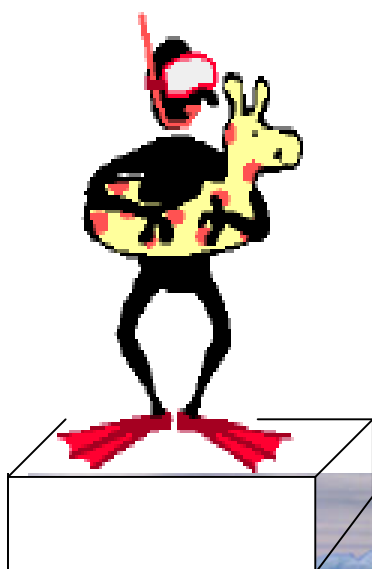


# PANORAMA DES OUTILS DE RECHERCHE SUR LE WEB VISIBLE ET INVISIBLE



*Ce document est librement exploitable. Merci de citer la source.*

# SOMMAIRE

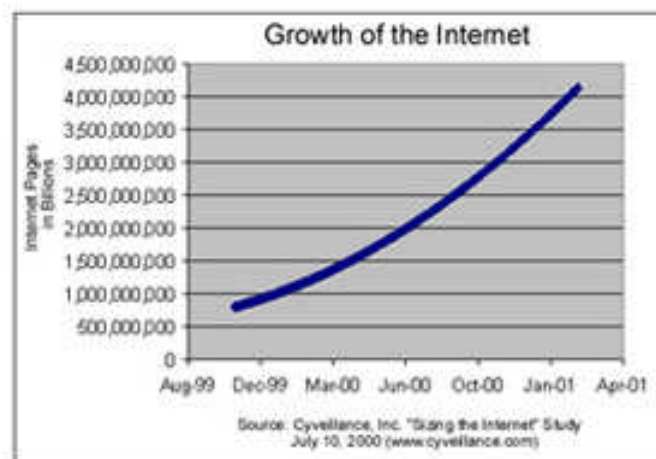
|                                                              |    |
|--------------------------------------------------------------|----|
| Introduction-----                                            | 3  |
| La recherche sur le web visible : les sources généralistes   |    |
| Les annuaires-----                                           | 5  |
| Les moteurs de recherche-----                                | 8  |
| Les métamoteurs -----                                        | 15 |
| La recherche sur le web invisible : les sources spécialisées |    |
| Définition du web invisible -----                            | 22 |
| La simple citation de liens vers d'autres sites -----        | 25 |
| Les annuaires sélectifs et thématiques -----                 | 25 |
| Les portails, les vortails, les sites fédérateurs-----       | 27 |
| Conclusion-----                                              | 30 |
| Sources -----                                                | 33 |

Depuis son invention en 1989 par des membres du CERN (Centre Européen de Recherche Nucléaire), Tim Berners-Lee et Robert Cailliau, la toile d'araignée ou World Wide Web n'a cessé de s'étendre.

Selon la source Médiamétrie, le nombre d'internautes en France est passé de 11,9 millions en mai 2001, 16,4 millions en février 2002 et 20,1 millions en janvier 2003, soit une quasi augmentation de 100 % sur les 60 millions de Français. A titre comparatif, ils sont 166 millions sur 278 millions aux Etats-Unis, 2,235 millions sur 5,2 millions en Finlande et 41,8 millions sur 83 millions en Allemagne. En France, 14,1% des foyers sont connectés à Internet contre 9,3% au Royaume-Uni et 5,6% en Allemagne.

Parallèlement, en 1999, une étude réalisée par le NEC Research Institute et la société Inktomi avait recensé plus de 800 millions de pages web statiques.

En 2002, la société américaine Cyveillance estimait à 3 milliards le nombre de pages web disponibles sur internet dont au moins 3 millions de sites anglophones, 160 000 sites francophones, 60 millions de pages web en français. Chaque jour, plus de 7 millions de nouvelles pages seraient créées.



Source : Cyveillance, Inc "xxxx the internet study July 10, 2000  
<http://www.cyber-strategie.com/formations/ri-2.htm>

Grâce à cette technologie, les auteurs des pages web sont aussi nombreux et différents que les types de contenu mis en ligne. L'internaute pourra aussi bien consulter un site créé par un fan sur Johnny Hallyday que les pages web du Ministère de l'économie et des finances français ou bien celles de documentalistes réunies pour diffuser gratuitement l'information économique en ligne.

Toutefois, le visage du web a évolué depuis quelques années. A ces débuts, ses utilisateurs, issus majoritairement des milieux universitaire et scientifique, prônaient la libre et gratuite circulation des connaissances, présentée alors comme la force d'internet. Aujourd'hui, les sites commerciaux ont énormément investi l'espace web. De plus en plus, l'information devient payante tandis que la gestion des sites relève d'un modèle économique rentable sur le moyen terme.

Néanmoins, une étude a estimé que 80 à 95% de l'information utile à l'entreprise pour surveiller son environnement était publique et disponible.

Au regard des enjeux commerciaux, stratégiques et économiques, l'intérêt pour la recherche d'information sur le web ne cesse de croître. Microsoft est un exemple significatif : une étude américaine annonçait que l'entreprise doublerait dans les prochaines années son équipe travaillant sur ce marché afin de devenir un acteur majeur.

Actuellement, les internautes ont le choix entre deux types d'outils pour trouver la bonne information dans cet incroyable et immense catalogue : les moteurs de recherche et les annuaires. Leur principe est d'indexer les pages web des sites. Or, seulement 16 % des pages sont référencées par les moteurs de recherche et les annuaires. Autrement dit, la part interrogeable et donc visible du web est faible. De plus, outre le nombre de pages à traiter, les moteurs n'intègrent pas de nombreuses autres sources d'information telles que les bases de données, les fichiers word, PDF, etc,.. Ces derniers constituent le web dit invisible.

Pour pallier cette faiblesse, d'autres logiques sont alors apparues pour retrouver ces informations. Peu connues du grand public, la connaissance de ces outils par les documentalistes accroît leur savoir-faire en matière de recherche sur internet.

En conséquence, l'exposé commencera par présenter les trois sources généralistes pour chercher l'information dite visible : les annuaires, les moteurs de recherche et les méta-moteurs. Puis, dans une seconde partie, seront abordées les sources spécialisées sur le web invisible, fondées sur une logique d'accès et de communication des informations différente.

## LA RECHERCHE SUR LE WEB VISIBLE : LES SOURCES GENERALISTES

Il existe trois sortes d'outils disponibles pour la recherche sur le web visible : les annuaires, les moteurs de recherche et les méta-moteurs. Chacun d'eux possède des particularités dans la constitution de leurs fonds, d'interroger ou de traiter les informations.

### Les annuaires

Les annuaires sont des **répertoires de sites internet** organisés en catégories. Ce ne sont donc pas des pages mais des sites qui sont référencés. Ce travail est effectué par une personne qui rédige une courte description du site, la valide pour son indexation dans le répertoire. Ce mode de fonctionnement basé sur l'**humain** donne une dimension qualitative à ce type d'outils structurés. Enfin, la **recherche** se fait sur les **termes de la description du site** et **non sur le contenu de la page**.

Les annuaires sont très utiles pour faire une recherche sur un sujet méconnu ou la dégrossir. L'internaute accède par la première page aux principales catégories. **La navigation s'opère hiérarchiquement du général au particulier**. Au fur et à mesure, l'annuaire affiche une sélection de sites dans la catégorie correspondant à la recherche. L'annuaire s'utilise pour une **recherche générale**.

Il existe plusieurs types d'annuaires :

- des annuaires généralistes francophones. Par exemple : yahoo France, Looksmart, Guide voilà, francité, etc ;
- des annuaires internationaux. Par exemple : yahoo.com, open Directory, search ;
- des annuaires spécialisés sur un domaine précis français ou mondial ;
- des annuaires d'annuaires thématiques.

