TEXTE DE PRESENTATION

Ce présent document est composé de différentes parties distinctes qui sont destinées à fournir une vision riche et globale de l'activité et du milieu dans lequel j'ai fait mes premiers pas pendant ce stage, de façon à faciliter le débat sinon l'écoute lors de la soutenance. Libre à chacun de choisir quelle(s) partie(s) il lira : Du plus global au plus précis chacun des textes ci dessous, devra trouver son lecteur, du moins averti au plus curieux.
Retrouvez aussi la doc sur http://www.ex-pression.com (optimisé IE 5)

Descriptif des parties destiné à orienter votre lecture

- **Généralités : 4**

structure d'accueil 4

Période de stage 4

Entreprise d'accueil 4

Encadrement 4

Objectifs de la mission 4

- **Petit état de l'art 6**

Pour ceux qui ne connaissent pas du tout l'univers des moteurs et de la recherche sur Internet, un panorama sur une quinzaine de pages de l'activité de référencement, sa problématique, son intérêt et différentes choses qui y sont liées.

- **1-Structuration des documents sur le web : Du visible à l'invisible....6**
  - Les méta données : une aide précieuse pour qualifier un document sur Internet....*6*
  - Un exemple *7*
- **2-L'indexation par un moteur de recherche....9**
  - Les visites des spiders....*9*
  - Le stockage des données.... *10*
- **3- La recherche d'information : de l'invisible au visible 10**
  - *Formuler une requête, c'est avant tout parler un autre langage*.... *10*
    - Un comportement qui varie avec l'habitude d'usage : Apprendre la langue.... 10
    - Un comportement qui varie selon l'internaute : Des usages personnels et communautaires 11
    - L'orthographe : Apprendre à ne plus faire de faute ou apprendre à en faire ?.... 11
    - Le langage des outils de recherche : Apprendre à dialoguer 12

- Savoir se reprendre et ne pas abandonner 13

❑ 4-Le quatrième acteur 14

▫ *Avec le document* *14*

▫ *Avec les outils* *14*

▫ *Avec l'internaute* *14*

❖ Rapport de l'activité de référencement 16

▫ *secteur de l'entreprise* *16*

▫ *L'entreprise* *16*

▫ *Place au sein de l'entreprise* *17*

▫ *Travail et réalisations* *17*

❖ Utilisation logicielle 18

❖ Supports explicatifs et pistes : 19

Ce document est destiné à fournir quelques explications, quelques pistes concernant les différents sujets abordés dans le rapport, mais ne prétend en aucun cas répondre à toutes les questions que vous pourrez vous poser en le lisant. Il regroupe donc tout ce qui a pu me sembler intéressant de citer, des résultats d'instituts de sondages, des articles de journaux Internet, des réflexions personnelles et des documents de référence, dont les sources, s'il y a lieu, sont explicitement citées.

❑ Qu'est ce qu'un cyber-documentaliste ? 20

❑ Quel différence entre un moteur et un annuaire ? 20

Annuaires, généralités *20*

Moteurs, généralités *21*

Méta moteur *22*

Autres : *22*

Petite note sur les moteurs en langage naturel *24*

❑ Le langage HTML 26

❑ Les méta données 28

❑ Petite note sur l'indexation des images 30

❑ La charte de déontologie 30

❑ Abus de référencement 34

❑ L'option -En cache- de Google 34

❑ Panorama du web actuel 36

- **DEFINITION DES TERMES 38**
- **TABLE DES LIENS 46**
- **TABLE DES FIGURES et TABLEAUX 47**
- **BIBLIOGRAPHIE ET BOOKMARKS 48**

Généralités :
structure d'accueil

Période de stage

Mon stage a débuté le 4 février 2002 et se termine le 4 juin de la même année, soit un stage de 4 mois.

Entreprise d'accueil

Pendant ces quatre mois de stage, j'ai été accueillie par une Web Agency Lilloise, **Better way web-Agency**. L'entreprise est en charge de l'accompagnement de ses clients dans leur stratégie Internet :

- De l'audit du site à la proposition de solutions adaptées selon les résultats mis en évidence.
- Ergonomie, architecture de navigation.
- Programmation, choix et réalisation des plates-formes techniques.
- Rédaction des textes clients.
- Et l'activité de référencement à laquelle nous allons particulièrement nous intéresser.

Encadrement

Durant cette période d'apprentissage, j'ai été placée sous la tutelle de **Michel HUBERT**, directeur technique de **Better way**. La fonction de mon tuteur dans l'entreprise est de coordonner tous les aspects techniques inhérents à la vie de l'entreprise et à celle des projets clients : Choix matériels et logiciels, orientation et conseil dans ses choix techniques et technologiques. **Michel HUBERT** supervise dans ce but deux équipes : l'équipe de développement et l'équipe de référencement que j'ai intégrée.
Dans cette équipe j'ai été également encadrée par **Gaétan TOUYON** avec qui j'ai travaillé en collaboration sur l'activité de référencement.

Objectifs de la mission

La mission de référencement qui m'a été proposée au début du stage se décomposait en deux grands axes :

1- Etude de visibilité

On entend par visibilité, une évaluation du « rayonnement » d'un site , sur un moteur, un annuaire, sur un mot clef et la toile dans sa globalité. Est-il visible des internautes ? Plus la

visibilité d'un site est grande et plus le nombre de point d'entrée sur ce site est grand (et inversement). Plus un site est visible plus il va être visité.

Dans ce premier axe, ont été développés les travaux d' :

- Analyse concurrentielle.
- Analyse de positionnement.
- Analyse de popularité*.
- Utilisation des outils.

2- Activité de référencement

Au delà des activités d'analyse et d'étude des sites clients (et concurrents), se trouve le cœur même de l'activité de référencement qui se base sur les choix stratégiques mis en évidence par la première partie d'analyse :

- Recherche des mots-clés.
- Distribution et intégration des mots clés dans l'univers et les sites du client.
- Inscription manuelle des sites clients dans les moteurs et annuaires.

S'ajoutent à ces deux axes le suivi des sites travaillés dans l'activité de référencement, ainsi que tout un travail périphérique qui sera présenté avec plus de précision dans la documentation ci-jointe et le jour de la soutenance.

Petit état de l'art

s'y retrouver

Le nombre des sites disponibles sur Internet ne cesse de s'accroître. Avec plus de 2 milliards de pages, le réseau devient une gigantesque base documentaire, une mine d'informations.
Le problème de savoir s'il existe bien une production qui puisse répondre à nos attentes est presque devenu obsolète et le plus gros problème de l'internaute, lorsqu'il se connecte est plutôt de savoir comment trouver **LE** document dont il soupçonne au plus profond de lui-même, l'existence.

Il y a quelques années, la masse de documents sur le net était beaucoup moins importante (le nombre de foyers connectés également), la manière de percevoir la recherche d'information, bien différente et les outils de recherches d'une moins grande technicité. Le fait est que les enjeux de l'indexation, du référencement, sont apparus grandissant, ainsi que le nombre d'entreprises désirant rentabiliser leur investissement. Pourquoi créer un site marchand, un site vitrine s'il n'est visible de personne ou des mauvaises personnes ??
A partir de cet état de fait les choses se sont enchaînées les unes aux autres : augmentation du nombre de sites mis en ligne, du nombre d'internautes, de l'impact des sites…Et la boucle fût bouclée.
Il n'est pas nécessaire de justifier plus longtemps de la nécessité grandissante et effective de rendre accessible, et de manière pertinente l'information disparate du Web, de pouvoir structurer les données, de les construire de manière à ce qu'elles soient le plus facilement traitables.

1- Structuration des documents sur le web : Du visible à l'invisible

- *Les méta données* : une aide précieuse pour qualifier un document sur Internet

Un document type, le plus classique que l'on puisse trouver sur le web peut, mais ce n'est pas obligatoire, se composer de méta données* qui peuvent fournir une orientation sur le sujet du document, mais aussi sur ce qu'il contient dans sa globalité (et voire bien plus lorsque les métas sont exploitées au maximum, mais c'est assez rare). Ces méta données ne sont pas lisibles par l'internaute directement, mais jouent un rôle dans la façon dont les documents vont être traités par le moteur ou plutôt par ses spiders*.

Il y a deux manières d'indexer un document sur Internet : Soit en l'enregistrant auprès d'un moteur de recherche (Google par exemple), soit en le soumettant à un annuaire (Yahoo ! en est un).[1]

[1] A noter que bien souvent les bases des moteurs sont couplées avec celles de certains annuaires et vice versa. Google ,par exemple, couple ses bases avec Dmoz et jusqu'à peu l'annuaire de YAHOO ! avec celles de Google.

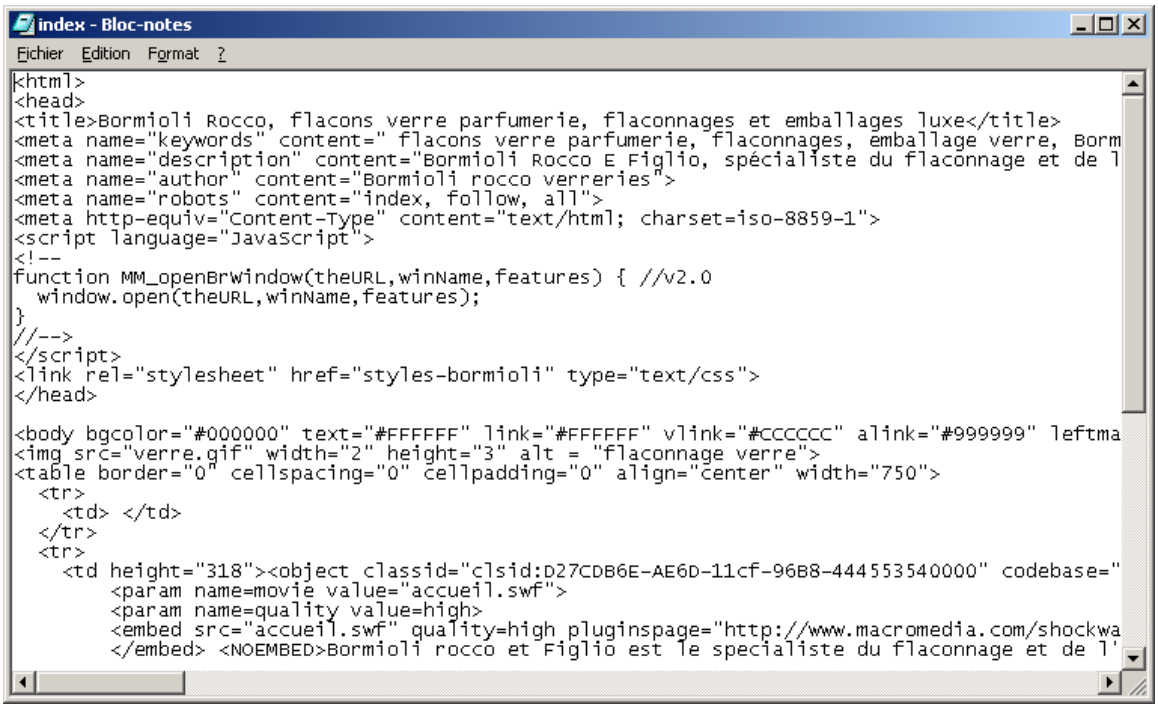
Pour l'annuaire, la question des méta données ne se pose pas réellement, puisque c'est un cyber-documentaliste, un surfeur, qui va entrer, après une visite minutieuse, le site soumis dans telle ou telle catégorie de l'annuaire (certains moteurs viennent ' piocher ' dans les bases des annuaires, c'est le cas de Google et de DMOZ).

En ce qui concerne l'indexation dans les moteurs de recherche, les avis sont mitigés et les rumeurs-vérité ne font que se contredire. Néanmoins, il s'avère que les méta données restent encore un point important, dans l'indexation des sites et des pages (sinon fortement apprécié), même si elles ne demeurent pas essentielles, leur absence peut parfois être pénalisante.

