

ZONE ABONNÉS : Archives de la lettre "R&R"

Janvier 2000

■ Les responsables du Go Network, et par voie de conséquence du moteur Infoseek, ont déclaré que l'outil de recherche allait se consacrer dans les mois qui viennent à certains domaines uniquement, tournant autour des loisirs et du divertissement (Disney oblige). L'information de type professionnel risque donc d'être rapidement bannie très rapidement de cet outil de recherche.

■ Les résultats de Lycos pour sa partie Recherche Avancée (<http://lycospro.lycos.com/>) sont fournis par AllTheWeb. On commence à s'y perdre un peu sur cet outil qui propose des données en provenance de sa base propre, d'AllTheWeb, de l'Open Directory, de Direct Hit, etc. Nous essaierons d'y voir plus clair le mois prochain dans un dossier qui lui sera consacré dans cette lettre.

■ Yahoo! France et Nomade semblent avoir quelques problèmes avec leur syntaxe d'interrogation.

Sur Yahoo! France, depuis le début de l'année, si vous cherchiez, par exemple, un site sur la Guerre des Étoiles (le film), vous pouviez saisir la requête suivante, avec les guillemets : "guerre des étoiles"

Et, oh, surprise, Yahoo! France passait la main à Inktomi sans rien trouver...

Alors que la syntaxe :

+guerre +étoile

trouvait bien des sites et même une catégorie où l'expression était bien orthographiée :

"guerre des étoiles" !! Un autre essai avec :

"moteur de recherche"

ne trouvait rien non plus, alors que :

+moteur +recherche

fonctionnait, semblant bien annoncer l'ébauche d'un bug dans le système de recherche de Yahoo! France.

Le problème a été soumis à l'annuaire et, effectivement, corrigé quelques jours plus tard.

Remarquons ici la réactivité du service "production" de l'outil qui a non seulement corrigé le bug, mais qui m'a également envoyé un message pour m'en avertir...

Malheureusement, début février, un nouveau bug semblait apparaître. En effet, une requête sur :

"la centrale" (avec les guillemets) était équivalente à : la centrale (sans guillemets), ce qui induit énormément de bruit dans la recherche (grosso modo, toutes les fiches descriptives de sites qui contiennent le mot "centrale" !!).

Le bug a une fois de plus été signalé. La suite le mois prochain...

Du côté de Nomade, il semble y avoir quelques problèmes également.

Lorsque l'on propose la requête suivante :

disque vinyl

(sans les guillemets)

Nomade trouve les expressions :

disque vinyl

et

disques vinyl

mais pas :

disque vinyls

et

disque vinyle

donc, a priori, disque trouve disques mais vinyl ne trouve ni vinyle ni vinyls. Ça ressemble fort à un problème de troncature, car avec un mot ça marche mais pas avec l'autre.

Le problème a été signalé à Nomade, qui a effectivement conclu lui aussi à un bug (bravo à eux également pour la réactivité : réponse dans la journée !) et travaillent sur sa correction.

■ Rappelons que, si Voila.fr va chercher dans son annuaire propre (l'ex-QuiQuoiOù) les premiers résultats à une requête, le site international Voila.com utilise, pour sa part, l'annuaire Looksmart pour cette première investigation.

■ Nouvelles fonctionnalités pour CTrouvé

L'annuaire CTrouvé annonce de nombreuses nouvelles fonctionnalités pour l'an 2000, dont certaines sont déjà disponibles en ligne : la mise en place d'un système de moteur de recherche appliqué sur les 85 000 sites de l'annuaire, une rubrique sur les 100 mots clés les plus recherchés par les internautes, des enchères en ligne avec Eurobid. Parmi les projets à court terme : recherche sur des plans cartographiques avec ISMap, référencement immédiat de sites, nouvelles options de recherche et nouveau look. Ça bosse chez CTrouvé !

<http://www.ctrouve.com/>

■ Google pour les Mac-maniaques

Google a lancé en janvier un moteur de recherche spécifique pour les aficionados des machines Apple. Ce service vient s'ajouter à ceux spécifiques du Gouvernement américain et sur Linux, déjà lancés par Google dans les mois précédents.

<http://www.google.com/mac.html>

■ Nomade bascule vers Inktomi

Depuis janvier, Nomade (qui, soit dit en passant, a dépassé il y a peu le cap des 80 000 sites Web répertoriés avec un rythme de 3000 nouveaux sites par mois) utilise le moteur de recherche Inktomi à la place d'AltaVista, et ce dans deux domaines :

- Pour une nouvelle recherche en texte intégral sur le Web international sur la page d'accueil (par l'intermédiaire du menu déroulant situé sous le formulaire de saisie de la requête).