## TECHNOLOGIES DES PRINCIPAUX ANNUAIRES FRANCOPHONES Mars 2003

Source : Les moteurs de recherche, <http://www.lesmoteursderecherche.com>

| Annuaire             | Database/spider | Annuaire            |
|----------------------|-----------------|---------------------|
| AOL France           | Exalead         | propriétaire        |
| Club Internet        | Fast            | propriétaire        |
| Francité             | DeepIndex       | propriétaire        |
| Lycos France         | Fast            | Propriétaire + DMOZ |
| MSN France           | Inktomi         | La Mine             |
| Nomade.fr/Tiscali.fr | Fast            | propriétaire        |
| Voilà                | propriétaire    | Propriétaire        |
| Yahoo! France        | google          | propriétaire        |

## TABLEAU COMPARATIF DES PRINCIPAUX ANNUAIRES FRANCOPHONES Septembre 2002

Source : Comparaisons quantitatives : Annuaire par J.P. Lardy, <http://urfist.urfist-lyon1.fr/risi/outils1.htm>

| Annuaire             | Taille        | Nombre de référenceurs        | Nombre de catégories |
|----------------------|---------------|-------------------------------|----------------------|
| Club Internet        | 150 000 sites | 5                             | 600                  |
| Francité             | 220 000 sites | nc                            | nc                   |
| Lycos France         | 75 000 sites  | 10                            | 8 000                |
| MSN France           | 65 000 sites  | nc                            | nc                   |
| Nomade.fr/Tiscali.fr | 150 000 sites | 6 en interne et 17 en externe | Plus de 11 000       |
| Voilà                | 67 000 sites  | 8                             | 7 500                |
| Yahoo! France        | 150 000 sites | 14                            | 50 000               |
| Open Directory       | 33 000 sites  | nc                            | nc                   |

Les deux principaux inconvénients des annuaires sont : leur manque d'exhaustivité du web et la mise à jour des références.

L'étude de la société eStat de mars 2002 indique que 43% des utilisateurs d'annuaires trouvent le site recherché sur la première page de résultats. Les 54% et 3% restants naviguent soit dans le guide, soit en cliquant sur le bouton "Suite" ou

"Réponses suivantes". Les requêtes comportent en général un mot ou deux mots-clés.

Toutefois, les études sur le comportement des internautes démontrent une désaffection croissante pour ce type d'outil.

## Les moteurs de recherche

En mars 2002, 7,18% des points de départ du trafic internet était généré par les moteurs de recherche. En mars 2003, ce chiffre est passé à 13,45% selon le Journal du Net. Leur importance ne cesse de croître.

Les moteurs sont des **machines** assistées par des robots (spider ou araignée) qui scrutent le contenu du web à la recherche de pages nouvelles. Le contenu des pages est analysé et stocké dans de **larges index**. Ce sont des outils **quantitatifs** qui référencent un nombre déterminé de **pages** de sites web.

Les moteurs de recherche sont les outils à privilégier quand **la personne sait exactement** ce qu'elle cherche. Lorsqu'elle tape une requête, le moteur interroge d'abord dans son index les termes de la requête. Il recherche ensuite dans les pages avant de les classer par ordre de pertinence.

La différence entre les moteurs réside dans leur algorithme de **calcul de la pertinence** des résultats. Ce calcul se fait en comptabilisant le nombre d'occurrence d'un terme dans un texte tout en prenant en compte leur proximité. D'autres paramètres peuvent être pris en compte : la présence du mot demandé dans le titre du document, sa présence dans le texte de la page et sa graphie (gras, taille des caractères, etc,...), sa présence dans les balises Meta ou dans l'adresse de la page. Les **critères pour classer les documents** sont basés sur le contenu de la page. Ils sont dits "**in the page**".

Puis, pour déjouer les détournements de classement faits par des sites, d'autres méthodes furent élaborées à partir de **critères "off the page"**. Ce fut l'indice de popularité (PageRank de Google) qui résulte d'un calcul se basant sur le nombre de fois qu'un site est cité sur d'autres sites. Autrement dit, plus il existe de liens vers un site, plus celui-ci sera populaire. Cet indice est aussi calculé en fonction de la page sur laquelle clique l'internaute au cours de l'affichage des résultats. Le moteur donnera ainsi une pondération plus forte à un lien souvent sélectionné. Le temps passé sur un site peut être aussi un élément déterminant à condition de revenir après la lecture vers la pages des résultats du moteur.



Le système de classement des moteurs de recherche s'appuie donc sur des calculs de pertinence qui associent de nombreux critères.

Cependant, il est indispensable de bien connaître la façon d'utiliser les moteurs de recherche. Sinon, les résultats peuvent s'avérer très décevants. En effet, même si beaucoup de critères d'interrogation sont identiques d'un moteur à l'autre, ce n'est pas systématique ; et une requête peut s'avérer totalement inefficace si l'internaute n'utilise pas les bons critères de recherche.

Ceux-ci peuvent être : les opérateurs booléens (ET OU SAUF), l'expression (guillemets), la troncature, la recherche linguistique, la recherche sur le titre (title:), le domaine (domain:, site: ), le nom du serveur (host:, site: ), l'URL (URL:) ou encore sur les adresses de liens (link:, linkdomain:).

## TABLEAU COMPARATIF DES DIFFERENTS MOTEURS DE RECHERCHE Août 2002

Source : les moteurs de recherche, <http://www.lesmoteursderecherche.com/tabella.htm>

|                                           | AltaVista                                                            | AllTheWeb<br>(FAST)                                                                                                          | Google                                                                                                                                 | Lycos.com<br><small>Analyse faite par N.Sobieski<br/>Avril 2003</small>                                                                  | Hotbot                                                        |
|-------------------------------------------|----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------|
| <i>Nom de la société<br/>propriétaire</i> | Overture                                                             | Overture                                                                                                                     | Google                                                                                                                                 | Terra Lycos en<br>association avec le<br>groupe Bertelsmann                                                                              | Lycos                                                         |
| <i>N° de pages (millions)</i>             | 1 100                                                                | 2 100                                                                                                                        | 2 200                                                                                                                                  | 340 (**)<br>50 (*****)<br>650 (**)<br><div>* nombre<br/>de fois cité</div>                                                               | 500                                                           |
| <i>Contenu</i>                            | Web, usenet, images,<br>MP3, Audio, vidéo                            | Web, news, photos,<br>vidéos, MP3, FTP, PDF,<br>audio                                                                        | Web, usenet, news, PDF,<br>DOC, XLS, PPT, TRF                                                                                          | Web, annuaires, images,<br>sons, vidéo, MP3, FTP                                                                                         | Web, usenet, news                                             |
| <i>Version française</i>                  | OUI                                                                  | NON                                                                                                                          | OUI                                                                                                                                    | OUI                                                                                                                                      | OUI                                                           |
| <i>Recherche avancée</i>                  | OUI                                                                  | OUI                                                                                                                          | OUI                                                                                                                                    | OUI                                                                                                                                      | OUI                                                           |
| <i>Opérateur standard</i>                 | AND                                                                  | AND                                                                                                                          | AND                                                                                                                                    | AND                                                                                                                                      | AND                                                           |
| <i>Recherche de phrases</i>               | Oui, automatique                                                     | Oui, en utilisant les<br>guillemets                                                                                          | Oui, en utilisant les<br>guillemets                                                                                                    | Oui, en utilisant les<br>guillemets                                                                                                      | Oui, en utilisant les<br>guillemets                           |
| <i>Pluriel/<br/>singulier</i>             | Oui, en utilisant ""*                                                | NON                                                                                                                          | OUI seulement en anglais                                                                                                               | NON                                                                                                                                      | OUI seulement en<br>anglais                                   |
| <i>Mots ignorés</i>                       | OUI                                                                  | OUI                                                                                                                          | OUI                                                                                                                                    | NON                                                                                                                                      | OUI                                                           |
| <i>Champs de recherche</i>                | Oui : langue des<br>pages, titre, mot-clé,<br>url, lien, site, image | Oui : langue, filtre de<br>mots et domaines, date<br>de mise à jour, taille des<br>pages, format des<br>fichiers, adresse IP | Oui : titre, mot-clé, url, lien,<br>site, image, format du<br>fichier, date de mise à jour,<br>pages spécifiques (pages<br>similaires) | Oui : index, langue des<br>pages, titre, domaine,<br>mot-clé, lien, site, image,<br>annuaire, date de mise à<br>jour, taille du document | Oui : position<br>géographique, type<br>de fichier, date, URL |