- *Un exemple*

Voici la page Internet telle que vous la voyez dans votre navigateur (figure a) suivi du fichier source (figure b),le codage HTML brut

Figure a : Le site tel que vous le voyez


index - Bloc-notes

Fichier Edition Format ?

```
<html>
<head>
<title>Bormioli Rocco, flacons verre parfumerie, flaconnages et emballages luxe</title>
<meta name="keywords" content=" flacons verre parfumerie, flaconnages, emballage verre, Borm
<meta name="description" content="Bormioli Rocco E Figlio, spécialiste du flaconnage et de l
<meta name="author" content="Bormioli rocco verreries">
<meta name="robots" content="index, follow, all">
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<script language="JavaScript">
<!--
function MM_openBrWindow(theURL,winName,features) { //v2.0
  window.open(theURL,winName,features);
}
//-->
</script>
<link rel="stylesheet" href="styles-bormioli" type="text/css">
</head>

<body bgcolor="#000000" text="#FFFFFF" link="#FFFFFF" vlink="#CCCCCC" alink="#999999" leftma
<img src="verre.gif" width="2" height="3" alt = "flaconnage verre">
<table border="0" cellspacing="0" cellpadding="0" align="center" width="750">
  <tr>
    <td> </td>
  </tr>
  <tr>
    <td height="318"><object classid="clsid:D27CDB6E-AE6D-11cf-96B8-444553540000" codebase="
        <param name=movie value="accueil.swf">
        <param name=quality value=high>
        <embed src="accueil.swf" quality=high pluginspage="http://www.macromedia.com/shockwa
        </embed> <NOEMBED>Bormioli rocco et Figlio est le specialiste du flaconnage et de l'
```

Figure b : le code source

La *figure b*, nous montre la page Internet (*figure a*) telle qu'elle est stockée sur le serveur.
Les méta données interviennent dans notre exemple dans les 3 à 8 premières lignes du code :
Le titre, *ligne3* : que l'on va retrouver en haut à gauche du navigateur (*figure a*).
Les mots clés (keywords*), *ligne 4* : Ces mots sont en général lus par les spiders*, mais on ne les retrouve pas dans le navigateur.
Description, *ligne 5* : Cette description est en général lue par les spiders*, mais, tout comme les mots clés, elle n'est pas visible à partir du navigateur.
L'auteur, *ligne 6* : Destinée à mettre en avant l'auteur (question de droit), cette information n'est toujours pas accessible par le navigateur.
Les robots, *ligne 7* : cette information est en fait strictement réservée au spider* du moteur de recherche, elle l'informe sur ce qu'il doit faire avec cette page : INDEX (l'indexer), FOLLOW (suivre les liens qu'elle contient), ALL (sans aucune restriction).
Content-type, *ligne 8* : C'est une méta donnée qui informe sur le format des données qui sont utilisées.

Ces six méta données sont les plus couramment utilisées, les trois premières (titre, keywords, description) étant celles que l'on essaye de composer au plus juste, d'autant plus que les sites à contenu purement textuel sont rares quand toute l'attention s'est souvent portée sur l'aspect graphique et que la seule information disponible pour indexer une page est celle portée par ces balises.

2- L'indexation par un moteur de recherche

- Les visites des spiders

Pour indexer les pages internet qu'on lui soumet, le moteur envoie un spider. C'est un programme qui va visiter la page en regardant son code source et qui va lire l'intégralité de ce qu'il contient, (l'intégralité de ce contient la figure b) et donc les méta données.
Le spider lit donc ce qui se trouve dans les méta données, le texte (entre les balises <body> </body>), certaines balises dont nous ferons mention, (peut être les commentaires et les noms de fichiers, mais le fait n'est pas avéré).

Et il indexe ce qu'il y lit :

- Le titre de la page
- Les mots clefs
- La description
- Le corps du texte en lui-même
- + voire d'autres informations portées par les méta données
- + voire les commentaires et les noms de fichiers dans les chemins d'accès
- + voire d'autres balises dont l'impact n'est pas nettement défini dans l'indexation

Pour notre exemple :

- **Un titre**: Bormioli Rocco & Figlio, flacon verre parfumerie, flaconnages et emballages luxe
- **Des mots clés**: Flacons verre parfumerie, flaconnages, emballage verre, bormioli rocco et figlio ...
- **Une description**: Bormioli Roco & Figlio, spécialiste du flaconnage et de l'emballage ...
- **Le texte, le corps du document**: Version Html, Télécharger Flash, Mentions Légales, credits, Optimisé 800x600
- **Un auteur pour les droits**: Bormiolli rocco verreries
- Les autres informations susceptibles d'être lues par le spider* et indexées :
  Le contenu de la balise <noembed></noembed>, <noframe></noframe>

Il est facile à comprendre que les méta données peuvent jouer un rôle très important, pour preuve, notre exemple, dont le contenu lisible par le browser[2], '*Version Html, Télécharger Flash, Mentions Légales, crédits, Optimisé 800x600*' n'est en aucun cas représentatif du contenu

[2] Browser = Navigateur (IE, nescape...)

sémantique de la page, plutôt porté par les méta données et les images (ici, c'est une animation flash qui tourne en back ground).

Même si tout paraît se passer très sobrement, les choses sont en réalité beaucoup plus ambiguës et il n'existe aucun absolu. Lorsqu'un spider visite une page, on ne peut jamais vraiment être certain de ce qu'il a lu, ni de ce qu'il a indexé, cela va dépendre du moteur et de ses critères d'indexation qui restent bien souvent assez mystérieux et le fait est qu'aucun moteur ne révèle ses stratégies et qu'il faut apprendre à composer avec ces incertitudes, apprendre aussi à connaître les tendances, les nouveautés en s'abonnant à des listes de diffusion mais aussi en tentant au mieux de vérifier les informations qui y sont données. Il est aussi parfois utile de tester les indexations : utiliser telle technique sur telle moteur et voir quel résultat ressort, quelle technique peut être éventuellement plus efficace qu'une autre.

- Le stockage des données

Selon les moteurs donc, les données récoltées par le spider ne sont pas stockées de la même façon. Certaines bases ne prendront peut être pas en compte certaines données, d'autres encore les privilégierons. Ces bases dont le traitement reste inconnu (classement ? - établissement d'équivalences ? - de synonymies ? - de groupes ?) servent de support lors d'une requête. Le moteur reçoit la requête de l'internaute (qu'il a tapé sur l'interface web du moteur) et va établir une liste de documents pertinents en allant « fouiller » la base de données constituée par le spider en utilisant un algorithme de pertinence* différent selon les moteurs.

3- La recherche d'information : de l'invisible au visible

Pour trouver le/les document(s) désiré(s), l'internaute va donc taper une requête, un mot ou plusieurs. La requête en elle-même est un obstacle qui se dresse dans la recherche d'informations sur le Web.

- Formuler une requête, c'est avant tout parler un autre langage

Le monde de la requête 'grand public' n'est pas assez vieux pour être étudié de fond en comble, mais il existe néanmoins des comportements de langage qui peuvent être relevés

- Un comportement qui varie avec l'habitude d'usage : Apprendre la langue

Un internaute néophyte se plaint souvent du temps qu'il passe à chercher des informations sur le net et il arrive bien souvent à la conclusion rapide et sans équivoque: ' De toute manière, ça marche pas ! '. Pourtant, en persévérant un peu, on finit par 'apprivoiser' la bête et à se réconcilier avec nos anciens démons.

La requête moyenne de l'internaute français est constituée **d'un seul mot**, mais les requêtes contenant **deux mots** sont de plus en plus usitées. Cette évolution, aussi peu significative semble-t-elle, nous prouve que les comportements vis à vis des O-R (outils de recherche) changent et avec lui le langage des requêtes. (et nos méthodes de référencement !!)

Exemple de comportement de recherche[3] :

(…)

- 40% des recherches durent entre 30 minutes et une heure[4].
- Plus de 50% des internautes utilisent deux outils de recherche.
- Lorsqu'ils ne trouvent pas "du premier coup", 70% essayent un autre mot clé sur le même outil de recherche, plutôt que de s'orienter vers un nouveau site.
- Plus de 50% des internautes consultent plus de 4 pages de résultats.[5]

Même si les recherches se font encore, pour une majorité, sur **un mot** il ne serait pas étonnant de voir se multiplier les requêtes à **deux mots** (quand c'est possible). Ce qui élimine d'emblée un grand nombre de documents, nous l'avons souligné. Et participe a une bonne utilisation des O-R.

- Un comportement qui varie selon l'internaute : Des usages personnels et communautaires

De même, il semble évident que pour une même chose, un même sujet, deux internautes ne vont pas utiliser la même requête. D'abord parce qu'il existe des synonymies, des usages culturellement plus fréquents dans certaines régions/pays que d'autres.

Par exemple, *Passiflore* et *Fleur de la passion*, qui sont deux mots pour signifier un seul et même objet. Alors qu'un internaute tapera *Passiflore*, l'autre utilisera *Fleur de la passion* et les résultats que chacun trouvera seront bien différents. (ce qui souligne l'importance, pour celui qui référence un site de pouvoir anticiper les requêtes éventuelle).

Par conséquent, même *ce qui se conçoit clairement s'énonce aisément*, il n'en va pas moins de chacun d'utiliser son propre vocabulaire, sa propre connaissance du monde et aux référenceurs d'éclairer ces mystères en anticipant des comportements langagiers qui peuvent parfois être curieux.

- L'orthographe: Apprendre à ne plus faire de fautes* ou apprendre à en faire ?

Lorsqu'on fait une requête sur un moteur ou sur un annuaire, il peut arriver qu'une faute se glisse dans le ou les mot(s) que nous utilisons : fautes d'orthographe, fautes d'inattention ou fautes

[3] Source : *abondance*,http://etudes.abondance.com/recherche-res.html
[4] Alors que la consultation des documents est souvent bien plus brève.
[5] A savoir qu'il vaut mieux ,pour être vu des 50% autres, se positionner dans les quatre premières pages.

de frappe. Parfois la faute est si discrète qu'on ne comprend pas pourquoi les résultats ne sont pas ceux auxquels on s'attendait, d'autre fois c'est en tapant une requête erronée (avec une faute) qu'on trouvera notre document (celui qui correspond le mieux).

Pourtant, globalement la faute est une barrière à la recherche et de nombreux outils (dont Google) propose un « correcteur de requête » qui souligne l'éventuelle faute (mais n'empêche pas de la conserver dans la requête). Pour palier à ce phénomène la solution des correcteurs est un bon compromis entre l'utilisateur et la langue, et permet aussi, en quelque sorte, de rééduquer les usagés et les webmasters.

Une autre solution est souvent envisagée par les référenceurs :

L'internaute tape une requête erronée , le nombre de réponses que lui propose le moteur est moins grand (en général) que si la requête avait été correcte, le site référencé avec la faute apparaît donc plus facilement et récupère un visiteur qu'il n'aurait pas eu outre mesure. Même si la méthode permet de récupérer des visiteurs, elle n'est pas sur le long terme une solution.

- Le langage des outils de recherche : Apprendre à dialoguer

La communication qui est établie entre l'interface d'un O-R et l'internaute est une communication qui passe uniquement par le texte et qui bien souvent est dénuée de grammaire[6]. Ce n'est pas pour autant que les possibilités de recherche se trouvent limitées aux nombres de mots, figés, qui se trouve dans le dictionnaire. Et il existe, comme dans la vie plusieurs moyens de s'exprimer. L'exemple le plus simple est celui qui différencie annuaires et moteurs. Deux formes d'O-R, deux formes de requête.

o Annuaires

Pour peu que l'internaute soit un peu averti, il saura que l'outil sur lequel il est, est un annuaire, un répertoire et que tout y est classé thématiquement, il peut donc partir d'un thème de recherche très général pour affiner ensuite sa recherche. Ce qui lui permet de rester relativement vague dans un premier temps et de se faire aiguiller par les propositions de rubriques de l'outil.

Exemple : Je veux des informations sur le *remboursement de la sécurité sociale*.

- En passant par plusieurs requêtes avec un annuaire (YAHOO !)[7] :
« *Santé* » > choix de la rubrique **Accueil > Santé**[8] **> «** *Remboursement* »

- Avec la requête *Remboursement* sur HOTBOT[9], moteur de recherche :
Présence de résultats concernant le:
Remboursement (d'achat)

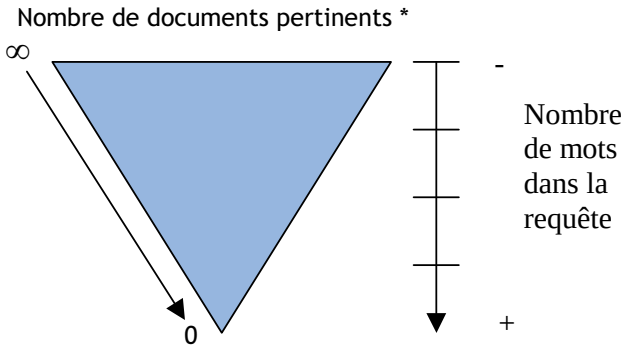
[6] Sauf éventuellement pour les moteurs utilisant une technologie de langage naturel.
[7] http://fr.yahoo.com/
[8] http://fr.dir.yahoo.com/Sante/
[9] http://www.hotbot.lycos.fr

Remboursement(de taxes)
Remboursement (de frais)
Remboursement (de sécurité sociale), (de santé)
Remboursement (de prêt) (de franchise), (de dettes), (de créances)
Remboursement (autres significations diverses et minoritaires)

Sur ma page de résultats YAHOO !, je ne trouve que des adresses de sites ayant directement rapport avec ce que je cherche : Des informations sur le remboursement (de sécurité sociale). Alors que sur le moteur HOTBOT, je récolte une pléiade de documents dont le sujet est assez varié.

o Moteurs

Avec un moteur on ne peut pas affiner une requête aussi facilement, on utilise donc une suite de mots, plus les mots de la requête sont nombreux, plus le sens va être précis.



