ce dernier avait tenté d'extorquer 100 000 $ à Google. Il menaçait de diffuser un logiciel provoquant des bugs lors des clics sur les bandeaux publicitaires du moteur. En cas de refus du moteur, ce développeur voulait vendre son programme aux sociétés de spam, jugement le 8 avril.
Un Hollandais a été appréhendé par le FBI à New York pour avoir vendu des fausses actions de Google, le 5 mars. Le montant de l'escroquerie s'éleve à 2,8 millions de $. Il risque 30 ans de prison.

LE MOIS PROCHAIN : Nouvelles rubriques dans NETCHERCHEUR : L'expert du mois et Veille sectorielle (concurrentielle pharma, techno, juridique, sociétale…). 16 pages

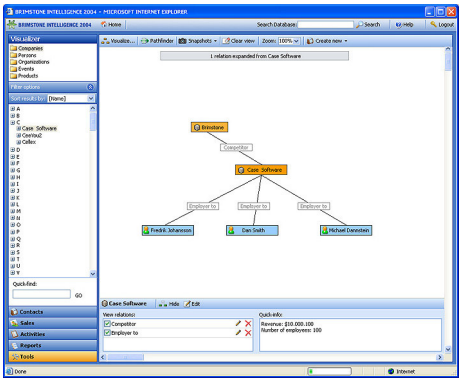
LE WEB FOUNTAIN D'IBM
« Le fossé entre une entreprise et les évènements en cours qui peuvent l'affecter va s'agrandir… Il devient impératif pour les sociétés de réduire ce fossé pour maintenir sa compétitivité » explique le site du Web Fountain. Lancé il y a 4 ans, le programme de recherche Web Fountain d'IBM vient de se finaliser. Une présentation des produits issus de cette R&D a été effectuée en février 2004 dans le centre de recherche d'Almaden.
Ce projet a permis la création d'outils de traitement des informations du web: crawler, moteur d'indexation et outils de filtrage de l'information. Aujourd'hui the Big Blue loue l'accès à son moteur de recherche capable de rapatrier des térabits de documents du web et de les organiser pour être intelligible par l'entreprise.
En savoir plus www.almaden.ibm.com/webfountain

LES CHINOIS CONTRE GOOGLE
« Mon travail est de botter les fesses de Google » expliquait au magazine Forbes Zhou Hongyi, créateur du moteur chinois 3721 (filiale de Yahoo). Il s'est associé à deux autres moteurs chinois, Baidu et Zhongsou, pour contrer l'arrivée de Google dans l'Empire du milieu fin février.

L'EUROPE LANCE UN PROJET SUR LE WEB SEMANTIQUE
La communauté européenne a lancé le programme SEKT (Semantic Knowledge Technologies). D'une durée de 36 mois, ce programme réunit 12 centres de recherche en Angleterre, Allemagne et Espagne. Annoncé le 22 mars, le labo British Telecom Exact va coordonner ce projet à 12 Millions d'Euros.
Sur le plan technique, SEKT vise à concevoir un moteur de recherche plus pertinent que les moteurs plein texte de rigueur sur le web. Il sera capable, grâce à la technologie sémantique, de filtrer en temps réel les réponses aux requêtes. En amont, un extracteur terminologique pourra reconnaître les noms de personnes, de produits ou de sociétés. Ces données seront ensuite classées automatiquement par concept.

INTELLIGENCE 2.0 DE BRIMSTONE

Intelligence 2.0, logiciel d'analyse du suédois Brimstone, autorise la création de graphes relationnels sur la concurrence, les technologies ou sur d'autres thèmes. Basé sur des tableaux Excel, il oblige à structurer une veille et à définir les entités et leurs relations entre elles. De plus, grâce à sa dimension visuelle, il offre au décisionnel une vision globale sur un thème. Un test effectif a permis de vérifier ces points :



- L'importation d'un fichier Excel ou txt est simple
- Les entités prédéfinies intègrent des personnes, des sociétés et organisations, des produits, des évènements…
- Sept types de liens sont possibles entre 2 entités
- Des fichiers joints sont associables à chaque entité
- Un moteur de recherche consulte tous les éléments (entités, relations) contenus dans la base du logiciel
- La fonction de Crosstabs automatise un comparatif entre 2 entités et génère un tableau de synthèse exportable vers les formats bureautiques ou HTML
- La fonction Infopacks autorise l'exportation de tous les éléments associés à une entité
- Les graphes de synthèse s'exportent au format jpeg et s'intègrent ainsi à des présentations powerpoint ou des rapports de synthèses

POUR QUOI FAIRE ? Intelligence 2.0 s'adapte tout autant à la veille concurrentielle que technologique. L'espace *Research Area* permet de l'adapter à d'autres types de veille.