|                                                      |                                                                                                                 |                                    |                                                                                                                        |                                                                                                                                                                                        |                                                          |
|------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|------------------------------------|------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|
| <i>Recherche d'images, sons, Java, etc.</i>          | OUI                                                                                                             | OUI                                | OUI                                                                                                                    | OUI                                                                                                                                                                                    | OUI                                                      |
| <i>Recherche par langue</i>                          | OUI                                                                                                             | OUI                                | OUI                                                                                                                    | OUI                                                                                                                                                                                    | OUI                                                      |
| <i>Filtre pour les enfants</i>                       | OUI                                                                                                             | OUI                                | NON                                                                                                                    | OUI                                                                                                                                                                                    | OUI                                                      |
| <i>Regroupement de résultats</i>                     | OUI                                                                                                             | OUI                                | OUI                                                                                                                    | NON                                                                                                                                                                                    | OUI                                                      |
| <i>Personnalisation des résultats des recherches</i> | OUI : description de base                                                                                       | Oui : nombre de résultats par page | Oui : seulement les titres ; recherche à l'intérieur des résultats                                                     | Oui : nombre de résultats par page                                                                                                                                                     | Oui : description brève, complète ou seulement l'adresse |
| <i>Mémorisation des préférences personnelles</i>     | OUI                                                                                                             | OUI                                | OUI                                                                                                                    | NON                                                                                                                                                                                    | OUI                                                      |
| <i>Fonctions spéciales</i>                           | <u>Refine</u> : pour ordonner les résultats d'une recherche ;<br><u>Translate</u> : pour traduire des pages web | NON                                | <u>Traduction</u> des pages web ; bouton " <u>j'ai de la chance</u> " permet d'accéder directement au premier résultat | <u>Lycos voyeur</u> : affichage des mots saisis en temps réel par les utilisateurs<br><u>Lycos Top 100</u> : présentation chaque mois des mots clés les plus tapés par les internautes | Meta Words pour des recherches avec plus de critères     |
| <i>Catalogue sites</i>                               | OUI                                                                                                             | OUI                                | OUI                                                                                                                    | OUI                                                                                                                                                                                    | OUI                                                      |
| <i>Temps d'insertion dans les sites</i>              | 7-45 jours                                                                                                      | 9-14 jours                         | 20-30 jours                                                                                                            | 2-5 jours officiellement<br>2 à 3 semaines officieusement                                                                                                                              | 20-30 jours                                              |

Quant aux requêtes saisies par les internautes, l'observation faite par baromètre ad'oc en mars 2003 révèle les faits suivants :

- 40,1% des requêtes ne comportent qu'un mot,
- 28,9% des requêtes sont composées de 2 mots,
- 14,1% ont 3 mots
- 8,6% ont plus de 3 mots
- et 8,3% consultent les catégories des annuaires.

L'enquête confirme la poursuite du déclin de l'intérêt pour la recherche par catégories, dû notamment à sa moindre valorisation au profit de la cellule de saisie des requêtes. De plus, la manière de rechercher des internautes a aussi évolué ; 55% d'entre eux expriment leurs demandes sur plus de mots en avril 2003 contre 44,2% en avril 2002.

L'étude de la société eStat de mars 2002 indique que 82% des utilisateurs de moteurs de recherche trouvent le site recherché sur la première page de résultats.

Même si les importants index des moteurs ne couvrent la totalité du web, leur couverture est beaucoup plus importante que celle des annuaires.

Des robots comme Inktomi parcourent le web et mettent leurs indexes au services des plus grands moteurs.

## **TECHNOLOGIES DES MOTEURS DE RECHERCHE Mars 2003**

*Source : Les moteurs de recherche, <http://www.lesmoteursderecherche.com>*

| <b>MOTEURS DE RECHERCHE INTERNATIONAUX</b> | <b>DATABASE/SPIDER</b> | <b>ANNUAIRE</b>   |
|--------------------------------------------|------------------------|-------------------|
| <b>Google</b>                              | propriétaire           | DMOZ              |
| <b>AOL Time Warner</b>                     | Google                 | DMOZ              |
| <b>Yahoo.com</b>                           | Google                 | propriétaire      |
| <b>MSN.com</b>                             | Inktomi                | DirectHit         |
| <b>Altavista.com</b>                       | propriétaire           | Looksmart         |
| <b>Ask Jeeves</b>                          | TEOMA                  | Looksmart et DMOZ |
| <b>Lycos.com</b>                           | FAST                   | DMOZ              |
| <b>Alltheweb.com</b>                       | FAST                   | Ø                 |

Les moteurs généralistes sont classables en fonction de la localisation géographique et linguistique des sites.

Il existe au moins deux catégories : les moteurs internationaux (ex : google.com), les moteurs par pays (ex : google.fr) et, dans le cas français, des moteurs francophones.

#### **PART DE MARCHE DES PRINCIPAUX MOTEURS Février 2003**

*Source : Netbooster*

| <b>Moteurs de recherche</b> | <b>Part de marché en %</b> |
|-----------------------------|----------------------------|
| Google                      | 53,33                      |
| Yahoo                       | 16,45                      |
| Wanadoo                     | 9,36                       |
| Voilà                       | 3,52                       |
| msn.fr                      | 5,12                       |
| Aol.fr                      | 4,07                       |
| Lycos.fr                    | 2,12                       |
| free                        | 1,58                       |
| Nomade.fr                   | 1,34                       |
| Altavista                   | 0,81                       |
| Autres                      | 2,30                       |
| - alltheweb                 |                            |
| - toile Q                   |                            |
| - netscape                  |                            |
| - excite                    |                            |

#### **CLASSEMENT FRANÇAIS DE LA POSITION DES MOTEURS ET ANNUAIRES**

**(.FR ET .COM CONFONDUS) Février 2003**

*Source : La revue du référencement, mars 2003, <http://www.revue-referencement.com>*

| <b>Moteurs et annuaires de recherche</b> | <b>Evolution du positionnement</b> |
|------------------------------------------|------------------------------------|
| 1 google                                 | ➔                                  |
| 2 yahoo                                  | ↗                                  |
| 3 wanadoo / voilà                        | ↘                                  |
| 4 msn.fr                                 | ↗                                  |
| 5 aol.fr                                 | ↗                                  |
| 6 lycos.fr                               | ↘                                  |
| 7 nomade.fr                              | ➔                                  |

En indexant chaque mot d'une page web et par leur couverture, les moteurs de recherche permettent d'effectuer des interrogations relativement exhaustives sur le web à l'aide d'options de recherches avancées.

En revanche, cette puissance génère ses faiblesses que sont un nombre de réponses très élevé, de nombreuses inexactitudes et la nécessité de savoir utiliser ces outils.

## Les métamoteurs

Les métamoteurs permettent d'interroger simultanément plusieurs moteurs de recherche. Les résultats de la requête sont donc la synthèse des références trouvées dans les différents outils.

Outre le gain de temps, ils permettent de débroussailler rapidement un sujet et de repérer les outils qui y répondent le mieux.

La seule restriction des métamoteurs est liée à son principe d'interrogation simultanée auprès de plusieurs moteurs de recherche. En effet, comme la même requête est envoyée à tous les moteurs, il est impossible de faire une requête complexe puisque chaque moteur utilise une syntaxe propre.

Il existe deux types de métamoteurs : ceux utilisés directement sur internet, appelés **métamoteurs on-line**, et ceux installés sur son ordinateur ou **métamoteurs off-line**.