Exemple :

« *Remboursement* » avec le moteur de voilà[10] : **81 289** réponses
« *Remboursement sécurit é* » avec le moteur de voilà : **25 327** réponses
« *Remboursement sécurité social e* » avec le moteur de voilà : **13 708** réponses

Le comportement que peut avoir un internaute dans la façon de formuler ses requêtes peut donc dépendre de l'outil qu'il utilise. Dans notre exemple, nous avons opposé Annuaires et Moteurs, car la différence nous a semblé être la plus facile, mais de la même manière le comportement de recherche d'un internaute pourra être variable d'un moteur à l'autre (d'un moteur en langage naturel à un qui ne l'est pas, par exemple).

- Savoir se reprendre et ne pas abandonner

[10] http://search.ke.voila.fr

Comme nous l'avons souligné, il y a énormément de façons de trouver un même document et contrairement au proverbe, toutes *les routes* ne *mènent* pas *à Rome*…

Il est très fréquent d'être confronté à un échec lors des recherches. Certains ne persévèrent pas et concluent qu'il n'y a pas de documents qui correspondent à leur recherche. D'autres au contraire, reformulent leur requête, la précisent ou la rendent plus généraliste. La pauvreté des réponses d'un moteur n'est pas une fin en soit et il faut savoir parfois mettre en cause nos méthodes de recherche.

4- Le quatrième acteur

Le panorama qui vient d'être dressé met en évidence trois grands acteurs dans le domaine de la recherche sur Internet : Le document, les outils et l'internaute. Le quatrième étant celui du chargé de référencement dont le rôle est d'entrer en inter-compréhension avec ces différents acteurs.

- Avec le document

Le rédacteur et le chargé de référencement peuvent être une seule et même personne. Ce n'est néanmoins ni l'un ni l'autre qui va décider de quel format sera le site dont il sera chargé et même s'il peut participer à l'élaboration du document, il n'en est jamais maître et doit ajuster sa stratégie, sa manière de procéder aux contraintes techniques et artistiques qui lui sont imposées.

- Avec les outils

De même, nous avons souligné la multiplicité des outils de recherche disponibles sur Internet. Chacun d'entre eux a un fonctionnement qui lui est propre. Il faut donc pouvoir évaluer les points communs, les différences qui peuvent exister entre les O-R pour optimiser au maximum la visibilité du site que l'on référence.

- Avec l'internaute

Le chargé de référencement ne rentre pas en contact direct avec l'internaute, même s'il doit au mieux comprendre ses comportements de recherche et la formulation de ses requêtes

C'est au regard de ces différents points que va s'orienter mon rapport où il sera question de ma place[11] chez BETTER WAY Web-Agency en tant que stagiaire de l'activité référencement. Avant

[11] Comprendre : travail, réalisations, relations, collaborations…

d'aborder le sujet et de décrire l'ensemble des travaux, vous avez maintenant une vision déjà cultivée de la problématique du référencement, ce qui vous permettra d'avoir une compréhension plus intuitive du reste du document et de la présentation orale.

Rapport de l'
Activité de référencement : Résumé

Il n'est pas facile d'expliquer dans sa globalité ce qu'a été mon expérience chez BETTER-WAY en la résumant en quelques mots, je vais donc tenter de décrire les taches qui m'ont été confiées, ce qui permettra de vous faire une bonne représentation de l'activité que j'ai exercée pendant ces quatre mois de stage.

- secteur de l'entreprise

Comme nous l'avons présenté dans la première partie du document, Better–way est une agence web, une web agency comme beaucoup furent crées il y a quelques années, le rôle des web agency peut varier d'une entreprise à l'autre, certaines proposent un hébergement d'autres non, certaines se spécialisent dans tel ou tel type de site d'autres n'ont pas de spécialisation, mais globalement, ce qui définie une Web agency est :

- Sa spécialisation dans le média Internet
- Ses connaissances en communication et marketing
- Son équilibre créativité & technicité
- Sa réactivité face aux évolutions du net, du marché et des tendances
- Ses rapports rapprochés avec le client

- L'entreprise

Le Groupe **BETTER WAY a été** créé en Janvier 2000 à partir de la diversification d'un groupe textile roubaisien et se compose de trois activités distinctes :

- **BETTER WAY Web Agency** Pour le conseil et l'accompagnement dans la stratégie Internet de ses clients.
- **BETTER WAY CRM (Custumer Relationship Management)** Pour la relation avec les clients d'entreprises diverses via un Web Call Center, destiné à répondre aux questions des entreprises et particuliers.
- **BETTER WAY Training** Pour la formation à distance, l'e-learning

L'agence de Lille de better-web agency qui comportait à sa création une dizaine de personnes a maintenant plus que doublé ses effectifs et se partage en grands pôles :

- *Un pôle technique* : Le directeur technique, l'équipe de développement et l'équipe de référencement réduite depuis peu d'un membre.
- *Un pôle de la création*, le directeur de la création, les responsables artistiques, les graphistes et web designer
- Et *un pôle marketing* constitué des chefs de projets qui sont en quelque sorte une interface permanente entre le client et les différents pôles de l'entreprise.

- Place au sein de l'entreprise

C'est au pôle technique que j'ai été intégrée, pour suivre l'activité de référencement alors tenue par **Gaétan TOUYON**, avec qui j'ai donc travaillé pendant quatre mois.

- Travail et réalisations

- *Pour les prospects*

Pour les prospects, les appels d'offres : Audit de référencement, étude de visibilité, étude concurrentielle, proposition de stratégie de référencement, conseil, rédaction de rapport.

- *Pour les clients*

Audit de référencement, étude de visibilité, étude concurrentielle, proposition de stratégie de référencement, rédaction de rapport, optimisation* du site client, recherche de partenaires web, établissement de la stratégie de référencement, soumission aux annuaires, enregistrement auprès des moteurs

- *Veille*

Recherche d'informations sur les moteurs et les méthodes de référencement et établissement de rapports et documents internes. Veille permanente ou sur demande.

- *Test logiciel*

Test de nouveaux logiciels liés au référencement et évaluation, contact éventuel avec les sociétés éditrices.

Tous ces aspects seront abordés lors de la présentation orale du stage, des travaux et réalisations que j'essaierai de documenter au mieux.

Utilisation logicielle :
Du traitement de texte à l'outil spécialisé

Les logiciels usités pendant mon stage sont très hétéroclites comme diversifiées ont été les tâches que j'ai pu accomplir. La quasi totalité des outils que j'ai utilisés sont des outils d'environnement Microsoft, tout comme la plate-forme de travail (NT 2000) ou des outils disponibles sur le web. Voici donc un panorama de ces logiciels et un petit explicatif de l'utilisation de chacun d'entre eux. La mention « TEST » précise que ce logiciel a été testé, sans forcément être utilisé dans l'activité de référencement (s'il a été jugé comme mauvais ou inadapté).

- **Les classiques Microsoft:** Word, Excel et la messagerie outlook
- **Le**(s) **classique**(s) **Macromédia** : Dreamweaver
- **Adobe :** visionneuse PDF
- Visionneuse Post script
- **autres:** Ultra-Edit (éditeur), Visio (arborescences et diagrammes Microsoft)
- **Agent web ranking** : utilitaire du référencement
- **All in 1 submission**: Aide à la rédaction de contenu et aide au référencement

- **GRKDA**:densité de textes.
- **Windex**: indexation automatique.
- **gkwproi**: interface "Good keywords" liée à Overture, logiciel qui permet de se connecter à la base de données d'OVERTURE.
- **Termextraxtor** : extracteur de termes et d'associations de termes. On ouvre un fichier et le logiciel en extrait les termes de prédilection.
- **Copernic Summarize** : logiciel de résumé automatique.
- **skyfox** : logiciel de soumission automatique.
- **human-links** : Logiciel devant permettre de cartographier des textes en ligne.
- **linkpromoter.html**: Logiciel de soumission automatique.
- **GDI DB** : logiciel de transformation de pages dynamiques (a partir de base de données) en pages statiques.

EN TEST

Supports explicatifs et pistes :

En savoir un peu plus…

Ce document est destiné à fournir quelques explications, quelques pistes concernant les différents sujets abordés dans le rapport, mais ne prétend en aucun cas répondre à toutes les questions que vous pourrez vous poser en le lisant.

Il regroupe donc tout ce qui a pu me sembler intéressant de citer, des résultats d'instituts de sondages, des articles de journaux Internet, réflexions personnelles et des documents de référence, dont les sources, s'il y a lieu, sont explicitement citées.

- **Qu'est ce qu'un cyber-documentaliste ?**

Comme son nom l'indique, le cyber-documentaliste est un (ou une) documentaliste qui travaille avec l'outil de stockage et de communication qu'est Internet. Aussi appelé Surfeur, son rôle est d'opérer le même rôle de classification de documents qu'un documentaliste classique, mais avec des documents numériques soumis par des internautes.

Une fois le site soumis[12], il est visité, dans un délai plus ou moins long, par ces fameux cyber documentalistes qui valident (ou réécrivent) la description proposé par l'internaute, ainsi que la catégorie de l'annuaire dans laquelle il apparaîtra.

- **Quel différence entre un moteur et un annuaire ?**

Annuaires, généralités

Un annuaire est un répertoire, une base de données construite manuellement, qui indexe selon des rubriques les sites qui leur sont soumis. Comme un annuaire (pages jaunes) classique :

Un site (une URL) est soumis à un annuaire dans une rubrique, il est visité par un être humain qui va juger et décider de la rubrique dans laquelle il apparaîtra (C'est le rôle du Net-surfeur que nous venons de décrire). L'internaute navigue sur l 'annuaire au travers des rubriques par exemple.

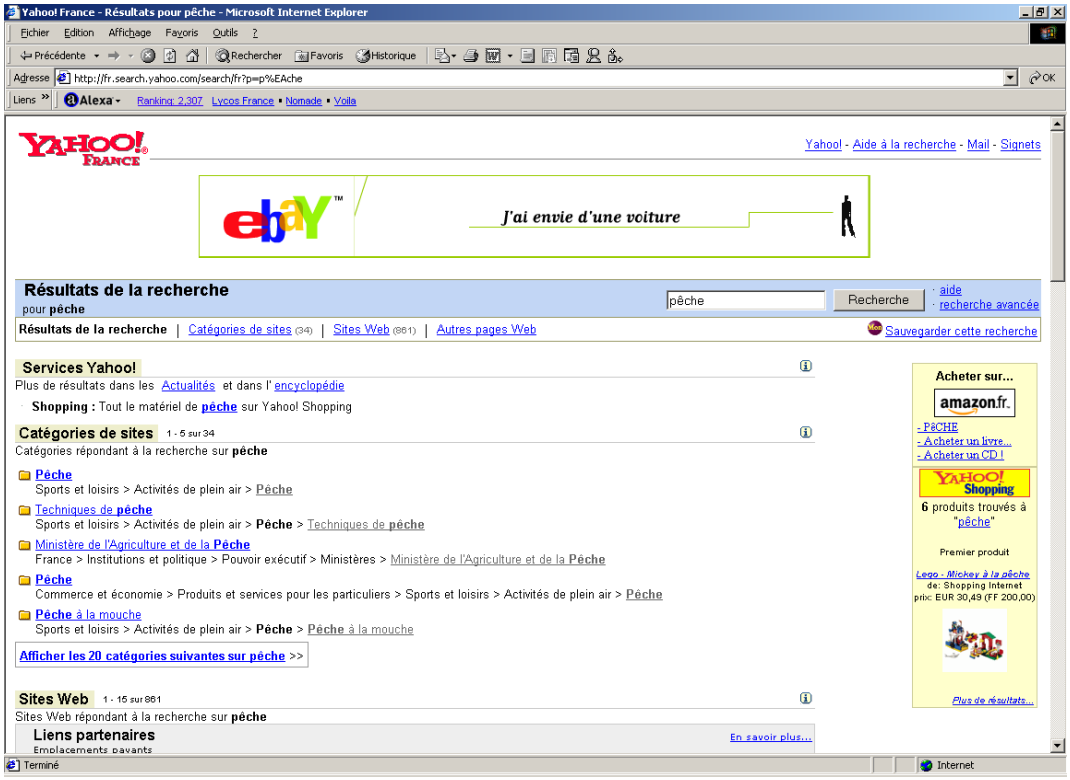


Figure 3 : L'utilisateur des annuaires peut formuler sa requête sous forme de mots

[12] Une fois que l'internaute à inscrit l'adresse de son site dans les listes de sites désireux d'entrer dans le moteur.