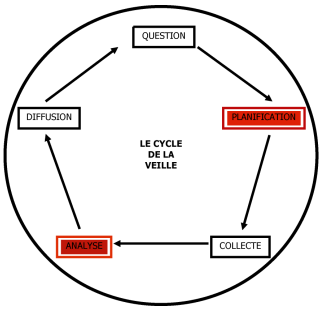
POUR QUI ? Le logiciel de Brimstone se destine avant tout aux analystes. En amont, il exige un investissement en temps pour l'apprentissage. Une connaissance préalable d'Analyst Notebook facilite grandement la prise en main du logiciel.

LE POINT FAIBLE : L'interface n'existe qu'en anglais et Brimstone n'a pas de bureau dans l'hexagone. Des négociations se poursuivent pour diffuser Intelligence 2.0 en France.

AVIS DE NETCHERCHEUR: Concurrent direct d'Analyst Notebook d'I2, il a déjà conquis 2500 veilleurs de part le monde. Son principal argument reste le prix : 560 Euros. Il est à la portée des veilleurs de PME. A la différence d'Analyst Notebook, Intelligence 2.0 n'exige pas une broche pour fonctionner. Fondé en 2001, Brimstone ne dipose pas encore de la richesse graphique d'Analyst Notebook : on ne y peut intégrer ses propres visuels encore moins de cartes géographiques.

Pour en savoir plus : www.brimstone.se

INTERETS D'INTELLIGENCE 2.0 DANS LE CYCLE DE LA VEILLE



PLANIFICATION : Intelligence 2.0 facilite la planification des plans de recherche (fonction Task)

COLLECTE : Importations depuis des fichiers Excel ou txt

ANALYSE : Grâce à sa fonction Visualizer, Intelligence 2.0 cartographie des réseaux. Il facilite aussi le classement des données obtenues lors de la collecte.

INTRODUCTION AUX STRATEGIES DE RECHERCHE

Savoir construire un plan de recherche, c'est la finalité des stratégies de recherche. Elles visent l'exploitation systématique des différentes strates de l'Internet.

Internet est devenue une telle *terra incognitae* qu'aucun expert ne se risque plus à évaluer sa taille. En 2001, une étude scientifique estimait à 4 milliards de pages le volume total d'Internet. C'est aujourd'hui ce que référence le seul Google. Certaines sociétés de marketing parlent aujourd'hui de 10 milliards de pages web. Mais, considérer cette masse d'informations comme un ensemble homogène constitue une vraie erreur. Entre le web, les forums de discussion, les listes de diffusion, le chat, Internet accumule depuis 25 ans de nouveaux médias, dernier en date les weblogs (voir tableau). Cette stratification du net oblige le professionnel à rationaliser ses recherches d'où la nécessité d'adopter des stratégies de recherche structurées.

Tableau : Les différents médias de l'internet classés par date d'apparition.

Date	Application	Nature du contenu
1972	Email	Communication, inaccessible aux moteurs
1979	Forum de discussion (usenet)	Communication, consultable sur Google Groups
1981	Liste de discussion	Communication, accessible seulement via un abonnement
1981	Liste d'information	Information, accessible avec un abonnement
1993	Premier navigateur web	Information et communication, une partie traitée par les moteurs
1996-1997	Migration des bases de données vers le web	Information, web invisible pour les bases payantes
1998	Apparition des pages perso	Communication, accès au travers des moteurs
1998	Vogue des portails	Information, accès souvent payant
1998	Développement de la presse en ligne	Information, accès direct ou via les moteurs d'actualité
1999	Peer to peer	Communication, nécessite un logiciel spécialisé
2000	Chat	Communication, pas de moteur à ce jour, mais un module d'acquisition peut être intégré
2003	Weblog	Communication, accessible via moteurs spécialisés comme Feedster

Basées sur plusieurs années d'expérience, six stratégies de recherche ont été formalisées. L'axe *moteurs généralistes* est le plus connu et le plus employé. Plus rares sont les professionnels qui exploitent systématiquement les stratégies *moteurs thématiques et métamoteur*. Adaptées à des recherches de longue haleine, les *axes source et communautés* s'avèrent particulièrement payantes. Enfin, la stratégie *expert* se révèle être la plus difficile mais la plus riche en terme de qualité. Trois autres axes de recherche résultent d'une combinaison des six premières stratégies. D'autres restent encore à formaliser.