Les métamoteurs on-line permettent d'interroger sur internet plusieurs moteurs de recherche simultanément. Certains offrent aussi la possibilité de paramétrer la requête (sélection des moteurs de recherche consultés, nombre de réponses par page, choix de l'opérateur booléen, sauvegarde de la recherche, etc...), de la relancer ultérieurement, de créer des alertes ou d'envoyer le résultat par mail.

Les métamoteurs off-line permettent de paramétrer l'équation de recherche sans être connecté à internet. Puis, l'utilisateur fait exécuter le programme pendant son temps de connexion sur le web. Une fois les données rapatriées sur le disque dur de son ordinateur, il peut les visualiser hors ligne.

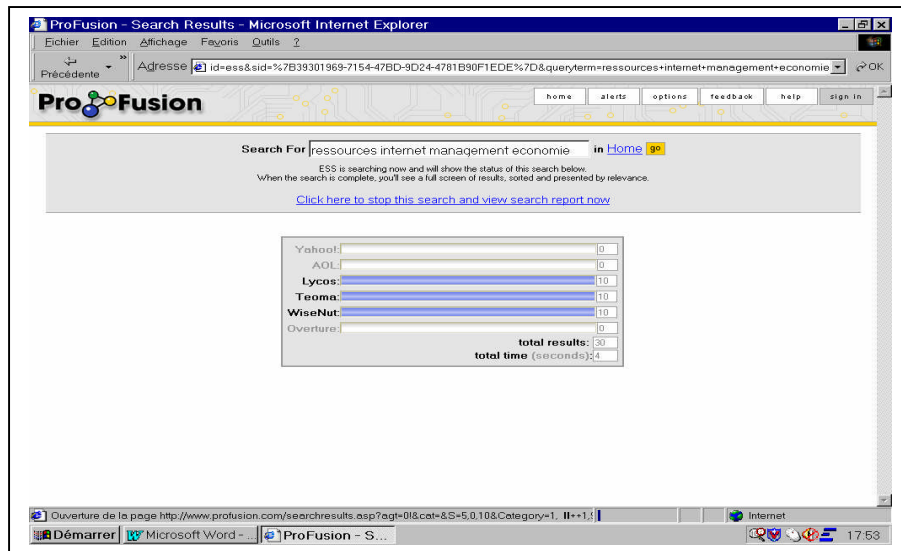
Il existe des métamoteurs off-line spécialisés en recherche d'images. Par exemple, le logiciel iwolf de la société Trellian est un outil de recherche qui localise les fichiers images et films sur le web en explorant les anneaux Web à la recherche de fichiers.

Puissant, ce type de moteurs de recherche met désormais l'accent sur le traitement de l'affichage des réponses obtenues. Les informations sont présentées sous forme cartographique en privilégiant l'aspect thématique (ex. : Kartoo *France*) ou bien par des vignettes représentant les pages des sites trouvés.