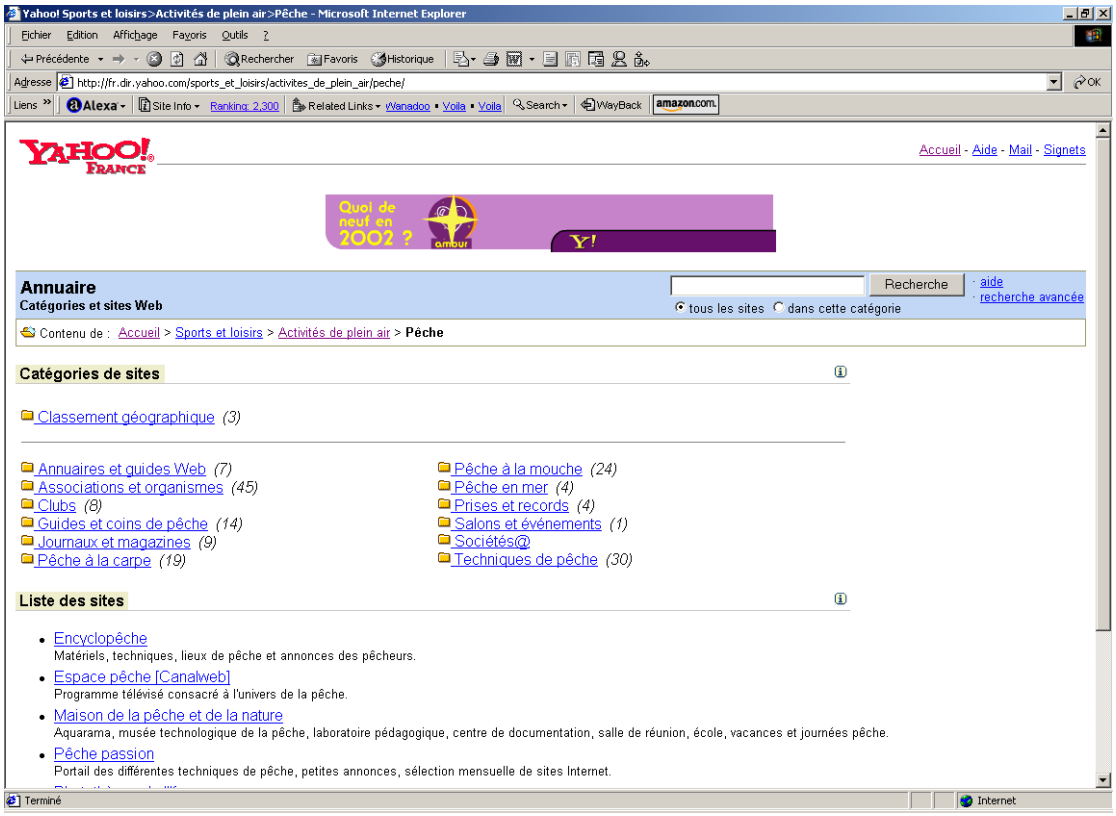


Figure 4 : Ou en se promenant dans les rubriques de la Home Page

Deux méthodes pour accéder à une seule information, deux résultats différents.
Un annuaire présente donc une liste de sites classés dans des thématiques qui doivent guider l'internaute dans ses recherches, mais il permet aussi une recherche plus « sauvage », identique à celle d'un moteur.

Moteurs, généralités

Un moteur est un outil de recherche qui indexe les pages d'un site, mais pas uniquement la page d'accueil, mais aussi toutes les autres pages du site (contrairement à un annuaire, qui ne répertorie que la home page d'un site). Chaque moteur de recherche possède ce que l'on appelle un (plusieurs) spider (une araignée* - une chenille) qui va visiter les pages et qui stocke sur un serveur (plusieurs) une image des pages Internet qui lui sont soumises. (Utiliser l'option en cache de Google pour voir quelle version d'un site a été enregistrée, cf ; p34).

Lors d'une requête, les Moteurs renvoient des listes de liens correspondants à la demande. Pour ce faire, le moteur sollicité va parcourir sa base de données et en extraire les pages qui semblent répondre à la requête de l'internaute. Il les ordonne selon un algorithme* propre à chaque moteur (ce qui peut expliquer qu'une page soit très bien placée sur Voilà et mal sur Google) et les affiche.

Meta moteur

Un méta moteur* est un outil de recherche qui permet d'effectuer une recherche sur plusieurs moteurs de recherche simultanément. Il affiche les meilleurs des résultats de chacun d'entre eux. Ce qui doit permettre en théorie d'avoir des résultats plus pertinents (puisque censés être un « Best of »), (figure 5.)

Les Méta-moteur permettent de synthétiser une information de recherche et évitent à l'internaute de consulter plusieurs O-R, il évite également les doublons.

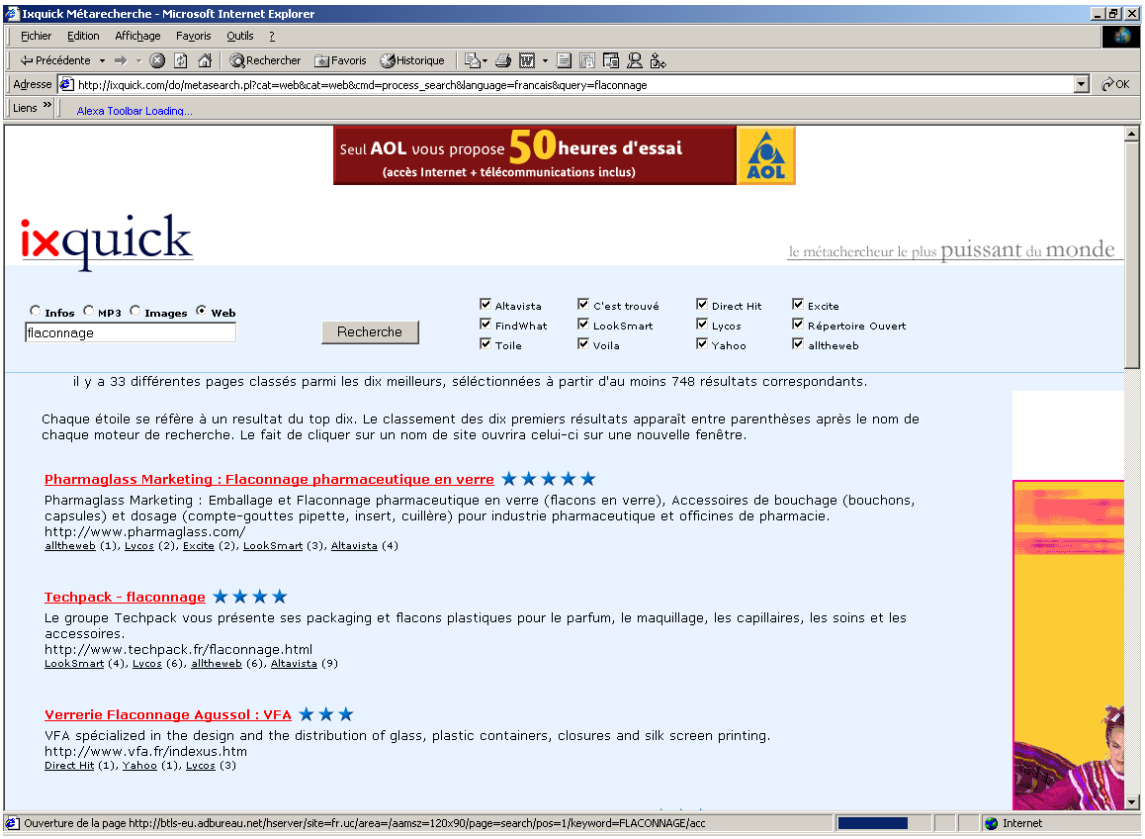


Figure 5 : IXQuick, méta moteur

Autres[13] :

Les moteurs sémantiques généralistes	
Hummingbird EIP, éditeur	Capable d'effectuer des recherches internes ou externes en intégrant des fonctions d'analyse linguistique, le moteur EIP, issu du rachat de l'éditeur Fulcrum, indexe les données structurées et non-structurées.

[13] SOURCE : Le journal du Net , http://solutions.journaldunet.com

Verity Portal One ou K2 Enterprise, éditeur	Prenant en charge plus de 225 formats de documents, il s'agit d'un moteur de recherche plein texte offrant nombre d'opérateurs (booléens, de proximité, etc.) et de fonctions sémantiques (synonymie, termes connexes, etc.).
Arisem OpenPortal4U, éditeur	Moteur sémantique multilingues, OpenPortal repose sa logique sur un référentiel de concepts. Intégrant des fonctions de personnalisation du processus de recherche, il explore les liens hypertextes associés à une source Web.
Sinequa Intuition,editeur	Intuition repose notamment sur un moteur d'analyse morpho-syntaxique. Fort de fonctions d'analyse sémantique en français, allemand et espagnol, il est capable de contextualiser les termes d'une requête en langage naturel .
Convera RetrievalWare éditeur	Ce moteur multilingues fournit des fonctions de recherche floue reposant sur l'analyse de la valeur binaire des séquences de lettres. Prenant en compte les formats image, il peut être enrichi de thésaurus, et offre la possibilité de restreindre une requête à certains domaines.
Inktomi Inktomi Search	Prenant en charge 35 langues, Inktomi propose également un analyseur de requêtes en langage naturel. Incluant des thésaurus personnalisables, ce moteur de recherche plein texte gère les termes connexes.
Les moteurs sémantiques généralistes/Web	
Atomz Atomz	Disponible en mode hébergé, ce moteur de recherche plein texte est particulièrement adapté à une problématique de site Web et à de fortes montées en charge. Intégrant des fonctions sémantiques, il est capable d'indexer jusqu'aux documents au format Flash.
Les moteurs de recherche statistiques	
TripleHop MatchPoint , éditeur	Nouveau venu sur le segment des moteurs statistiques, MatchPoint génère des réseaux de concepts par le biais de fonctions statistiques. Il couple ce système de listes avec des fonctions vectorielles afin de répondre à des requêtes multicritères.
Autonomy KnowledgeServer, éditeur	Ciblant les domaines du Knowledge Management, du commerce électronique et du newsmedia, ce moteur utilise aussi des algorithmes probalistiques indépendants des langues -s'appuyant notamment sur les théories de l'inférence bayésienne et de l'information (Shannon).
Les moteurs de recherche multi-dimensionnels	
Instranet Instranet 2000, éditeur	Plate-forme de gestion de contenu Web, Instranet 2000 cumule un moteur multidimensionnel, qui permet de manipuler les informations sur plusieurs axes sémantiques, et un moteur de recherche plein texte (Oracle).
Les assistants de requêtes	
LexiQuest LexiQuest , éditeur	LexiQuest a choisi de venir compléter les solutions existantes. Objectif : enrichir le travail de base des moteurs avec des fonctions d'analyses grammaticales et

	syntaxiques (anglais et français), des réseaux sémantiques, et un traducteur booléen.
Albert Albert Meaning Interpreter(AMI)	S'adaptant à divers outils (Verity, Fast, etc.), AMI analyse la demande de l'utilisateur afin de constituer une requête qui soit compréhensible par les système d'indexation. Pour ce faire, il s'appuie sur une base de connaissances personnalisée qui s'auto alimente au fil des réponses renvoyées.
Alogic Alcalimm	Grâce à son assistant visuel, Alogic permet d'affiner une demande. Couplant thésaurus et indexation plein texte, ce moteur repère dans ses bases les mots contenus dans les requêtes, ainsi que les expressions débutant par l'un de ces mots. L'utilisateur validant ensuite les concepts qui l'intéressent.
Auracom Auraweb	Traitant données structurées et non-structurées, Auraweb offre de nombreuses fonctions d'assistance à la recherche: calcul du nombre de réponses disponibles, liste de des expressions de l'index contenant un ou plusieurs mots clés demandés, etc.
Les moteurs de recherche intégrés	
Mediapps Net.Portal	LIEN MORT*
Opentext Livelink	LIEN MORT*
Microsoft Sharepoint Portal Server	S'appuyant sur Microsoft Index Server, ce moteur assure une indexation plein texte des documents contenus sur des serveurs Web, des arborescences de fichiers et des bases Notes. Des algorithmes de classement et des filtres de recherche permettant d'affiner les requêtes.
Reef Real Knowledge	Le module de recherche de cette solution de gestion documentaire s'articule autour du moteur de Microsoft SQL Server, ou de celui de Verity, et le complète d'une couche d'algorithmes : une requête sera par exemple traitée en fonction de l'audience des documents.