Par la combinaison de plusieurs stratégies, le professionnel sera à même d'élaborer rapidement de complexes plans de recherche. Grâce à cette méthode, les netchercheurs concevront rapidement un plan de recherche structuré. Chaque mois, une nouvelle stratégie de recherche sera passée au crible sur les 9 prochains numéros de Netchercheur. A l'issue de cette méthodologie, le professionnel apprendra à choisir la bonne stratégie selon le délai imparti. Ce facteur temps guidera ses choix selon que la recherche soit express ou de longue haleine. Le mois prochain, cette rubrique développera la première stratégie, l'axe *moteurs généralistes*, la plus simple à mettre en oeuvre. EC

BREVES

ANACUBIS PATENT

Anacubis vient de lancer son plug-in Destop Intellectual Property Analysis, un ajout à son programme de cartographie adapté à la visualisation de brevets. La licence de cet outil est commercialisée 750 $ pour 12 mois. Depuis un an, Anacubis est accessible sur la base brevet de Questel-Orbit. (www.anacubis.com)

NOUVELLE INTERFACE POUR DIALOG

Etes-vous prêt pour un profond changement ? C'est le du nouveau Dialog, l'un des plus gros fournisseurs de données. Netchercheur a pu tester cette nouvelle interface. Les premiers tests ont été laborieux. Pendant plusieurs semaines, il était impossible d'interroger les archives sur un mois. Le paramétrage des recherches sauvegardées révélait des bugs. Il a fallu s'y prendre à plusieurs reprises pour combiner des thèmes et des titres de presses. Une fois sauvegardées, ces requêtes personnalisées avaient du mal à s'afficher correctement. La navigation entre la recherche et les alertes n'est pas toujours aisée. Quant aux alertes par mail, la déception domine. Seuls les titres des articles sont envoyés par courriel sans hyperlien pour accéder directement aux articles.

LOOKSMART LANCE FINDARTICLES

Le 1 mars, l'annuaire américain Looksmart a lancé une nouvelle version de Findarticles. Le nouveau moteur stocke plus de 3,5 millions d'articles de presse, provenant de 700 sources US. A la différence de Google ou Yahoo, Findarticles conserve des archives (jusqu'en 1998) grâce à un partenariat avec Thomson Gale.

GOA SCRUTE LE WEB INVISIBLE

"Closer Look trouvera tout ce que Google, Altavista et même les métamoteurs tels que Copernic ne peuvent trouver". Si cette affirmation de Goa technologie peut sembler présomptueuse, elle n'en est pas moins vraie. Le moteur de recherche crée par Goa technologie, société montréalaise, est plus performant que tous les autres systèmes, du moins sur le créneau du web invisible. Sumithra Jagannath, la présidente de Goa explique que Closer Look va "chercher l'information là où certains ne cherchent pas". Le web invisible comprend l'ensemble des documents, textes, vidéos, images et autres qui ne sont indexés ni par les outils de recherche classique ni par les annuaires comme Yahoo.
Ces données peuvent être non-indexables - comme les animations dynamiques - ou bien présentes dans les bases de données des gouvernements, universités et banques, ou encore tout simplement non-référencées.
"On estime que le web invisible est de 10 à 100 fois plus grand que le web visible" souligne Melle Jagannath. S'il existe d'autres firmes qui commercialisent des moteurs de recherche pour le web invisible, Goa se distingue par sa stratégie de commercialisation.
Au lieu de vendre le moteur tel quel, ils développent une interface personnalisée. Par exemple, Goa a développé pour les centres d'enquête de solvabilité une application permettant de connaître le profil financier complet d'un individu ou d'une entreprise. Le dernier point fort de ce produit est l'analyse, le traitement et l'organisation des résultats de recherche, ce qui économise un maximum de temps. ZATAZ
http://www.zataz.com/zatazv7/news.php?id=2443

KEEPMEDIA : 150 SOURCES POUR 5$

Le fournisseur de données, Keepmedia, vient de créer un nouveau service presse doté de 1540 journaux et magazines. Il s'agit surtout de magazines américains comme Newsweek, Forbes, US Today. Pour 5$, l'abonné disposera d'un accès illimité aux archives qui remontent à 1992. Mais il faut payer pour les articles récents.

EVALUER UNE SOURCE SUR INTERNET

Plus d'un demi million, c'est le nombre de weblogs aujourd'hui accessibles sur Internet. En un mois, près de 41 000 nouveaux weblogs ont vu le jour. Sur Internet, chaque internaute peut produire de l'information sur Internet. Mais comment évaluer une source d'information diffusée sur les weblogs ? Sur quels critères?