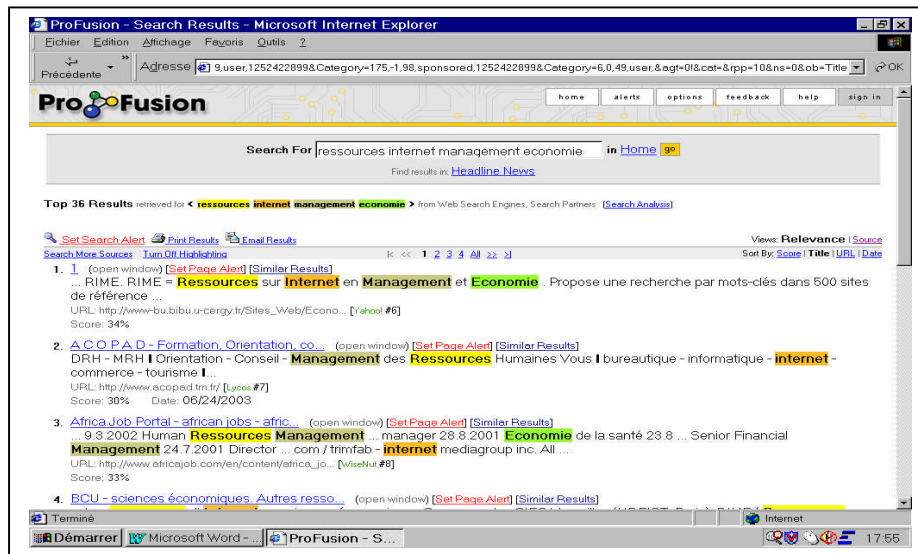




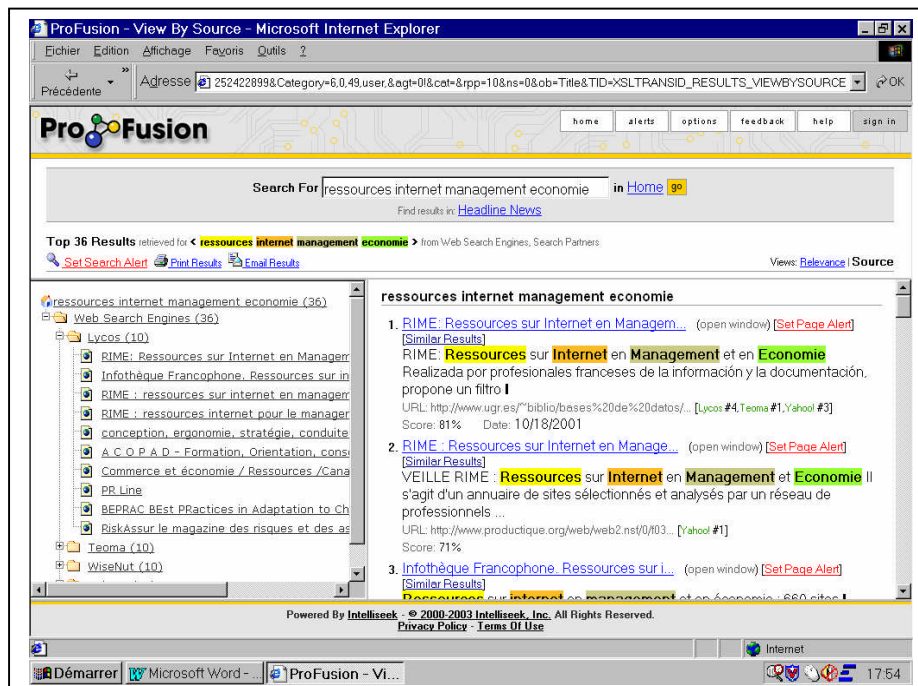
Recherche en cours



Affichage de la liste des réponses

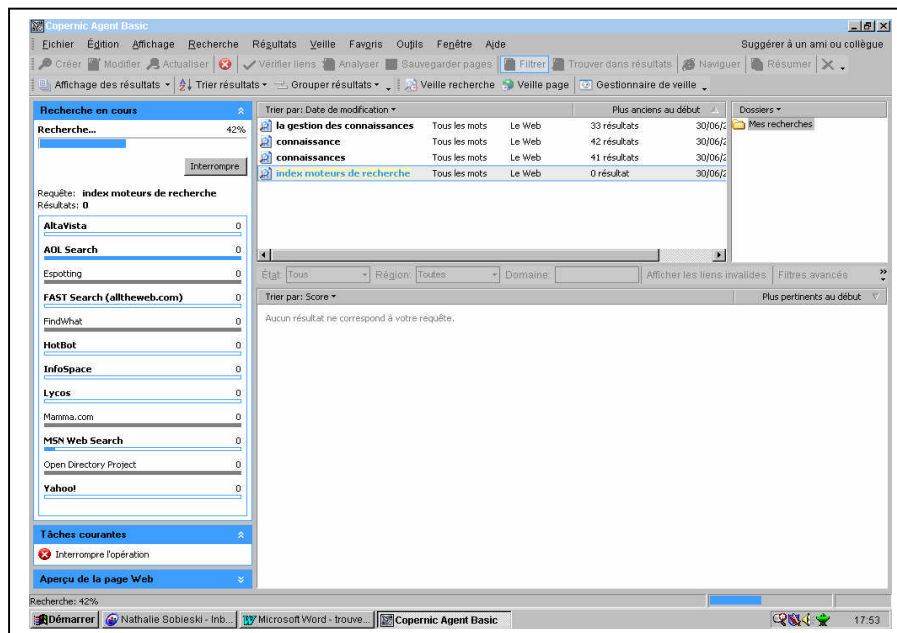


Affichage des réponses par sources

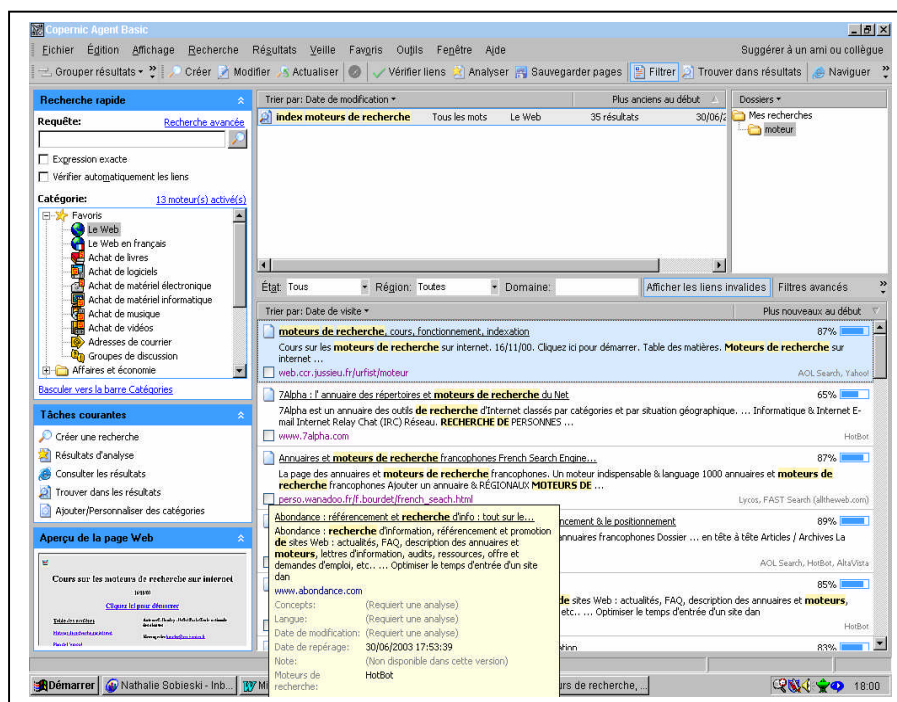


ECRANS DU META-MOTEUR ON-LINE PROFUSION Juin 2003  
Source : <http://www.profusion.com>

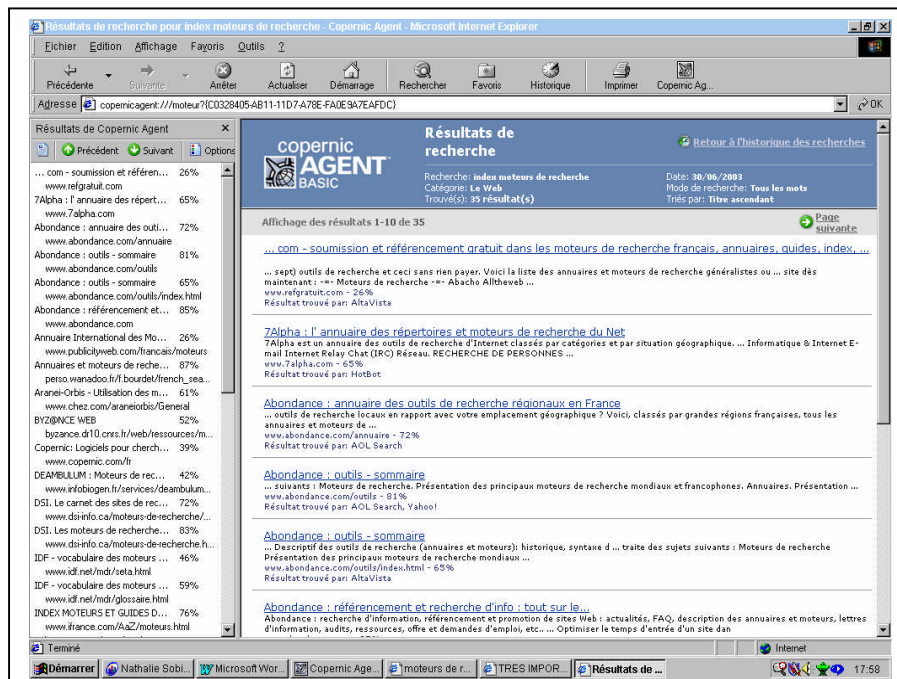
Recherche en cours



Affichage de la liste des réponses avec options



Affichage uniquement de la liste des réponses



ECRANS DU META-MOTEUR OFF-LINE COPERNIC Juin 2003  
Source : logiciel copernic agent basic

## TABLEAU COMPARATIF DES DIFFERENTS META-MOTEURS Avril 2003

Source : Nathalie Sobieski

|                                       | <b>Webcrawler</b>                                                                                                                                          | <b>ProFusion</b>                                                                                               | <b>Copernic</b><br>Version professionnelle<br>Version personnelle<br>Version basic (gratuite)                                                                                                                            |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Nom de la société propriétaire</i> | InfoSpace                                                                                                                                                  | Intelliseek, Inc                                                                                               | Copernic                                                                                                                                                                                                                 |
| <i>Type de métamoteur</i>             | Métamoteur on-line                                                                                                                                         | Métamoteur on-line                                                                                             | Métamoteur off-line                                                                                                                                                                                                      |
| <i>Outils de recherche interrogés</i> | Google, Looksmart, Inktomi, Ask Jeeves, About, Overture, FindWhat, FAST, Open Directory, Search Hippo, Sprinks, WebCrawler Picks, WebCrawler Search, Teoma | Altavista, MSN, Teoma, Raging Search, yahoo, About, Adobe PDF Online, AOL, LookSmart, Lycos, Netscape, WiseNut | De 90 (version gratuite) à plus de 1000 moteurs de recherche + web invisible couvert par des catégories spécialisées donnant accès à plus de 1 000 sources supplémentaires qui sont groupées dans quelque 90 catégories. |
| <i>Version française</i>              | NON                                                                                                                                                        | NON                                                                                                            | OUI                                                                                                                                                                                                                      |
| <i>Recherche avancée</i>              | OUI                                                                                                                                                        | OUI                                                                                                            | OUI                                                                                                                                                                                                                      |
| <i>Opérateur standard</i>             | AND                                                                                                                                                        | AND                                                                                                            | AND                                                                                                                                                                                                                      |
| <i>Recherche de phrases</i>           | OUI                                                                                                                                                        | OUI, en utilisant les guillemets                                                                               | OUI                                                                                                                                                                                                                      |
| <i>Pluriel/ singulier</i>             | NON                                                                                                                                                        | OUI                                                                                                            | NON                                                                                                                                                                                                                      |
| <i>Mots ignorés</i>                   | NON                                                                                                                                                        | NON                                                                                                            | NON                                                                                                                                                                                                                      |
| <i>Recherche par langue</i>           | NON                                                                                                                                                        | NON                                                                                                            | NON                                                                                                                                                                                                                      |
| <i>Filtre pour les enfants</i>        | NON                                                                                                                                                        | NON                                                                                                            | OUI                                                                                                                                                                                                                      |
| <i>Options de recherche</i>           | Sélection du fond de la recherche : Web,                                                                                                                   | Choix du nombre de sources interrogées ;                                                                       | OUI                                                                                                                                                                                                                      |