Petite note sur les moteurs en langage naturel

Un moteur en langage naturel est un moteur qui doit être capable d'interpréter un requête faite en langage naturel :

- **AskJeeves**[14] (*figure 6 & 7*)

L'internaute fait une requête en langage naturel, une requête de type « How to make a pancake ? » pour notre exemple. Dans ses réponses, Askjeeves nous fait une proposition déduite de l'interprétation de la requête en langage naturel.

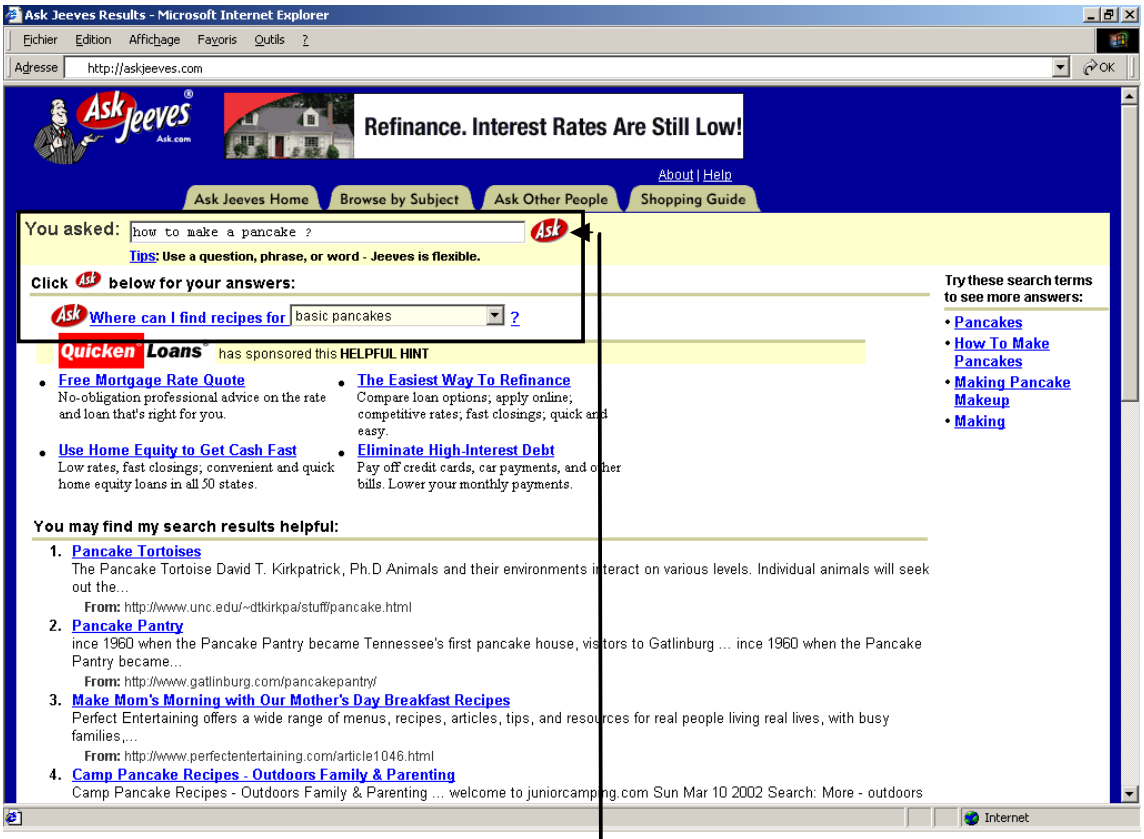


Figure 6 : Exemple d'un moteur en langage naturel

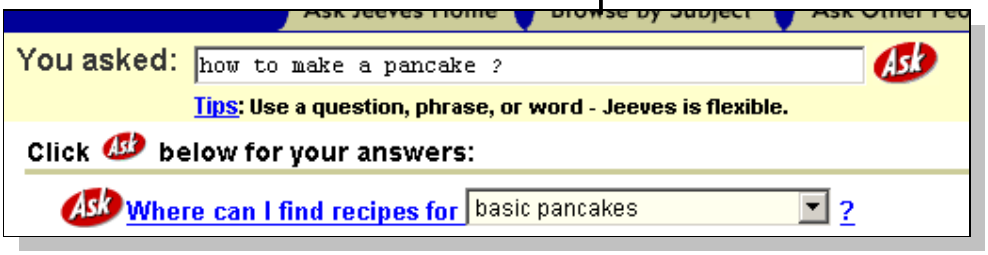


Figure 7 : Détail d'ASKJEEVES

Pour ce type de moteurs, ce n'est pas tant la façon de retrouver les documents stockés qui diffère du moteur « classique », mais plutôt la façon dont est interprété la requête. Parmi d'autres moteurs en langage naturel : INFOCLIC[15], MAXIMINI[16]

[14] http://askjeeves.com
[15] http://www.infoclic.fr/
[16] http://maximini.net/

- **Le langage HTML**

Le langage HTML est langage compris par le navigateur que vous utiliser. Quand vous créez une page, un site (enregistré en .html ou .htm) avec des outils WYSIWYG tels que DREAMWEAVER ou même FRONT PAGE ou WORD, vous générez un fichier dont le langage est du HTML (Hypertext Mark-up language). Lorsque vous visitez un site, pour avoir son code source (le code HTML brut), il suffit de cliquer sur *AFFICHAGE > SOURCE (figure 8)*. Vous obtenez un fichier de texte plus ou moins compliqué (*figure 8*). Le langage en lui même est assez simple dans son fonctionnement puisqu'il consiste à apposer des propriétés à des éléments.

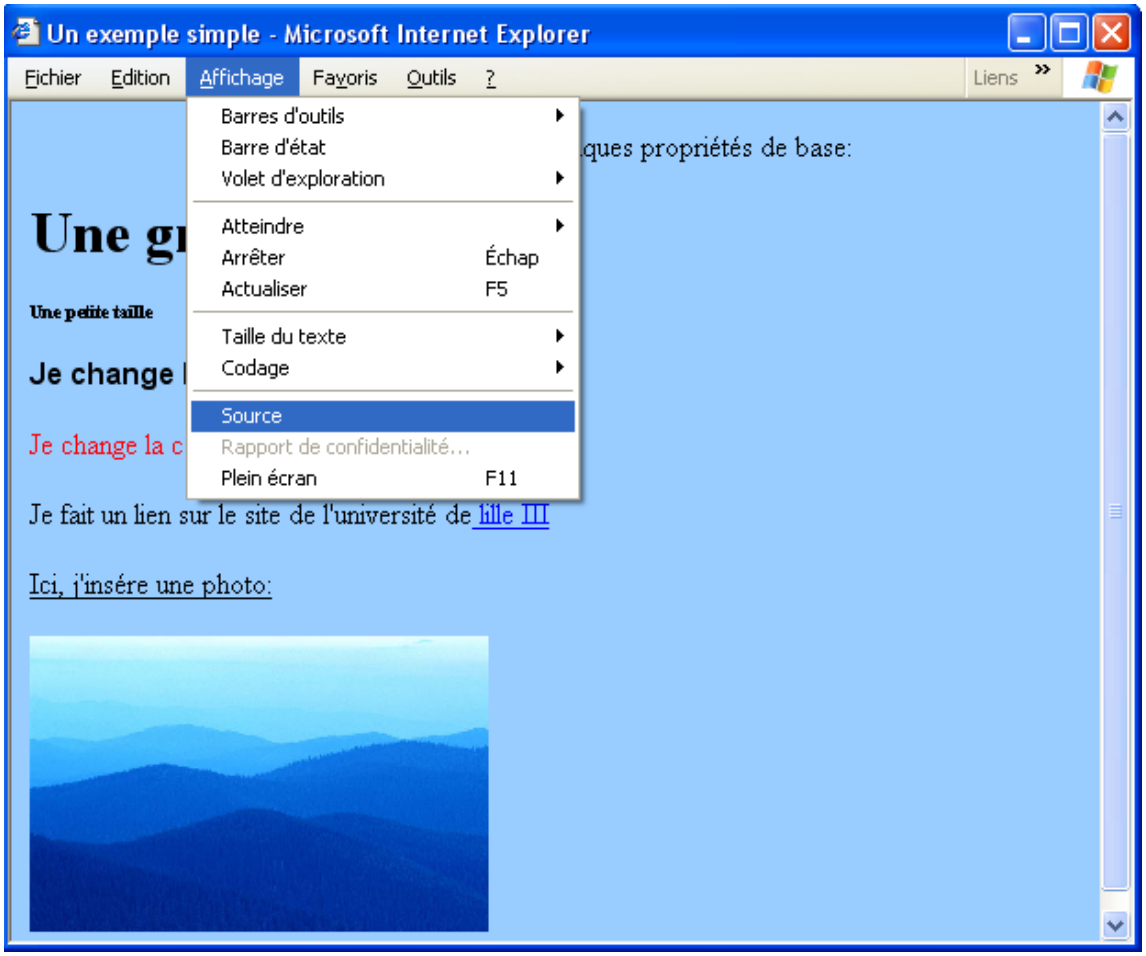


Figure 8 : exemple d'une page toute simple en HTML

- Petit exemple simple commenté qui ouvert dans le navigateur nous donne la page ci-dessus (*figure 8*):

```
<html> <!--ceci est un commentaire --> <!—Je dis que j'utilise le langage HTML -->
    <head> <!—première partie, la balise head dans laquelle je vais déclarer en quelque
    sorte mon  document, c'est aussi dans cette balise que j'insère mes méta données (voir
    point  sur les méta données) -->
        <title>Un exemple simple</title> <!—Le titre du document (qui va apparaître en
        haut à gauche dans la barre du navigateur -->
        < content="text/html; charset=iso-8859-1"> <!—les formats éventuellement
        contenus dans le code et la table de caractère utilisée iso-88-59-1 -->
    </head> <!—Je ferme ma balise head, j'ai fini de déclarer mon document -->

    <body bgcolor="#99CCFF" text="#000000"> <!—J'entre dans le corps de mon document -->
        <p <!—La balise P indique un nouveau paragraphe --> align="center" <!—Dont
        la propriété d'alignement est d'être centré -->>Voici un petit texte et les
        quelques propri&eacute;t&eacutes; de base:</p  <!—Je ferme la balise de
        nouveau paragraphe-->>
            <h1>Une grande taille</h1> <!—J'affecte au texte entre les balise la
            propriété H1 -->
            <h6>Une petite taille</h6> <!—J'affecte au texte entre les balise la
            propriété H6 -->
        <p><font face="Geneva, Arial, Helvetica, san-serif" <!—J'affecte au texte entre
        les balise le changement d police --><b <!—J'affecte au texte entre les balise la
        propriété Bold (gras) -->>Je change la police</b></font <!—Fin de changement
        de police (font) -->></p>
        <p><font color="#FF0000" <!—Je change la couleur de la police grâce à un code
        couleur -->>Je change la couleur</font></p>
        <p>Je fait un lien sur le site de l'universit&eacute; de<a href="http://www.univ-
        lille3.fr"<!—J'utilise la fonction 'a href=' pour faire un lien  --></>
        lille III</a <!—Je ferme la balise a href pour dire que mon lien s'arrête aux mots
        entre les balises -->>p>
        <p><u<!—Balise de soulignement (underline)-->>Ici, j'ins&eacute;re une
        photo:</u></p>
        <p><img src="images/collines.jpg" <!—J'utilise la propriété 'img src=' (image
        insertion) en donnant le chemin d'accès de cette image -->></p>
    </body> <!—Je ferme la balise de mon corps de texte -->

</html> <!—Je ferme mon document HTML -->
```

- **Les méta données**

Les méta données sont des informations sur les contenus du document aussi bien au niveau du contenu textuel que les types de ressources que le document utilise, son auteur, la date de création, le format, la source, mots clefs…Ces informations ne sont pas visibles dans le document lui même, mais font office de données « satellitaires ».