Qu'est-ce que l'information crédible sur Internet? De nombreux spécialistes de la documentation ont élaboré des grilles d'évaluation d'une source sur Internet. Les documents pdf foisonnent sur ce sujet.
« Une nouvelle forme de pollution s'installe un peu plus chaque jour sur le Web. Elle est due à la prolifération de données non vérifiées entraînant des effets indésirables voire néfastes. » détaillent 4 étudiants de l'Ecole des Mines de Saint Etienne. Ils préconisent la mise en place d'un service d'experts chargés d'évaluer les données du web. La vérification d'une information peut exiger un vrai travail d'investigation et peut exiger des compétences pointues. Une information du net peut masquer des finalités commerciales (promouvoir un produit ou nuire à un concurrent), défendre une cause politique quand elle ne s'intègre pas à un vaste plan de désinformation.
Un professionnel américain a identifié 4 groupes de critères pour évaluer une information sur internet. Le premier s'intéresse au média et à l'auteur de la page. Par exemple l'internaute s'intéressera au site qui l'héberge, à la fréquence des mises à jour ou son audience. Ensuite, 2ième groupe, il s'agira d'identifier le public du site et de la page pertinente. Le 3ième groupe de critères se penche sur la pérennité du site, son indexation éventuelle par les moteurs de recherche. Enfin l'ultime groupe s'intéresse aux hyperliens.

LES CRITERES D'EVALUATION D'UNE SOURCE

Groupe	Critères
1-Média et auteur	Audience, objectivité d'information, qui est l'auteur, qu'a-t-il écrit d'autres ?
2-Site de diffusion	Contenu global du site, pertinence du site par rapport à l'information à évaluer, expertise du webmaster
3-Pérénnite du site	Date de naissance du site, délai des mises à jour, structure du site, récompenses du site
4-Hyperliens du site	Liens pointant sur le site et les hyperliens extérieurs. Ces sites pointés sont-ils reconnus ?

Nota : Une autre grille d'analyse (en français) est proposée sur le site EDUCNET (www.educnet.education.fr/dossier/rechercher/évaluation1.htm)

COPERNIC TRACKER POUR LA VEILLE DE PAGES WEB
Lancé fin février, Copernic Tracke génère des alertes lors de changements de pages. Ce suivi peut être paramétré dans le temps : chaque heure, au quotidien voire toutes les semaines. Ces alertes sont ensuite diffusées par mail. Les modifications de la page apparaissent surlignées. Copernic Tracker est vendu 50$.

DIGIMIND : CHOISIR LES MEILLEURS OUTILS DE VEILLE
L'éditeur de solutions de veille diffuse depuis le 11 février une grille d'évaluation des logiciels de veille. Ce document recense sur 18 pages un ensemble de critères pour comparer les différentes plateformes de veille disponible. Chaque critère est classé dans l'une des 3 étapes du processus de veille (acquisition, traitement et analyse, diffusion). Une dernière catégorie couvre aspects d'administration et de sécurité de la plateforme.

ISCOPE & ATLANTIC INTELLIGENCE PARTENAIRES
Iscope et Altantic Intelligence (AI) ont formalisé un partenariat le 18 mars. Ce rapprochement entre l'éditeur de Keywatch et le cabinet de veille confirme une tendance initiée avec la fusion Startem/Datops, à savoir la convergence entre les éditeurs d'outils et les cabinets d'études. AI va exploiter Keywatch alors qu'Iscope bénéficiera de domaines d'expertise du cabinet d'IE de Mr Legorgus.
« Cette nouvelle solution très puissante [keywatch] nous permettra d'être plus efficace dans le recueil pour qu'ils se concentrent davantage sur l'analyse de l'information stratégique » témoigne Nathalie Spillmann, responsable du pôle veille et IE chez AI.

EXPLOITER LES ANNUAIRES ET GUIDES

Avril 1994, deux étudiants de l'université de Standford ont l'idée de classer par thème tous les sites intéressants qu'ils découvrent et lancent **Y**et **A**nother **H**ierarchical **O**fficous **O**racle. Le succès est immédiat : dès l'été 1994, plus de 10 000 internautes consultent Yahoo. Dix ans plus tard, ils seront 250 millions à visiter ce portail chaque mois. Mais, l'arrivée de Google en 1998 met à mal le leadership de Yahoo. Très vite, ce portail a dû intégrer un moteur de recherche, car sa fonction annuaire est peu employée. Qu'est ce qu'un annuaire ?
L'annuaire appelé aussi répertoire ou guide a la particularité de classer les sites web dans des catégories thématiques. En amont, des spécialistes de la documentation effectuent cette classification (voir encadré Nomade). Aujourd'hui, Nomade n'est plus mis à jour depuis plusieurs mois. Depuis son rachat par le portail Tiscali, l'équipe du guide a été réduite comme une peau de chagrin. Dans ses dernières semaines, Nomade s'appuyait sur une seule personne ! Les annuaires coûtent chers à maintenir en raison des ressources humaines qualifiées nécessaires. La logique financière a eu raison du principal annuaire 100% français qui laissent place à des versions locales de sites américains (voir Tableau 2). Le succès des moteurs au détriment des annuaires s'expliquent aussi par ses inconvénients.