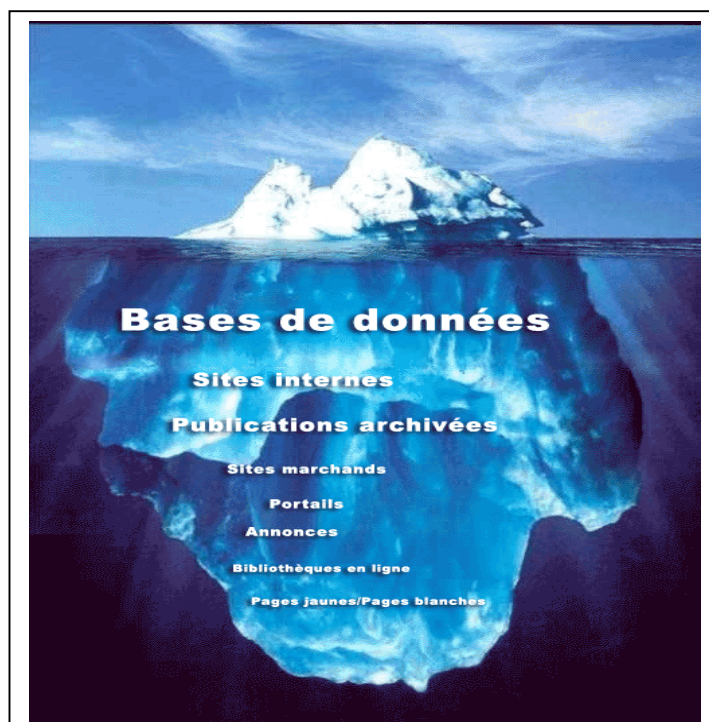
|                                                      |                                                                                                                                                   |                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                              |
|------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                      | news, photo ; sélection du nombre de sites maximal affichés par source ; sélection du temps de recherche ; choix du nombre de sources interrogées | sélection du nombre de sites maximal affichés par source ; sélection du temps de recherche                                                                                                    | <u>Barre de filtres</u> comprenant plusieurs options pour filtrer les listes de résultats (les fonctions diffèrent selon la version installée).<br><br>Dans la version professionnelle :<br>détection de langues dans les pages Web, détection des pages en double ayant des adresses différentes, extraction des concepts clés dans les pages Web trouvées                                  |
| <i>Regroupement de résultats</i>                     | OUI par degré de pertinence ou par source                                                                                                         | OUI                                                                                                                                                                                           | OUI<br><br>En version personnelle et professionnelle : regroupement des résultats de recherche (par scores, dates de visite, domaines, contenus d'annotation, états: nouveaux, visités, annotés, cochés, sauvegardés) ;<br>uniquement professionnelle :<br>regroupement avancé des résultats de recherche (par langues, dates de modification des pages, contenus de pages identiques, etc.) |
| <i>Personnalisation des résultats des recherches</i> | Envoi de la liste des réponses affichées par mail                                                                                                 | OUI : par <u>score</u> (présence en % des mots de la recherche), par <u>titre du document</u> , par <u>adresse du site web</u> (URL), ou par <u>date</u> . Classement par degré de pertinence | Possibilité de <u>trier</u> différemment les listes de résultats selon divers champs: titre, extrait, adresse, score, date de repérage, date de visite, date de                                                                                                                                                                                                                              |

|                                                         |     |                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                |
|---------------------------------------------------------|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                         |     | mais aussi en fonction du concept ( <u>catégorisation des résultats</u> ) ou des moteurs de recherche ( <u>Source</u> ) ;<br>Dédoublonnage des réponses                                                  | modification, moteur de recherche.<br>Possibilité de <u>grouper</u> différemment les listes de résultats selon divers champs: score, état - nouveau, état - visité, état - sauvegardé, état - annoté, état - coché, date de repérage, date de visite, domaine, contenu des annotations, langue, date de modification, contenu identique.<br>Consolidation automatique des résultats avec des liens identiques. |
| <i>Mémorisation des préférences personnelles</i>        | OUI | OUI                                                                                                                                                                                                      | OUI<br>Possibilité de créer des dossiers et des sous-dossiers de recherches                                                                                                                                                                                                                                                                                                                                    |
| <i>Fonctions spéciales</i>                              | NON | Format d'impression des résultats, envoi par mail des résultats, création et enregistrement d'alertes ou DSI (pages alerts ou search alerts), création de groupes de moteurs de recherches personnalisés | Automatisation de tâches<br>Veille de recherches<br>Veille de pages Web<br>Possibilité d'obtenir un résumé de la page Web d'un résultat sélectionné                                                                                                                                                                                                                                                            |
| <i>Catalogue sites</i>                                  | OUI | OUI : thématique                                                                                                                                                                                         | OUI                                                                                                                                                                                                                                                                                                                                                                                                            |
| <i>Affichage du temps de recherche dans les moteurs</i> | NON | OUI                                                                                                                                                                                                      | NON                                                                                                                                                                                                                                                                                                                                                                                                            |

## LA RECHERCHE SUR LE WEB INVISIBLE : LES SOURCES SPECIALISEES

### Définition du web invisible

Les sources généralistes interrogent donc des index construits après la visite des sites web par les robots. Or, certaines informations sur les sites ne peuvent pas être indexées par ces derniers. Ces pages web constituent la partie immergée de l'iceberg, appelée deep web (web profond) ou web invisible.



Source : [http://c.asselin.free.fr/french/invisible\\_web.htm](http://c.asselin.free.fr/french/invisible_web.htm)

Ce web invisible comprend un grand nombre d'éléments laissés de côté par les moteurs qui sont :

- les documents trop volumineux pour être entièrement indexés et la sélection d'un certain niveau de pages dans l'arborescence du site ;
- des pages protégées par l'auteur (balise meta qui stoppe le robot) ou interdites au référencement pour des raisons de confidentialité (ex : extranet, pages personnelles) ;

- des pages générées dynamiquement lors d'une requête par exemple ou par un script (perl, php,...) ;
- des formats de documents qui ne sont pas reconnus par les robots (fichiers de traitement texte, tableur, animation flash, pdf, etc,...) ;
- des sites trop récents ou n'ayant pas encore fait l'objet d'un référencement ;
- des pages qui demandent une identification de l'internaute ;
- des pages situées dans un frame ;
- les archives dont l'accès peut être payant (ex : la presse) ;
- les banques de données professionnelles ;
- etc,...

Chris Sherman et Gary Price distinguent ainsi 4 types de web :

- The *Opaque Web* : les pages qui pourraient être indexées par les moteurs mais qui ne le sont pas (limitation d'indexation du nombre de pages d'un site, fréquence d'indexation, liens absents vers des pages ne permettant donc pas un crawling) ;
- The *Private Web* : les pages webs disponibles mais volontairement exclues par les webmasters (mot de passe, metatags ou fichiers dans la page pour que le robot du moteur ne l'indexe pas).
- The *Proprietary web* : pages seulement accessibles pour les personnes qui s'identifient. Le robot ne peut donc pas y accéder.
- The *Truly Invisible Web* : contenu qui ne peut être indexé pour des raisons techniques.

Autrement dit, le problème n'est pas tant leur visibilité que l'accessibilité de ces pages qui rend difficile leur indexation par les moteurs de recherche. C'est pourquoi l'étude BrightPlanet désigne un web de surface que les robots parcourent facilement et un web profond **que leur technologie ne parvient pas encore à explorer, bien que visible par d'autres outils**. Pour Laurent Soron, dans son livre intitulé "Le web invisible, cet inconnu..", le web invisible correspond à l'ensemble des documents (textes, vidéos, images, ..) qui n'est pas indexé par les outils de recherche (moteurs, annuaires). Ces données peuvent être de nature non indexable (animations), non référencées (volontairement ou non), ou dynamiques.

Selon l'étude BrightPlanet, le web invisible serait 260 fois plus vaste que le web visible. L'enjeu est donc énorme.