La liste ci dessous est un extrait du modèle de méta données proposé lors du **Dublin CORE** [17], meeting dont la mission est d'établir, de proposer des normes pour améliorer la recherche d'information sur Internet dans le but de pouvoir générer de manière uniforme et pertinente des indexes automatiques ou des catalogues de documents.
A l'heure actuelle peu de référenceurs utilisent toutes les méta données possibles et les balises les plus fréquentes sont celle de la description, du titre et des mots clefs.

Title

```
<meta name    = "DC.Title"
     content = "Hamlet in Iceland; being the Icelandic romantic Ambales saga">
<meta name    = "DC.Title.Alternative"
     content = "Ambales saga">

<meta name    = "DC.Title.Alternative"
     lang    = "en"
     content = "The Well-tempered Klavier">
```

Subject (topic or keyword)

```
<meta name    = "DC.Subject"
     content = "heart attack">

<meta name    = "DC.Subject"
     scheme  = "MeSH"
     content = "Myocardial Infarction; Pericardial Effusion">

<meta name    = "DC.Subject"
     content = "Vietnam War">

<meta name    = "DC.Subject"
     scheme  = "LCSH"
     content = "Vietnamese Conflict, 1961-1975">

<meta name    = "DC.Subject"
     content = "Friendship">

<meta name    = "DC.Subject.Classification"
     scheme  = "DDC"
     content = "158.25">
```

Description (textual description or abstract)

```
<meta name    = "DC.Description.TableofContents"
     lang    = "en"
```

http://dublincore.org

```
        content = "The Author gives some Account of Himself and Family
                -- His First Inducements to Travel -- He is
                Shipwrecked, and Swims for his Life -- Gets safe on
                Shore in the Country of Lilliput -- Is made a
                Prisoner, and carried up the Country">

    <meta name    = "DC.Description.Abstract"
        content = "The kinematics of the jaws and hyolingual apparatus in
                Caiman crocodilus were examined by cineradiography and
                electromyography. After catching, caimans position their
                prey between the teeth by a series of inertial bites and
                then kill and crush it by a forceful bite.">
```

Date (of creation or availability of resource; [WTN8601] recommended)

```
    <meta name    = "DC.Date.Created"
        content = "1998-05-14">
(...)
```

Type (category or genre)

(...)

Format (data format, plus optional dimensions)

(...)

Identifier (of the resource)

```
    <meta name    = "DC.Identifier"
        scheme  = "URI"
        content = "http://catalog.loc.gov/67-26020">
```

Language (of the content of the resource; [RFC1766] recommended)

```
    <meta name    = "DC.Language"
        content = "en">
    <meta name    = "DC.Language"
        content = "en;fr">
(…)
```

Relation (relationship to a second resource plus its identifier)

(...)

Coverage (spatial or temporal features of the intellectual content)

```
    <meta name    = "DC.Coverage.Temporal"
        content = "US civil war era; 1861-1865">
    <meta name    = "DC.Coverage.Spatial"
        content = "Columbus, Ohio, USA; Lat: 39 57 N Long: 082 59 W">
  (...)
```

- **Petite note sur l'indexation des images**

Pour l'instant, aucun moteur ne couple réellement indexation automatique des images et indexation des textes bruts, et ce n'est de toute manière pas le sujet de ce rapport. Néanmoins, il peut être intéressant de souligner cet aspect qui, dans l'avenir, pourrait révolutionner, sinon améliorer l'indexation et la recherche des documents sur le Web (et sur tout support numérique). On retrouve cette technologie à différents niveaux, de la reconnaissance « simple » de chaînes de caractères à une reconnaissance plus complexe de formes.

Pour la reconnaissance « simple », un bon exemple sur le moteur de Google à l'adresse suivante : http://catalogs.google.com , répertoire qui utilise une technologie de reconnaissance de caractères, (version beta).

Et pour en savoir plus sur cette technologie qui se base sur la reconnaissance de forme par le contraste : Le site[18] du LTU Technologie qui conçoit ce type de produit et en donne certains exemples d'utilisation.

- **La charte de déontologie**

On a souvent vu (et on verra toujours), des référencements frauduleux, un site indexé sur un thème fédérateur de requête (Souvent sexe ou musique gratuite ? voir point sur les abus de référencement). Pour palier à ce genre de comportement qui gêne en tout la pertinence des réponses des outils de recherches, l' **I.P.E.A** [19], Internet Positioning European Association, a crée et mis en place une charte de déontologie dont les adhérents s'engagent en quelque sorte dans une charte de qualité. Vous trouverez ci-dessous une copie de cette charte, mais elle est également disponible sur le site de l'association :

Les signataires de la "Charte de Qualité et de Déontologie sur le référencement de sites web" devront accepter les conditions suivantes (O-R = Outils de Recherche) :

Titre	Contenu
Réalisme	Les signataires s'engagent à une obligation de moyens à mettre en oeuvre et à ne pas promettre de résultats de positionnement issus d'un seul mot-clé générique, et en règle générale, ne pas promettre de résultats qui ne pourront être tenus ou vérifiés par le client.
Transparence	Les signataires s'engagent à tenir à disposition de leurs clients - et à leur faire connaître sa disponibilité - un document présentant leur méthodologie de travail.

[18] http://www.ltutech.com
[19] http://www.ipea.com

Conseil	Les signataires s'engagent à aider leurs clients dans la réflexion sur les informations qui seront fournies aux O-R (descriptif, mots clés, etc.) lors du référencement.
Méthodologie	Les signataires restent libres de la méthodologie mise en place pour référencer les sites de leurs clients, à partir du moment où elle respecte la présente charte, notamment en ce qui concerne la lutte contre le spam.
Loyauté	Les signataires s'engagent à suivre strictement les indications des O-R, publiées de façon spécifique sur leurs sites, dans le but d'effectuer une soumission efficace d'un site web dans leur index ou bases de données.
Suivi	Les signataires s'engagent à remettre à leurs clients de façon périodique des rapports clairs sur l'avancée des travaux de référencement de leur site web, sous la forme qui leur semble la plus appropriée.
Autonomie	Les signataires s'engagent à remettre tous les éléments relatifs aux travaux réalisés pour l'inscription dans les O-R de façon à permettre à leurs clients de changer de prestataire s'ils n'étaient pas satisfaits de la prestation effectuée. Comme pour toute prestation informatique ou de service, les travaux effectués appartiennent au client qui en a payé le montant.
Veille	Les signataires s'engagent à mettre en place des mécanismes de veille afin de se tenir au courant de l'évolution des outils de recherche.
Qualité	Les signataires s'engagent auprès des O-R à ne soumettre à leur indexation que des sites dont le contenu et la pertinence sont suffisamment riche pour alimenter leur base de données en vue d'apporter une information utile au visiteur.
Mode de fonctionnement	Les signataires s'engagent à n'effectuer pour leurs clients que des prestations de référencement manuel, sans l'aide d'aucun logiciel de soumission automatique, sauf dans le cas où cette prestation est explicitement indiquée dans la proposition commerciale, et uniquement si l'utilisation de logiciels n'intervient qu'en complément d'une prestation

	manuelle majeure et ce aussi bien en phase de référencement qu'en phase de suivi et de veille.
Respect de la concurrence	Les signataires s'engagent à ne pas nuire au référencement d'un concurrent pour le compte d'un client et à ne pas utiliser la marque de concurrents pour le référencement de leurs clients. En règle générale, les signataires s'engagent à ne pas nuire au référencement d'un site pour lequel ils n'auraient pas été mandatés.
Combat contre le spam	Les signataires acceptent de ne pas réaliser d'action de spamdexing (fraude sur les O-R). Une annexe (voir ci-après) décrit de façon précise ce qui est considéré comme du spamdexing et ce qui n'en est pas dans le domaine du référencement de site web.
Information	Les signataires s'engagent à remettre à leurs clients et prospects, dans leurs propositions commerciales, un exemplaire de la Charte ici présente, accompagnée d'un document expliquant le fonctionnement des O-R, détaillant en quoi consiste un référencement de site Web, ainsi que ses contraintes. Ce document pourra, par ailleurs, présenter l'association IPEA.

définition du spam :

Pour les Outils de recherche en général :

- Soumettre un site faisant la promotion d'activités illicites ou dégradantes (piratage, racisme, sectes, pédophilie, etc.).
- Soumettre un site inadéquat par rapport à la langue de l'O-R (site en anglais sur Yahoo! France ou Altavista France, par exemple).
- Soumettre un même site sous plusieurs noms de domaine différents (chaque adresse soumise pointant sur la même page) ou sous plusieurs adresses différentes (www.site.com et site.com)

Pour les Annuaires spécifiquement :

- Soumettre un même site plusieurs fois dans des délais inférieurs à 3 semaines.
- Soumettre un site pour des rubriques/catégories non appropriées par rapport au sujet principal du dit site.

- Soumettre un site déjà inscrit sur un annuaire dans une rubrique nouvelle sans utiliser le formulaire de modification s'il existe sur l'O-R.
- Soumettre un site avec un nom "litigieux" n'ayant a priori rien à voir avec l'intitulé "réel" du site et ne servant qu'à des vues de meilleur classement (ABC-Zirconium ou 123-Zirconium par exemple), sauf, bien entendu, s'il s'agit de l'intitulé KBis de l'entreprise.
- Remplir le formulaire de soumission en fournissant des indications fausses, inappropriées (résumé trop long par exemple) ou erronées (adresse e-mail inexistante par exemple) par rapport à la réalité du site.
- Tenter de "soudoyer" un netsurfeur (un cyber documentaliste) sous prétexte que le propriétaire du site est un gros annonceur. Demander un traitement de faveur sous le même prétexte.

Pour les Moteurs de recherche spécifiquement :

- Répétition abusive de mots clés.
- Répétition abusive de balises TITLE, META ou autres (commentaires).
- Répétition abusive de soumissions pour un même site dans des délais trop rapprochés (pour la même page) ou d'un trop grand nombre de pages dans un délai court (pour un même site). D'une manière générale, non-respect des limites indiquées par les O-R.
- Utilisation de mots clés inadéquats avec le contenu du site, utilisation de marques ou de noms sans liaison avec le site ou faisant mention de marques déposées sans autorisation.
- Utilisation de texte invisible (blanc sur fond blanc par exemple).
- Cloaking (technique de délivrance aux spiders* de pages spécifiquement réalisées pour les moteurs de recherche) à des fins de positionnement.
- Soumission multiple (ou actions quelconques) de sites concurrents ou d'entreprises non clientes.
- Copie de code HTML existant sur son propre site.
- Création de pages HTML ayant pour unique but d'augmenter l'indice de popularité* d'un site ou d'une page.
- Pages de redirection (balise Meta Refresh)

D'autre part, les pages alias et/ou satellites sont acceptées par les moteurs de recherche à la condition qu'elles soient réalisées pour pallier des difficultés techniques inhérentes à la construction des sites web référencés et à la structure des moteurs de recherche (pages dynamiques, pages Flash, URL exotiques, etc.) ou que leur contenu présente une pertinence avec le thème de la requête effectuée, dans le respect des différents termes de la présente charte.Enfin, il est autorisé, de façon exceptionnelle, de soumettre sur un annuaire des "parties de sites", sous une url spécifique (sports.site.com ou www.site-sports.com) ou non (www.site.com/sports/), si cette "zone" est indépendante du reste du site et propose un réel contenu pouvant répondre de façon précise aux questions des internautes.

- **Abus de référencement**

Les abus de référencement étaient plus fréquents par le passé mais existent toujours. L'abus le plus manifeste étant de donner dans les mots clés ou la description des mots, qui n'ont rien à voir avec le contenu du site. Pour exemple le site de Françoise de Panaffieu qui en 2000 avait défrayé la chronique en insérant (pas elle directement, mais son webmaster) dans les métas le nom de PAMELA ANDERSON[20] et autres noms de concurrents.

Il y existait donc lors des législatives de 2000 une probabilité de « tomber » sur le site de l'élu RPR[21] en tapant comme requête *Pamela Anderson*. Dans ce cas précis, le webmaster a fait une belle « coquille » puisque son tour de passe-passe a été dévoilé, mais d'autres procédés de ce style sont usités sans jamais être découvert (parce qu'on ne les cherche pas), en politique ou ailleurs.

- **L'option -En cache- de Google**

Le moteur de recherche Google propose dans sa page de résultats l'option – *En cache* grâce à cette option il est possible d'accéder à la page telle qu'elle a été stockée par le moteur au moment de l'indexation. Si elle a changé depuis la dernière visite du spider, on peut néanmoins (par cette option) avoir accès à la vraie page, celle qui à vraiment été indexée par le moteur qui nous la propose comme répondant à notre requête pour laquelle la page de résultats est affichée.