NOMADE : UN (EX) ANNUAIRE DANS LE DETAIL
par Géraldine Gourbin, ancienne responsable de la documentation chez Nomade:

"En fait, les annuaires sont plus complémentaires que concurrents des moteurs de recherche. Un annuaire comme Nomade ou Yahoo ne peut pas tout couvrir, surtout les domaines très spécialisés. Dans ce contexte, l'utilisation d'un moteur est plus pertinente. Les moteurs de recherche génèrent un bruit phénoménal. C'est le revers du tout automatique. Nous ne prétendons pas à l'exhaustivité. Chez un annuaire, l'accent est mis sur la qualité et non la quantité. C'est sa particularité.
J'ai dirigé une équipe de trois documentalistes appuyés de 2 stagiaires, tous ont suivi une formation en documentation ou communication. Chaque semaine, nous recevons en moyenne 1500 soumissions de sites web. Nous effectuons le référencement de ces sites soumis. Une bonne partie est rejetée, une majorité est classée parmi les 8000 sous catégories. En tout, notre arborescence contient 65000 sites web. Ce chiffre correspond, à mon avis, à 80% des sites représentatifs du web francophone. Pour éviter la saturation d'une sous catégorie, dès qu'une section contient plus d'une cinquantaine de sites web, nous la subdivisons. C'est ce qui distingue Nomade d'un Yahoo plus labyrinthique.
Quant à l'utilisation du guide, la recherche par mot clé demeure l'approche la plus utilisée même si la navigation dans l'arborescence offre d'intéressantes possibilités. Cela permet entre autres de localiser tous les sites abordant un même thème".
NOTA : Depuis septembre 2003, l'annuaire Nomade n'est plus mise à jour.

Tableau 2 : Les principaux annuaires généralistes

Nom	Remarques	Site
Looksmart	Un nouvel annuaire commercial qui rivalise avec Yahoo. N'existe pas en français	www.looksmart.com
Nomade	Plus mise à jour depuis plusieurs mois	www.nomade.fr
Open Directory	Plus complet que Yahoo, incontournable	www.dmoz.fr
Yahoo	Un pionnier surestimé	www.yahoo.com

QUANTITE VERSUS QUALITE
Les annuaires présentent des inconvénients : ils pêchent par leur manque d'exhaustivité. Il arrive qu'une requête très spécialisée ne génère aucun résultat (voir Encadré Quand et comment interroger les annuaires). Par essence, l'annuaire privilégie le silence (absence de

résultats). Basés sur une sélection humaine, les guides se distinguent des moteurs de recherche comme Google ou Exalead au mode de fonctionnement radicalement différent. Un moteur repose à 100% sur des programmes, aucun humain n'intervient dans les résultats. A la qualité d'une évaluation humaine des annuaires, les moteurs opposent la quantité. L'internaute a plébiscité les moteurs. Toutefois, l'apparition de Dmoz rend caduque le distingo entre moteur et métamoteur. Lancé en juin 98, cet annuaire s'appuie sur des milliers de volontaires pour classer les sites. Très vite, il dépasse Yahoo en terme de volume de sites référencés (voir Yahoo contre Dmoz).

Yahoo contre Dmoz

	Nombre de sites classés	Nombre de catégories	Nombre de contributeurs
YAHOO	2 700 000	?	150 (estimation)
DMOZ	4 500 000	590 000	62 200

De plus, Dmoz autorise les sites web à exploiter gratuitement son annuaire. La traditionnelle séparation entre moteurs et annuaires devient de plus en plus caduque. Les deux genres fusionnent. Le répertoire de Google s'appuie en grande partie sur DMOZ. Il n'est pas le seul, plus de 300 moteurs (dont Altavista, Alltheweb, Lycos, Teoma) reprennent le classement de l'annuaire « freeware ».

En parallèle à la montée en puissance de Dmoz, on assiste à la multiplication des annuaires spécialisés à l'instar des moteurs de recherche thématiques. Tout comme avec le web invisible, aucun site ne recense tous les annuaires disponibles sur Internet. Le site Les Annuaires constitue un bon point de départ pour trouver un répertoire francophone. Cette nouvelle génération d'annuaires démontre bien que les répertoires sont loin d'être morts. EC

Des annuaires à découvrir

Nom	Notes	Site
Les annuaires	Recense plus de 2000 annuaires professionnels français, à découvrir	www.lesannuaires.com
Bubl Link	Annuaire britannique élaboré par des professionnels de l'information	www.bubl.ac.uk/link
Librarians Index to Internet (LII)	Annuaire américain conçu par des bibliothécaires américains	www.lii.org
Infomine	Plus de 100 000 ressources spécialisées destinées aux enseignants	http://infomine.ucr.edu

ASTUCE DU NETCHERCHEUR: pour trouver un annuaire spécialisé sur un thème donné, il faut lancer la requête entre guillemets « annuaire du thème » sur un moteur généraliste.