- Pour la recherche sur les pages Web, lorsqu'il n'a pas trouvé dans sa base propre de sites répondant à la requête de l'internaute. Dans cette fonctionnalité, il rejoint par exemple Yahoo! qui fonctionne de la même façon avec Inktomi.

Nomade bascule donc vers Inktomi, comme MSN il y a quelques jours. Pourtant, ce dernier avait déjà changé de prestataire pour ses recherches il y a quelques mois. Il avait abandonné... Inktomi pour AltaVista pour changer de cap à nouveau à la mi-décembre 99. Alors, pourquoi ce changement ? A priori, cela n'a rien à voir avec la prochaine installation d'AltaVista en France dans quelques jours, le contrat ayant été signé entre Nomade et Inktomi depuis plusieurs semaines. Cependant, la stratégie d'Inktomi est claire sur ce point : il ne désire pas rentrer en concurrence avec ses clients et ne propose pas réellement lui-même de moteur de recherche. Il peut plutôt être considéré comme un "grossiste en URLs", ce qui n'est pas le cas d'AltaVista, à la fois portail, moteur et fournisseur de résultats pour ses clients extérieurs. Nomade regrettait surtout de la part d'AltaVista une forte dégradation du temps de réponse depuis plusieurs mois, allant parfois jusqu'à 4 secondes de la part du portail américain. Inktomi, de son côté, garantit aujourd'hui par contrat un temps de réponse inférieur à 250 millisecondes pour proposer ses résultats à ses clients. La recherche se fait de plus sur la base européenne d'Inktomi, située chez BT en Grande-Bretagne, ce qui accélère également le processus. Le but est de fournir un résultat, qu'il soit issu de Nomade ou d'Inktomi, en moins d'une demi-seconde. Les résultats fournis par Inktomi favorisent également les pages en langue française, comme sur Yahoo!, Nomade devant améliorer ce point dans les jours qui suivent, une fois la phase de rodage terminée.

Enfin, Nomade a basculé vers Inktomi car ses services techniques estiment que les résumés fournis par Inktomi sont de meilleure qualité (notamment lorsque les pages proposées ne contiennent pas de balises Meta Description). Si vous ajoutez à cela la perte de pertinence dans ses résultats que connaît AltaVista depuis sa dernière version du 25 octobre dernier, vous obtenez alors un ensemble de raisons qui ont fait pencher la balance du côté d'Inktomi.

Dans les cartons de Nomade, on trouve également un projet de recherche sur les ressources audiovisuelles (MP3 notamment) pour la fin du mois, ainsi que des recherches linguistiques et/ou géographiques pour le mois prochain.

■ Plus d'un milliard de pages

Début juillet, une étude du NEC Research Institute indiquait le nombre de 800 millions de pages Web disponibles sur l'Internet. Dans le même temps, le PDG d'Inktomi avançait le chiffre d'un million de nouvelles pages Web chaque jour. Les deux (NEC et Inktomi) viennent de se réunir pour travailler sur une étude sortie en janvier et annonçant la disponibilité de plus d'un milliard de pages Web à l'heure actuelle en ligne. Bien entendu, il ne s'agit que de pages "statiques", ce chiffre ne prenant pas en compte les pages "dynamiques", c'est-à-dire issues de bases de données en ligne interrogeables au travers de formulaires (les archives de presse en sont l'exemple parfait). Plus de 86% des documents trouvés sont en anglais et... 2,36% en Français ! 4 217 324 sites ont également été identifiés.

Dans le même temps, le moteur Alltheweb du Norvégien Fast a annoncé le chiffre de 300 millions de pages Web dans son index, histoire de dépasser les 250 millions d'AltaVista.

<http://www.inktomi.com/webmap/>

■ 250 millions de pages pour Excite

Le moteur Excite vient d'annoncer le passage de son index à 250 millions de pages, égalant ainsi celui d'AltaVista et restant juste en dessous de celui d'All The Web (300 millions de documents HTML). Excite crawlerait 920 millions de pages pour n'en garder finalement que 250 millions, une fois les pages spammées et/ou identiques éliminées. En fait, il s'agirait de 210 millions de pages Web et 40 millions d'"objets multimédia" (sons, images, vidéos). Rappelons que tous ces chiffres ne sont qu'indicatifs et sont parfois l'objet d'effets d'annonces. Rappelons qu' Excite ne prenait plus en compte les soumissions sur son site depuis de longs mois. Peut-être s'est-il brusquement réveillé !

■ Altavista France sera lancé le 15 février prochain, si ils ont récupéré leur "altavista.fr" (piqué il y a quelques semaines par un cyber-squatteur opportuniste) d'ici là !