Des outils sont alors apparus pour permettre à l'internaute de naviguer dans ce web profond. Leur distinction par rapport aux outils de recherche dits généralistes (ex : moteurs de recherche, annuaires) se fonde sur les critères suivants : le degré de spécialisation du sujet traité, le mode de communication, de diffusion et de sélection des informations.

Les outils peuvent être regroupés ainsi :

- la simple citation de liens vers d'autres sites : le maillage, le webring ;
- les annuaires sélectifs et thématiques ;
- les portails, les vortails, les sites fédérateurs.



## **La simple citation de liens vers d'autres sites**

Cet outil de recherche du web invisible reprend dans son expression la plus simple le concept fondateur du web : le lien. Deux logiques peuvent être appliquées selon le contexte.

### **Le maillage**

Le principe du maillage est d'indiquer dans les pages web du site des liens hypertextes vers d'autres sites jugés pertinents par le webmaster. C'est le premier mode de fonctionnement du web.

### **Le webring ou Anneau de sites**

C'est un "système de navigation en forme de boucle qui relie des sites abordant des sujets communs et qui permet à l'internaute de passer facilement de l'un à l'autre. Sur la page d'accueil de ces sites, un sélecteur permet de passer de site en site en utilisant des liens hypertextes." La mention Webring peut être un gage de qualité et de fiabilité car, en général, ce sont des sites membres qui se cooptent.

Ces liens sont donc le moyen le plus simple pour rebondir sur des pages ignorées par les moteurs de recherche et trouver de nouvelles pistes d'information.

## **Les annuaires sélectifs et thématiques**

Pour compenser les lacunes des moteurs de recherche, des personnes en groupe ou individuellement – souvent des professionnels de l'information issus du milieu universitaire, de l'enseignement ou de la recherche - ont commencé à établir des listes de sources d'information sur le web jugées pertinentes par rapport au domaine traité.

Ces répertoires ou ces guides ont notamment trois particularités essentielles : ils sont sélectifs, qualitatifs et largement ouverts sur le web invisible. Ils couvrent soit

plusieurs thèmes –annuaires généralistes -, soit un seul de manière très approfondi. Enfin, les ressources en ligne retenues peuvent appartenir à différents types de support : sites de référence, journaux électroniques, listes de diffusion, bases de données, catalogues, document de recherche, etc..

Très souvent, ces annuaires du web invisible proposent un accès thématique par un menu arborescent ou bien la possibilité d'interroger leurs références par un moteur de recherche sur leur site. En effet, par définition, il n'existe pas pour le moment de moteurs de recherche sur le web invisible puisque ce dernier ne peut pas être indexé par les moteurs. En conséquence, un moteur du web invisible désignera en fait la possibilité d'interroger les références d'un annuaire en ligne. Sinon ce serait un abus de langage.

#### **QUELQUES EXEMPLES D'ANNUAIRES SELECTIFS SUR LE WEB INVISIBLE :**

##### **Invisible-web**

<http://www.invisible-web.net/>

Plus de 10 000 sites sélectionnés sur le web invisible.

##### **INFOMINE**

<http://infomine.ucr.edu/>

Bibliothèque virtuelle proposant des ressources en ligne dans les domaines étudiantin, pédagogique et celui de la recherche dans le milieu universitaire.

Ce sont des bases de données, des journaux électroniques, des livres en ligne, des bulletins d'information, des lettres de diffusion, des catalogues en ligne, des répertoires de chercheurs, etc,.. L'annuaire est réalisé par des bibliothécaires de l'université de Californie, Wake Forest University, l'université de Détroit et d'autres administrations.

##### **Resource Discovery Network (RDN)**

<http://www.rdn.ac.uk/>

Fruit de la collaboration de plus de 70 organismes de l'enseignement et de la recherche, le site propose gratuitement une sélection commentée de sites web

pertinents dans les domaines de l'enseignement, la pédagogie et la recherche. RDN est fondé sur un réseau coopératif entre un noyau central et des partenaires.

### **Educasource**

*<http://www.educasource.education.fr/>*

Le site Éducasource a pour vocation de faciliter l'accès aux ressources électroniques utiles à l'enseignement. C'est une base de données de plus de 6000 références de ressources électroniques dans le domaine de l'enseignement. Le recensement s'étend à des ressources non francophones et à des ressources à accès payant. Un moteur de recherche permet d'interroger cette base.

Sa consultation est totalement gratuite.

## **Les portails, les vortails, les sites fédérateurs**

Comme l'information sur le web devenait de plus en plus abondante, non structurée, et les outils multiples, de nouveaux espaces de travail, d'accès à l'information et de communication ont émergé. Ceci amena le concept d'annuaire à évoluer. Sont alors apparus les portails, les vortails ou les sites fédérateurs. Offrant une approche spécifique des sources d'informations en ligne sur un sujet, ils permettent à l'internaute d'accéder rapidement et simplement à plusieurs types de supports classés "web invisible" et réunis sur un même site.

Les définitions proviennent de la rubrique "Encyclo" du Journal du Net.

### **Les portails et les vortails**

Un site portail est un site qui se propose d'être une "porte d'entrée" sur le Web.

Les objectifs d'un portail est donner un accès simple et unique à des contenus, des applications et des services. Sa force provient donc de l'agrégation de contenu et de services à valeur ajoutée.

Les catégories de fonctions ou services proposées par un portail sont :

- des outils de recherche
- des services d'information et de diffusion (ex : actualité, interview, reportages, dossiers, etc..),
- des outils de communication (email, liste de diffusion, forum, groupe de discussion),
- des accès thématiques à l'information : agrégation du contenu (collecte d'information de sources internes et externes, format de présentation unique) ; organisation du contenu (classement des ressources selon un plan de classement)
- des services personnalisables ou du contenu.

Le portail est souvent le site par lequel un internaute entame sa navigation.

Il existe deux catégories de portails :

- les **portails généralistes** construits autour d'un annuaire ou d'un moteur de recherche généraliste auquel sont associés des fonctions complémentaires ;
- les **portails spécialisés ou professionnels ou thématiques**, appelés aussi **vortails** qui ont le même rôle sur un secteur limité (ex : un portail éducatif, un portail high-tech). Ils ont une très forte valeur ajoutée car ils offrent des ressources sélectionnées pour un environnement professionnel précis et des usages propres à un groupe d'utilisateurs concernés. Ce sont des sites très pertinents pour découvrir les ressources du web invisible.

#### **QUELQUES EXEMPLES DE PORTAILS THEMATIQUES :**

<http://www.educasup.education.fr/index.html>

<http://ntic.org/>

<http://www.reseau-amerique-latine.fr>

<http://www.business.com>

## **Les sites fédérateurs**

Avec le développement de la sphère économique du web, le concept de portail a donné naissance à celui de site fédérateur. Ce dernier est son équivalent du point de vue marchand. En effet, le concept du site fédérateur est né des constats et des besoins suivants :

- faciliter la recherche d'information ;
- être rentable pour son propriétaire ;
- permettre à certaines entreprises de perdurer ou de se développer ;
- permettre de développer le commerce électronique.

C'est un site proche du portail thématique ou de la communauté virtuelle. Le fédérateur associe les acteurs, centralise les informations et les transactions d'un secteur particulier. Il permet à l'utilisateur d'acheter un produit, un service, ou de consulter une information en rapport avec l'activité fédérée. Le site fédérateur contient 4 piliers : un annuaire spécialisé, un espace de vente, un espace d'échanges et un contenu éditorial. C'est grâce à des outils interactifs spécifiques à l'Internet que le site acquiert une dimension fédératrice.

Les multiples sources de revenus associées à des dépenses raisonnables de promotion le définissent comme un modèle économique rentable pour ses acteurs.

La clé de sa réussite repose sur un développement par paliers, de partenariats, d'externalisation, etc.

Il comporte quatre sections : un annuaire spécialisé, un contenu éditorial (information, interviews, dossiers, reportages, etc,..), des échanges (forum, liste de diffusion, petites annonces, etc,..), un espace fidélisation (espace membre, boutique, catalogue, affiliation).