QUAND ET COMMENT EXPLOITER LES ANNUAIRES

- Pour cantonner ses recherches aux sites qualifiés
- Pour identifier des sources validés sur un thème (plan de sourcing)
- Pour découvrir les différentes catégories d'un thème
- Pour se familiariser avec un nouveau sujet
- Navigation dans l'arborescence de thèmes et sous thèmes pour découvrir des informations que l'on ne cherche pas à priori (serenpidity)
- Eviter les requêtes trop pointues sur les annuaires

85% DE L'INFORMATION EN ENTREPRISE RESTE INACCESSIBLE d'après une récente étude d'IDC. Les professionnels de l'information (veilleurs, responsables marketing, chercheurs et plus largement les décideurs) perdent entre 15 et 30% de leur temps à rechercher de l'information. Cela se traduit par la perte de 6 millions de $ par an pour une structure employant 1000 professionnels.

VERITY a racheté CARDIFF software pour 50 millions de $. Cardiff Software a développé des suites logicielles automatisant l'acquisition et la circulation de documents en internet. Le leader mondial des moteurs, Verity, poursuit cette politique de croissance externe après l'acquisition de l'activité Entreprise d'Inktomi.

IPSEN a choisi TEMIS. L'Institut Henri Beaufour d'IPSEN va intégrer la suite logiciel Insight Discoverer de Temis. Outre l'extracteur terminologique, cette suite comprend une fonction de classement automatique et un module de visualisation en plus de cartouches spécialisés dans le domaine pharma. Les équipes de chercheurs accéderont à plusieurs centaines de milliers d'articles scientifiques.

CINQ QUESTIONS A LINGWAY

1-QUI EST LINGWAY ?
Lingway est une société récente, née en août 2001. Elle a été créée par un groupe d'experts en ingénierie linguistique et documentaire. Son CA en 2003 a atteint 1,4 M€ pour 13 salariés.


B. NORMIER - PDG

2-QUELLE TECHNOLOGIE AVEZ-VOUS DEVELOPPE ?
C'est une technologie linguistique, plus particulièrement sémantique. Cela signifie qu'il y a dans le système des connaissances sémantiques explicites et externes aux données, qui ont été décrites par des experts linguistes. Par ailleurs notre technologie utilise aussi des méthodes statistiques, dont on ne peut de toute façon pas se passer (calcul de facteurs de pertinence, ranking, clustering etc.)

3-QUELS PRODUITS COMMERICALISEZ-VOUS
Notre produit principal est Lingway KM. Il s'agit d'une suite d'outils dédiés à la gestion de connaissances (Knowledge Management) et aux applications de fouille textuelle (Text Mining) et veille. Contenant une base de 100.000 concepts, Lingway KM traduit automatiquement la requête dans d'autres langues (recherche cross-language). En amont, Lingway KM peut crawler le web. Il peut extraire des entités, des phrases ou des terminologies d'un vaste corpus de données. Deux déclinaisons spécialisées de Lingway KM sont disponibles dans la pharma (Lingway Medical Dictionary Encoder) et les brevets (Lingway Patent Access)

4-A QUELS BESOINS REPONDENT-ILS
Lingway KM permet de mieux retrouver l'information interne et externe à l'entreprise, de mieux organiser cette information et surtout de lire plus facilement et de mieux comprendre cette information. Ces trois besoins (accès, organisation et lecture) doivent être traités simultanément. Il ne sert à rien de trouver plus d'information si l'on ne sait pas l'organiser et si l'on n'a pas le temps de l'analyser. L'INPI, le World Intellectual Property Organization à Genève, Questel-Orbit et l'INSEE ont intégré ce produit.

5-QU'EST-CE QUI VOUS DISTINGUE DE LA CONCURRENCE ?
Une approche résolument linguistique et sémantique, basée sur une ontologie (dictionnaires et réseaux sémantiques) adaptable. Cela n'exclut pas d'exploiter aussi les méthodes plus classiques de statistiques. Cette approche permet d'aborder des applications que les autres technologies ne peuvent prendre en charge car leur niveau d'analyse des questions et des textes est insuffisant.