Exemple : les figures 9 et 10, sur **www.google.com** , requête actualités France.

Pour cet exemple, Google veut me diriger vers la page *figure9*, mais l'indexation étant plus ancienne que la date de modification, j'aboutis sur une page qui n'est en fait pas celle que le moteur me proposait (*figure 10*). La fréquence de passage des spiders* peut être un véritable problème et il est assez fréquent d'accéder à un site dont la date d'indexation est antérieure à la réactualisation.

[20] Source : http://www.journaldunet.com/dossiers/elections/020205tags.shtml
[21] http://www.panaf.org

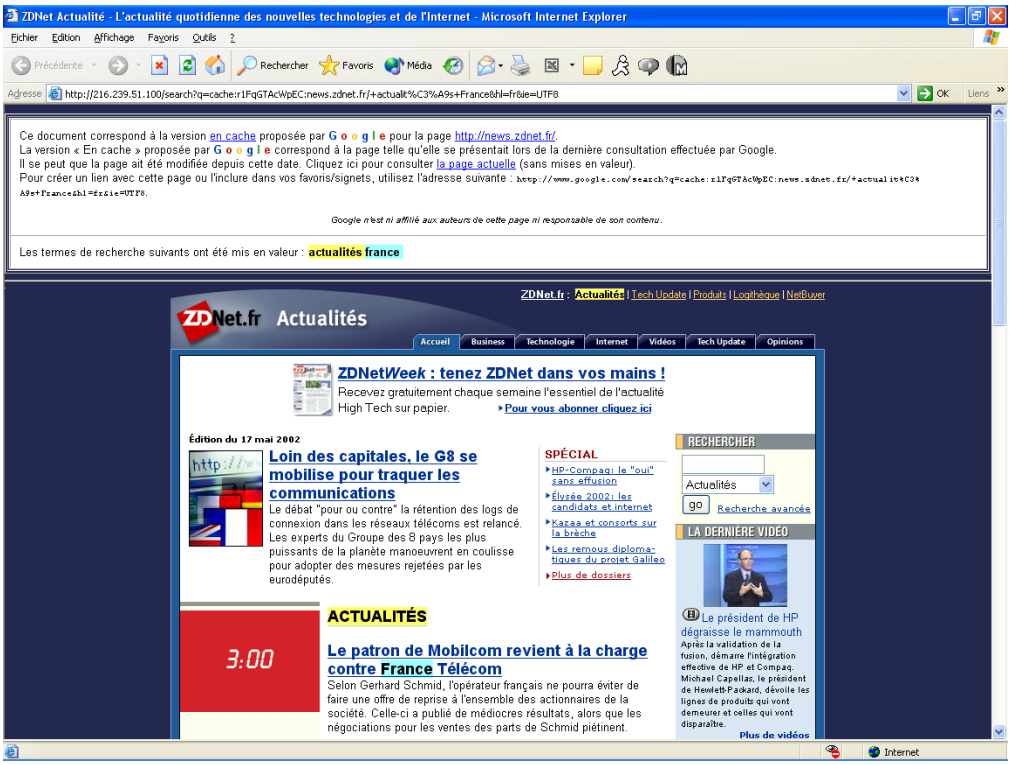


Figure 9: Page d'un site en cache de Google

Figure 10: page telle qu'elle est actuellement

- **Panorama du web actuel**[22]

Légère progression du nombre d'internautes en France

Le nombre d'individus âgés de 11 ans et plus s'étant connectés à Internet au cours du dernier mois a légèrement progressé en avril par rapport à mars 2002 : 16 591 000 individus, soit 32,6 % des Français âgés de 11 ans et plus.

Tableau 1: Progression du nombre d'internautes

	Avril 2002	Mars 2002	Février 2002	Janvier 2002
Nombre d'individus de 11 ans et plus déclarant s'être connectés à Internet au cours du dernier mois, quel que soit le lieu de connexion	16 591 000 soit 32,6 % des Français	16 472 000 soit 32,4 % des Français	16 433 000 soit 32,3 % des Français	16 985 000 soit 33,5 % des Français
Nombre d'individus de 11 ans et plus déclarant avoir utilisé Internet au cours du dernier mois pour télécharger des fichiers (musicaux, vidéo, logiciels, …)	6 170 000 soit 12,1 % des Français	6 527 000 soit 12,8 % des Français	6 080 000 soit 12,0 % des Français	6 632 000 soit 13,1 % des Français

Rappel sur l'équipement multimédia des foyers francais – 1er trimestre 2002

Tableau 2: Equipement des foyers

	1er tri 2002	4è tri 2001	3è tri 2001	2è tri 2001	1er tri 2001
Nombre de foyers équipés micro-ordinateur	8 806 000 soit 35,7% des foyers	8 685 000 soit 35,6% des foyers	8 680 000 soit 35,6% des foyers	8 420 000 soit 34,5% des foyers	7 890 000 soit 32,3% des foyers
Nombre de foyers ayant accès à Internet	5 384 000 soit 21,9% des foyers	5 196 000 soit 21,3% des foyers	4 387 000 soit 18,0% des foyers	4 445 000 soit 18,2% des foyers	4 074 000 soit 16,7% des foyers

Definition

Internautes : la population internaute de référence de ce communiqué est la population des « internautes dernier mois 11 ans et + », soit 16 591 000 individus s'étant connectés à Internet au cours des 30 derniers jours quel que soit leur lieu de connexion : domicile, travail, autres lieux (lieux d'éducation, lieux publics ou privés tels les cybercafés, les bibliothèques, chez des amis…)

Rappelons que « Les Baromètres Multimédia » est l'étude de référence des équipements et usages multimédia pour observer simultanément les taux d'équipement multimédia et le profil des utilisateurs. Elle permet une analyse qualifiée de l'ensemble des usages Internet ou téléphonie mobile selon un nombre important de critères socio-démographiques et comportementaux. Cela en

[22] SOURCE MEDIAMETRIE, http://www.mediametrie.com

fait un outil opérationnel pour anticiper les évolutions rapides des comportements face aux développements des nouvelles technologies.

DEFINITION DES TERMES[23]

Vous trouverez dans les pages suivantes quelques définitions de termes fréquemment utilisés dans les lettres d'actualité et le sites liés au référencement, ce petit lexique n'est pas une production personnelle et peut être trouvé sur le site d'IDF.net

[23] SOURCE : http://www.idf.net

A

Achat de mots-clés
L'achat de mots-clés dans les moteurs de recherche est uniquement possible à travers les bandeaux. Tous les principaux moteurs de recherches (sauf EuroSeek et GoTo) insistent que cet achat de mots-clés n'est lié qu'à l'apparition de bannières et n'influence en rien le résultat des requêtes. Un service "Bannerstake" proposé par Thomson et Thomson à l'adresse suivante : http://www.namestake.com permet de comprendre la logique d'affichage de bandeaux en fonction de la requête faite.

Algorithme de Pertinence
C'est la méthode qu'utilise un Moteur de Recherche ou un Répertoire pour relier les mots-clés d'une requête avec le contenu de chaque page, de telle sorte que les pages Web trouvées correspondent bien au sujet de la requête. Chaque Moteur de Recherche ou Répertoire est susceptible d'utiliser un algorithme différent et de le changer ou de l'améliorer de temps en temps.

Araignée
C'est la partie du Moteur de Recherche qui surfe sur la toile, enregistre les URLs, classe les mots-clés et le texte de chaque page qu'il trouve. En français, le terme plus souvent employé est robot. Vous pouvez trouver plus d'information sur chaque araignée* au niveau du Search Engine Watch.

ArchitextSpider*
C'est le petit nom de l'araignée* du Moteur de Recherche d'Excite.

B

Bait-and-Switch
C'est le fait d'"appâter" le visiteur et de le "transferer" ensuite vers une autre page. Une page d'un certain style est pris en compte par les outils de recherche mais une autre page apparaît à la même URL pour les navigateurs. Diffèrentes méthodes peuvent être utilisée comme la substitution de pages.

Bourrage de Mot-Clé
La répétition des mots-clés et des phrases-clés dans les Méta-tags ou ailleurs.

C

Client
Un ordinateur, un programme ou un processus qui demande des informations à partir d'un ordinateur, d'un programme ou d'un processus différent. Les navigateurs sont des programmes-clients. Les robots des Moteurs de Recherche sont des clients (ou peut-être on leur dit de se comporter comme tel!).

Cliquez à travers
Le fait de cliquer sur un lien dans un Moteur de Recherche et de sortir de la page pour visiter un autre site.

Contenu dynamique
Ce sont des pages web avec des informations qui changent ou sont changées automatiquement en fonction d'une base de données ou d'éléments provenant de l'utilisateur. Il est possible, certaines fois, de se rendre compte que cette technique est utilisée quand l'URL fini avec les extensions suivantes: .asp, .cfm, .cgi ou .shtml. Il est aussi possible d'avoir des pages avec un contenu dynamique et finissant avec les

extensions habituelles à savoir .html ou .htm. Les moteurs de recherche référencent ces pages dynamiques de la même manière que les pages avec un contenu statique. Attention, les adresses qui contiennent le caractère ? ne sont généralement pas indexées.

Crawler
ou Chenille, cf Araignée*

D

Délistage
La suppression d'un site d'un Moteur de Recherche.

La suppression peut s'expliquer par différentes raisons comme le manque de fiabilité de la machine qui héberge le site ou en raison de tentatives perçues de spamdexing.

Densité* des Mots Clés
Une des propriétés qui permet d'indiquer l'importance de certains mots dans le texte d'une page Web. Certains Moteurs de Recherche utilisent cette propriété pour le positionnement. Des analyseurs sont disponibles pour effectuer des comparaisons entre les pages. Des pages peuvent être alors créées avec la même densité* de mots clés que celles qui obtiennent un bon classement dans les résultats des Moteurs.
La formule de calcul de la densité* est simple en fait: nombre d'occurrences du mot demandé / nombre de mots total de la page.
Imaginons qu'une page Web comporte 10 mots dans son texte visible. Si la page contient une fois le mot microsoft, ce mot clé aura une densité* de 10%. Si le mot est

présent deux fois, sa densité* passe à 20%. Si les dix mots sont égaux et sont donc microsoft, la densité* de ce mot passera à 100%.

Autre façon de voir les choses : Si la page ne contient qu'un mot, celui-ci reçoit automatiquement une densité* de 100%. S'il y a deux mots différents, ils ont chacun une densité* de 50%, etc.

A noter cependant : le calcul ne prend pas en compte les "stop words" ou mots inutiles (at, the, I, of, to, etc.).

L'indice de densité* des mots clés est l'un des critères majeurs pris en compte lorsqu'on crée une page satellite (ou page alias), c'est-à-dire une page créée spécifiquement par rapport à un moteur de recherche précis.

La formule de calcul est quasiment toujours la même. Par contre, les moteurs diffèrent au niveau de l'importance donnée à cet indice.

Il existe un logiciel intitulé Keyword Density Analyzer qui vous permet d'analyser l'indice de densité* d'une page, ce qui fait gagner pas mal de temps.

Description
Le texte descriptif, qui est associé à une page web, est celui affiché (habituellement le titre de la page et son URL) quand un visiteur fait une demande à travers un Moteur de Recherche ou un Répertoire. Certains Moteurs de Recherche prennent cette description au niveau des méta tags, d'autres génèrent leur propre texte à partir de la page. Les Répertoires demandent souvent le texte à l'enregistrement.

Directory
Cf Répertoire

Dissimulation
C'est le fait de cacher, de dissimuler le contenu de ses pages. L'objectif est de généralement (je précise bien, généralement) de dissimuler des pages optimisées aux voleurs de contenu !!! Cf aussi Bait-and-Switch.

E

Ecriture
L'écriture du texte pour une page Web. Semblable à l'écriture pour n'importe quel autre type de publication, l'écriture correcte peut avoir un meilleur effet sur le Moteur de Recherche. Ainsi, l'écriture forme une partie importante de l'optimisation.

Enregistrement
C'est l'action d'informer un Moteur de Recherche ou un Répertoire qu'une nouvelle page ou un nouveau site doit être indexé.