Le site fédérateur peut constituer un autre accès à d'autres sources d'information.

### **QUELQUES EXEMPLES DE SITES FEDERATEURS :**

<http://www.federateur.com/index.shtml>

<http://www.ijardin.com/fra/index.php3>

## CONCLUSION

Le monde des outils de recherche poursuit son évolution.

L'utilisation de l'outil de recherche se banalise auprès du grand public. Les internautes les utilisent de plus en plus.

Les versions francophones des outils de recherche sont privilégiées : 72,5% des visiteurs d'un site venant d'un moteur ou d'un annuaire utilisent la version française de l'outil (*Source : eStat, mars 2002*).

Néanmoins, l'évolution comparée des moteurs de recherche et des annuaires généralistes démontre le relatif déclin des annuaires dont l'audience globale est en baisse constante. En effet, en mars 2002, sur l'ensemble de la population analysée par eStat, une nette préférence pour les requêtes se dessine : parmi les 21% d'internautes ayant utilisé un outil de recherche pour trouver un site web, une moyenne de 27% utilisent les annuaires et 73% les moteurs de recherche. Cependant, ces derniers tels que Lycos ou AOL ont bénéficié de campagnes publicitaires diffusés dans différents média. Ceci amène à s'interroger sur les effets générés en matière d'usages des internautes.

Si les annuaires généralistes sont moins bien connus, le concept a donné forme à de nouveaux types de sites web. De plus en plus spécialisés, ils ont évolué en fonction des besoins des internautes mais aussi selon la logique adoptée.

Ainsi, après l'annuaire, sont apparus le portail généraliste, thématique et le vortail. Puis, le désir de rentabiliser des portails a donné naissance à des sites fédérateurs dont le but était de regrouper les acteurs d'un domaine. La spécialisation s'est poursuivie dans la sphère économique du web par l'apparition des places de marché. Le Journal du Net en donne la définition suivante : « *Site portail exclusivement réservé au commerce interentreprises et dont l'objectif est de faciliter la mise en relation entre acheteurs et vendeurs. On peut toutefois distinguer les places de marché transactionnelles qui autorisent la commande en ligne (e-procurement) et la facturation ou le paiement en ligne, de celles qui sont plutôt des "infomédiaires".* »

Autrement dit, bien que son principe - c'est-à-dire référencer et regrouper des informations pour ensuite orienter- soit toujours identique, l'annuaire a subi des

transformations. L'ajout de nouvelles fonctions a modifié la forme tandis que de nouvelles logiques ont amené la césure entre le web marchand et celui non-marchand. Mal connue, cette évolution est une source de nombreuses confusions dans le vocabulaire employé pour ce domaine. Une place de marché est une sous-branche des sites fédérateurs, eux-mêmes issus de la famille des portails. Cette spécialisation illustre la confrontation entre le monde des professionnels de l'information et celui des acteurs de la sphère économique.

D'un point de vue technique, la troisième génération d'outils de recherche comme teoma ou exalead arrive. Leur capacité croissante de stockage et de traitement permettent d'envisager les premiers outils utilisant la sémantique des contenus pour la recherche ou le classement d'information. Les outils sont maintenant plus préoccupés par la pertinence des résultats qu'à l'exhaustivité. Des fonctions de spécialisation géographique ou sectorielle (ex : mirago) permettront de proposer des résultats pré filtrés et d'augmenter la qualité des réponses.

Les moteurs de recherche et les annuaires ont un bel avenir devant eux car plus de 40% des internautes visitent un site pour la première fois grâce à ces outils.

#### LES POINTS DE DEPART DU TRAFIC INTERNET

Source : Journal du Net, Mars 2003

| Origine du trafic                                                             | Mars 2002 | Mars 2003 |
|-------------------------------------------------------------------------------|-----------|-----------|
| <b>Navigation directe (URL tapée au clavier ou dans la liste des favoris)</b> | 50,21%    | 65,45% ↗  |
| <b>Liens entre sites</b>                                                      | 42,60%    | 21,04% ↘  |
| <b>Moteurs de recherche</b>                                                   | 7,18%     | 13,46% ↗  |

Or, ces nouveaux usages ne sont pas anodins pour les professionnels de l'information. Ils les amènent à s'interroger sur la valeur ajoutée de leurs bases de données, leur ergonomie, leur méthodologie de recherche ou leur positionnement par rapport à « cette grande base de données où l'on trouve tout » que représente le web pour 90% des étudiants.

Actuellement, le secteur des outils de recherche se concentre fortement. Le marché est dominé par trois principaux acteurs : **Google**, **Lycos** (lycos et Hotbot sous son contrôle) et **Overture**, spécialiste du référencement payant (marques Altavista et alltheweb/FAST).

Récemment, un journal en ligne a annoncé la volonté de Microsoft d'investir ce domaine amenant celui à se demander si Google sera le nouveau Netscape des outils de recherche.

Compétition accrue, spécialisation, puissance, la recherche sur le web est l'objet de très forts enjeux économiques et stratégiques. Que restera-t-il du principe de libre communication des informations et de la philanthropie originelle du web ?



## SOURCES

@position. (Page consultée le 27/06/2003). « La revue du référencement ». [En ligne]. Mise à jour : nc. <http://www.revue-referencement.com>

« Indicateur.com ». (Page consultée le 27/06/2003). [En ligne]. Mise à jour : quotidienne. <http://www.indicateur.com/default.asp>

« Journal du net ». (Page consultée le 25/06/2003). [En ligne]. Mise à jour : quotidienne. <http://www.journaldunet.com>

« Le barometre adoc : référencement et trafic : l'expérience de sites d'entreprises ». (Page consultée le 30/06/2003) [En ligne]. Mise à jour : mensuelle. <http://www.barometre.adoc.fr>

□□□ Alexandre Serres et Marie-Laure Malingre, URFIST Rennes. (Page consultée le 25/06/2003). « Recherche d'information sur internet : Approfondissement ». [En ligne]. Mise à jour : 01/02/2003. [http://www.uhb.fr/urfist/Supports/Rechinfo2/Rechinfo2\\_cadre.htm](http://www.uhb.fr/urfist/Supports/Rechinfo2/Rechinfo2_cadre.htm)

□□□ Christophe ASSELIN. (Page consultée le 25/06/2003). « Intelligence-Center.com » [En ligne]. Mise à jour : 2003. <http://c.asselin.free.fr>

Décisionnel.net. (Page consultée le 27/06/2003). « Les outils de recherche ». [En ligne]. Mise à jour : 09/04/2001. <http://www.decisionnel.net/recherche/index.htm>

Emmanuel Desmontils. (Page consultée le 26/06/2003). « Outils et techniques pour explorer le web invisible » [En ligne]. Mise à jour : 03/02/2003. <http://www.sciences.univ-nantes.fr/info/perso/permanents/desmontils/InvisibleWeb.html>

Internet Diffusion. (Page consultée le 27/06/2003). « Le Guide des Moteurs de Recherche ». [En ligne]. Mise à jour : nc. <http://www.lesmoteursderecherche.com>

Marc Duval. (Page consultée le 25/06/2003). « Service de recherche documentaire DSI » [En ligne]. Mise à jour : 2003-05-20. <http://www.dsi-info.ca>

Maxboxer. (Page consultée le 26/06/2003). « Internet et la recherche d'information ». [En ligne]. Mise à jour : nc. [http://www.chez.com/maxboxer/cours/ri\\_sommaire.html](http://www.chez.com/maxboxer/cours/ri_sommaire.html)

Olivier Andrieu. (Page consultée le 26/06/2003). « Recherche d'information et référencement ». [En ligne]. Mise à jour : quotidienne. <http://www.abondance.com>

Thibaut Deshayes. (Page consultée le 1/07/2003). « Asktibbs.com : Recherche sur internet ». [En ligne]. Mise à jour : 1999-2002. <http://www.asktibbs.com/>