LA TRADUCTION CONTRE LE TERRORISME

« Ils sont un maillon particulièrement vital de la sécurité nationale des Etats-Unis » écrivait Tech Review dans son article *La traduction dans l'age de la terreur,* article partiellement traduit par Courrier International du 24/02. On y apprend dans l'article en anglais que l'un des systèmes de gestion documentaire de la CIA ne pouvait traiter que les langues européennes. Les textes en arabe, russe ou chinois devaient être d'abord transcrit dans notre alphabet ! Le FBI n'est pas mieux loti. Son nouveau réseau information Trilogy, il a couté 600 millions de $ aux contribuables américaines, s'avère très limité sur le plan de la traduction. Partiellement déployée, Trilogy n'a pas été conçu pour traiter des langues exotiques (russe, chinois, arabe). Cette nouvelle fonctionnalité, rajoutée en cours de route, va rallonger la note, déjà dépassée de plusieurs dizaines de $ par rapport à son budget initial.

Le journaliste de Tech Review, Michael Erard, se focalise sur le nouveau National Virtual Translation Center (NVTC) basé à Washington, fondé l'an dernier. Sa création fait suite à de nombreux dysfonctionnements dans la traduction révélés lors du 11/9 : deux messages en arabe interceptés veille des attentats n'ont été traduit que le 12…

L'originalité du NVTC réside dans son approche, basée non sur une solution de traduction automatique (un vrai leurre) mais sur ses ressources humaines. Tout d'abord, le traducteur est avant tout un analyste et non plus un simple transcripteur : « le maître mot, ici, est l'analyse, et pas seulement la traduction » explique-t-on au NVTC. A ce jour, le noyau dur de 5 analystes/traducteurs compose cette nouvelle structure. Mais dans les prochains mois, près de 300 linguistes seront embauchés. D'ici 3 à 5 ans, NVTC sera à même de relier des dizaines de milliers de spécialistes via un réseau sécurisé.

Cette priorité donnée à l'humain rompt avec le tout informatique, de rigueur dans les années 90. On se souvient d'Intellink, l'intranet des 13 agences de renseignement américaines. Aujourd'hui, la technologie ne cherche pas à remplacer l'élément humain, redevenu central. En amont, des bases de données, des solutions de reconnaissance de langues et des archives des traductions antérieures facilitent le travail quotidien du traducteur/ analyste. De ce fait, la R&D dans la Traduction Assistée par Ordinateur s'oriente dans d'autres directions. La société Trados, fournisseur du FBI, élabore des outils personnalisables au traducteur qui offre aussi d'autres fonctionnalités comme le résumé automatique. Septique, le gourou des sources ouvertes, Robert Steele d'OSS, prédit que le NVTC rejoindra d'autres grands programmes fédéraux à la valeur douteuse…

AQUAINT : DES QUESTIONS SANS REPONSES

Ou se trouve Ben laden ? C'est le type de question auquel le programme de R&D Aquaint doit répondre. Une réunion de synthèse s'est déroulée à Tampa (Floride) en mars. Aquaint, aujourd'hui en phase II, vise l'élaboration d'un complexe système de questions-réponses à la Askjeeves. Il s'agit de concevoir un outil apte à répondre à des questions multiples posées par des spécialistes. Ou se trouve Ben Laden ? Il n'est pas à Paris.

SRD DANS L'ŒIL D'IN-Q-TEL

Dans son édition du 16 mars, le Wall Street Journal a consacré Systems Research & Development (SRD). Cette jeune pousse a très vite attiré l'attention du renseignement américain. Pourquoi ? La technologie élaborée par Jeff Jonas a de quoi séduire les agences US : il a conçu un logiciel capable de comparer les noms d'une base de données avec d'autres listes. Rien de nouveau pour certains, mais le programme de SRD n'exige pas que les bases soient divulguées à des tiers. On connaît la réticence des agences de renseignement à partager leurs données. Cette technologie a d'abord été exploitée par les casinos de Las Vegas. Les salles de jeu comparaient leurs listes des persona non grata aux réservations effectuées dans les hôtels de la région. Logiquement, In-Q-Tel a investi dans SRD en juillet 2002.

AGENDA

- **19-20 avril : Search Engine Meeting – La Hague (Hollande)**
C'est le rendez-vous annuel en matière de R&D sur les moteurs de recherche. C'est le passage obligé de toute nouvelle start-up dans le domaine. Netchercheur couvrira l'évènement. www.infonortics.com
- **21-22 avril : Library Information Show - Londres**
Plus de 130 exposants vont se presser à ce rendez-vous des professionnels de l'information et de la documentation. Plusieurs séminaires gratuits contenteront les spécialistes de la gestion documentaire. www.lishow.co.uk
- **4 mai : Les outils avancés de veille sur Internet** – Paris
Scip France organise une journée dédiée aux outils de veille: au menu présentation d'outils et retour d'expériences. www.scip-France.org
- **8 - 10 juin 2004 : I-EXPO 2004 – Paris**
Inutile de présenter ce incontournable salon de l'information numérique (ex IDT), tous les professionnels français de l'information et de la veille s'y retrouvent chaque année. www.i-expo.net
- **25 - 29 octobre 2004 : VSST 2004 -** Toulouse
Ce séminaire, organisée par l'IRIT, est très orienté sur la R&D en matière d'outils de veille, pour les spécialistes. http://atlas.irit.fr/COLLOQUES/VSST2001/manifs.html

A LIRE

* Tout réussir avec Google de F.Schneider, N.Blachman et E.Fredricksen (Ed First Interactive - 19,90 E)
C'est le troisième livre exclusivement consacré à Google paru en quelques mois. Celui-ci, écris par des salariés de Google, dévoile de nombreux trucs et astuces pour exploiter au maximum ce moteur numéro 1. Le néophyte de la recherche y apprendra comment mieux l'interroger tandis que le netchercheur avéré y découvrira surtout des points de détails comme Google Zeitgeist.

The Extreme Searcher's Internet Handbook de Randolph Hock (Ed Cyberage Book - 24,95$)
Cinq ans après son premier livre, Randolph Hock revient à la charge avec cet ouvrage dédié à la recherche sur Internet. Comme beaucoup d'autres livres sur le sujet, il se focalise sur huit moteurs de recherche et non sur les méthodes, malheureusement. En fait ce livre s'avère être un catalogue de ressources du web invisible. Mais son auteur ne sombre pas dans la Googlemania.

Puzzle n°10 : Fin mars, la revue espagnole de l'IE, Puzzle est paru long de 32 pages. Au sommaire du numéro 10 : on trouve des articles sur les brevets comme source d'intelligence, l'externalisation des services d'IE et les stratégies de recherche sur Internet (Netchercheur a rédigé cet article). Puzzle n°10 est téléchargeable gratuitement en PDF (www.revista-puzzle.com)

La nouvelle frontière : Google - Newsweek du 29-03-2004
L'hebdomadaire américain envoyé l'une de ses meilleurs plumes dans ce numéro spécial consacré à Google. Steven Levy, auteur de nombreux livres sur les nouvelles technologies, a rencontré les fondateurs du moteur. Un encadré se focalise sur les directions de recherche du moteur.

©Newsweek

ABONNEMENT

NETCHERCHEUR, c'est **LA** lettre mensuelle des professionnels de l'information, de la veille, 100% pratique, 100% recherche, 100% veille.

CHAQUE MOIS, c'est :

- L'actualité de la recherche sur Internet et des systèmes d'information
- Des dossiers de fonds pratiques
- Des méthodes détaillées et opérationnelles
- Des tests de logiciels pour optimiser ses recherches et ses veilles
- Des outils pour exploiter le web invisible
- Des procédures pour concevoir des veilles systématiques
- Des solutions pour mieux gérer son intranet et son information interne
- 16 pages par numéro - 11 numéros par an, parution tous les 20 du mois

POUR QUI ? Netchercheur intéressera en premier lieu les documentalistes, les journalistes, les veilleurs, les analystes et les praticiens de l'Intelligence Economique. Cette lettre sera un outil précieux pour tous les autres professionnels en prise quotidienne avec l'information : chercheurs, ingénieurs, avocats, responsables marketing et plus largement les décideurs.

✂----------

A retourner complété et signé à : EC PRESSES – 5 rue de DOUAI – 75009 PARIS

Nom : ..... Prénom :.....
Fonction : ..... Société : .....
Adresse : .....
Code Postal : ..... Ville : .....
Tel :..... Portable : ..... Email :.....

Je désire m'abonner à Netchercheur pendant un an (11 numéros)

- ☐ 147 Euros TTC Particuliers, PME-PMI, Secteur public (version papier, Europe)
- ☐ 248 Euros TTC Sociétés > 50 salariés (version papier, Europe)
- ☐ 347 Euros TTC Editeurs, Consultants IE, SSII, Grands comptes (version PDF, hors UE)

Une facture avec la TVA (2,10%) sera envoyée dès la réception du règlement

- ☐ Paiement par chèque à l'ordre de EC PRESSES
- ☐ Paiement par carte bleue

Type de carte : ❒ Carte bleue ❒ Visa ❒ MasterCard

N° Carte : |_|_|_|_| |_|_|_|_| |_|_|_|_| |_|_|_|_| Date d'expiration : /

Date et Signature : Cachet :

NETCHERCHEUR est une publication mensuelle de EC PRESSES, SARL DE PRESSE. RCS : Paris 451 095 764
COMMISSION PARITAIRE : EN COURS. ISSN : EN COURS. IMPRESSION : EC PRESSES
DIRECTEUR DE LA PUBLICATION ET REDACTEUR EN CHEF : Emmanuel CHANIAL, ACTUS : Maxime Paulet, Zataz EMAIL : contact@netchercheur.com