En-Tête
Plusieurs moteurs de recherche donnent plus d'importance et de poids au texte trouvé entre les commandes d'en-tête au niveau du html (heading tags en anglais). Il est généralement conseillé d'utiliser ces commandes d'en-tête dans une page web et d'y mettre des mots clés dedans. (<h1> à <h6>)

F

Fautes d'Orthographe
Les gens font souvent des fautes de frappe ou d'inattention dans leurs requêtes en utilisant les Moteurs de recherche. Les pages qui prennent en compte ces erreurs reçoivent souvent des hits supplémentaires. La technique consiste à inclure ces erreurs au niveau de la balise alt, des mots clés, du nom de la page et du titre. Un effet similaire est l'oubli d'espace entre deux mots.

Ferme de popularité des liens
Ce sont des pages qui contiennent de nombreux liens. L'objectif est d'augmenter la popularité dans les outils de recherche des

sites qui ont un lien sur ces pages.

Fichier Log
C'est un fichier présent sur votre serveur et qui conserve les informations concernant les accès sur votre site. L'analyse de votre fichier log peut-être un moyen très efficace pour mieux connaître le visiteur qui vient voir votre site. Il est possible de savoir d'où il vient, les pages qu'il a consulté, les mots-clés qu'il a utilisé dans sa requête, etc. Il existe de nombreux programmes pour analyser un fichier log. Il y a NetTracker proposé par Sane Solutions, WebLog proposé par Awsd, etc, etc.

Frames
Cf Trames

G

Gulliver
C'est le petit nom de l'araignée* du Moteur de Recherche de Northern Light.

H

HTTP
HyperText Transfer Protocol - le (principal) protocole de communication entre les serveurs web et les navigateurs (clients).

I

Image Cliquable
Cf Image Map

Image Map
C'est une série de liens hypertextes attachés à une image. Ils sont définis dans la page ou à travers un fichier externe.

Si l'image cliquable est définie comme un fichier externe, les Moteurs de recherche peuvent avoir des problèmes pour indexer vos autres pages, à moins que vous ayez défini aussi des liens hypertextes plus conventionnels.

Si l'image cliquable est définie dans votre page, les Moteurs de recherche n'auront pas de problèmes pour suivre les liens. Il est quand même

conseillé de fournir aussi sur votre page des liens au format texte pour aider ceux qui ont des problèmes de vision ainsi que ceux qui accèdent au site sans les graphismes ou en utilisant un navigateur texte.

Index
Cf Répertoire - Fait aussi référence à la base de données qui contient les pages web d'un Moteur de Recherche et/ou d'un Répertoire.

J

Javascript
Un langage simple interprété qui permet d'executer des petites taches au sein de vos pages web en HTML. Les scripts sont normalement interprétés (ou exécutés) sur l'ordinateur client (du visiteur) par le navigateur. Quelques moteurs de recherche ont, à un certain moment, indexés ces scripts, sûrement par erreur.

Juxtaposition
Le rapport entre les mots dans une requête dans un Moteur (ou Répertoire) de Recherche. Les

Moteurs de Recherche permettent souvent aux utilisateurs de préciser si les mots doivent être l'un à côté de l'autre ou séparés dans les pages Web recherchées.

K

Keyword
Cf Mot Clé

Keyword Density
Cf Densité* des Mots Clés

Keyword Domain Name
Mots Clés dans l'URL

L

Lien Mort
Un lien qui ne mène plus à une page ou à un site, probablement parce que le serveur est en panne ou que la page a été déplacée ou alors n'existe plus. La plupart des Moteurs de Recherche ont des techniques pour retirer de telles pages de leurs listes automatiquement. Internet

continuant à augmenter en taille quotidiennement, il devient de plus en plus plus difficile pour un Moteur de Recherche de contrôler régulièrement toutes ces pages. Reporter des liens morts aide à maintenir les Moteurs propres et précis. On peut le faire en soumettant le lien mort au Moteur de Recherche.

Link popularity farms
Cf Ferme de popularite des liens

Liste des Canaux
Ce sont des listes de liens de sites Web pré-choisis (et généralement, les plus populaires). Ces listes sont mises à jour par des Moteurs ou des Répertoires de recherche et sont triés à travers des catégories ou des canaux. Les sites sont sélectionnés par le responsable du canal, souvent en raison du bon placement du site dans les Moteurs de Recherche. Certains Moteurs, annuaires autorisent les visiteurs à nommer des sites pour les inclure dans leurs listes.

Log File
Cf Fichier Log

M

Méta-Moteur de Recherche
Un serveur qui passe des requêtes à plusieurs moteurs de recherche et/ou répertoires et résume les résultats. Ask Jeeves, Debriefing, Dogpile, Infind, Metacrawler, Metafind et Metasearch sont des exemples de méta-moteurs de recherche.

Méta Recherche
La recherche des recherches. Une requête est soumise à plus d'un Moteur de Recherche ou Répertoire. Les résultats de tous les moteurs sont affichés après élimination des doubles et un triage.

Méta Tag
Une construction placée dans l'entête HTML de votre page Web, fournissant des informations qui ne sont pas

visibles par les navigateurs. Les méta-tags les plus courants (et les plus utiles pour les Moteurs de Recherche) sont KEYWORDS (mots-clés) et DESCRIPTION.

Le méta-tag KEYWORD permet à l'auteur de souligner l'importance de certains mots et phrases utilisés dans sa page. Certains Moteurs de Recherche tiendront compte de cette information - d'autres l'ignoreront. N'utilisez pas des guillemets autour des mots ou phrases clés.

Le méta-tag DESCRIPTION permet à l'auteur de contrôler le texte affiché quand la page paraît au niveau des résultats d'une recherche. Certains Moteurs de Recherche peuvent ignorer cette information.

Le méta-tag HTTP-EQUIV est employée pour émettre des commandes HTTP et est fréquemment employée avec le tag REFRESH pour remettre à jour le contenu de page après un nombre donné de secondes. Les pages passerelle emploient parfois cette technique pour forcer les navigateurs à aller vers une page ou un site différent. La plupart des Moteurs de Recherche en sont conscients et classeront la page à la fin et/ou réduiront le placement du site. Infoseek est contre cette technique et pénalise le site ou même l'interdit.

D'autres méta-tags sont GENERATOR (pour ceux utilisant un logiciel d'assisté :-) à la création de pages) et AUTHOR (utilisé pour créditer l'auteur de la page qui contient souvent son adresse E-mail, l'URL de son site et toute autre information utile).

Moteur de Recherche
Un serveur ou un groupe de serveurs qui se consacre au référencement des pages Internet. Lors de requêtes particulières, ces Moteurs renvoient des listes de liens correspondants à la demande. L'enregistrement dans ces moteurs se fait par les robots, la plupart du temps. Les

principaux Moteurs de Recherche sont Altavista, Excite, Hotbot, Lycos, Infoseek, Northern Light et Webcrawler. Notez que Yahoo n'est pas un Moteur de Recherche mais un Répertoire. Le terme Moteur de Recherche est bien souvent employé pour décrire les deux, Répertoire et Moteurs de recherche.

O

Open Directory Project
Un projet de répertoire mis en place par des milliers d'éditeurs volontaires. En principe, c'est un moyen très intéressant et très puissant pour organiser le web. En pratique, il y a le problème du comportement de certains éditeurs qui ont causés quelques difficultés initiales aux organisateurs. Connu à la base comme NewHoo, ce projet est maintenant intégré à Netscape (et par conséquent, AOL). Voir http://directory.mozilla.org.

Optimisation
Les changements faits à une page pour améliorer le positionnement de cette page dans un ou plusieurs Moteurs de Recherche. C'est aussi les moyens d'aider les clients ou

les visiteurs potentiels à trouver un site Web. L'optimisation peut impliquer des changements de design, de contenu, un nouveau texte pour le titre, pour les méta-tags, pour les alt des images, les entêtes, etc... Une grande image cliquable (image-map) au début de la page doit être déplacée plus bas dans la page. Les trames doivent être évitées (à moins que des liens alternatifs pour la navigation sans trames soient également fournis).

P

Page Passerelle
Une page spécifiquement pré-préparée, soumise à un Moteur de Recherche en particulier pour donner à ce Moteur les données qu'il désire, dans le format qu'il a besoin, afin de placer le site à un bon niveau

pour un ou plusieurs sujets désirés (cette importance de classement que l'on appelle le "placement" ou "positionnement").
Une page passerelle (ou portail) est adaptée pour présenter une information spécifique au Moteur que le visiteur occasionnel ne peut percevoir. La page existe afin de permettre à un site de présenter un certain visage au Moteur et un autre aux visiteurs humains.

Il y a plusieurs raisons qui peuvent expliquer la mise en place de ce genre de page. Par exemple, l'auteur ne désire pas révéler publiquement sa logique de placement. Ou bien le format pour que la page soit acceptée par certains Moteurs ne correspond pas au format désiré pour les visiteurs, d'un point de vue esthétique. Ou alors un bon format pour un Moteur n'est pas forcément le bon format pour un autre Moteur, en utilisant des pages passerelles, vous pouvez présenter vos pages de la meilleure manière connue ou supposée connue pour un Moteur en particulier.

Page Vue
En anglais, c'est page view

Le nombre de pages vues est souvent utilisé dans les logiciels de statistiques et mis en avant par rapport aux hits du serveur. Bien souvent, plusieurs hits sont nécessaires pour mesurer l'affichage d'une seule page causant plusieurs entrées dans les fichiers logs. La plupart du temps, les logiciels d'analyse peuvent déterminer les hits appartenant au même visiteur et les regrouper pour fournir des informations plus pertinentes et ainsi compter plus facilement le nombre de personnes visitant un site.

Placement
Cf Positionnement

Popularité d'une Page, des Liens
Mesure le nombre et la qualité des liens pointant vers une page particulière (des liens à

l'arrivée). Plusieurs moteurs de recherche (et le plus connu étant Infoseek) utilisent de plus en plus ce procédé dans le processus de positionnement. Le nombre et la qualité des liens à l'arrivée se révèle être de plus en plus important comme l'optimisation du contenu de la page. Un service gratuit qui mesure la popularité de votre page peut être trouvé à
http://www.linkpopularity.com
.

Portail
Cf la page passerelle mais peut aussi désigner le site portail.

Positionnement
C'est le processus de classement des sites, des pages web dans un Moteur de Recherche ou un Répertoire de façon à ce que les sites les plus appropriés apparaissent en premier sur la page résultat lors d'une requête spécifique. Des logiciels tel que AgentWebRanking Freeware (Logiciel gratuit permettant de suivre la position de son site ou d'un site concurrent dans les principaux moteurs de recherche et annuaires en fonction de mots clés. Ce graticiel permet aussi d'améliorer sa position dans les moteurs, d'inscrire votre site, de vérifier les liens et vos pages web.) PositionAgent, Rank This et Webposition peuvent être utile. Ils vous aident à définir votre position dans la page résultat d'un Moteur de Recherche en utilisant pour la recherche une phrase ou un groupe de mots particulier. Le site GoHip Search vous permet d'avoir des informations sur votre positionnement dans les principaux moteurs de recherche et l'ensemble affiché sur une même page.

R

Rang
ou Ranking

Referer
Le referer, c'est l'URL d'où vient votre visiteur. Le fichier de referer-log de votre serveur

vous indique cette information. Si un visiteur vient directement d'un résultat de Moteur de Recherche, la requête utilisée pour trouver la page sera encodée dans le referer-log, rendant plus facile la connaissance des mots-clés qui amènent des visiteurs. L'information referer peut également être consultée à partir du document.referer dans JavaScript ou par l'intermédiaire de la variable d'environnement
HTTP_REFERER.

Refresh
La balise de rafraichissement. Voir le paragraphe sur les méta tags.

Registration
Cf Enregistrement

Regroupement
Le regroupement consiste en l'affichage d'une seule page et donc adresse pour chaque site web sur la page des résultats après une requête auprès des moteurs de recherche ou des répertoires. Cette méthode permet d'éviter qu'un petit nombre de sites occupe toutes les premières positions de résultats et par la même occasion, cela rend la liste affichée par les moteurs plus claire et surtout beaucoup plus pratique pour l'utilisateur.

Répertoire
Un serveur ou un groupe de serveurs dédiés à l'indexation des pages du web. Ces répertoires retournent une liste de pages de liens selon les requêtes particulières faites par le visiteur. Les répertoires (aussi connu comme des annuaires) sont généralement mis à jour manuellement, le plus souvent suite à la demande de l'utilisateur (comme à Whatsnew.com) et la plupart mettent en place un procédé éditorial de sélection et/ou de catégorisation.

Requête
ou bien query en anglais. Un mot, une expression ou un groupe de mots employés pour passer des instructions à un Moteur de Recherche ou à un

Répertoire afin de localiser des pages sur le sujet recherché.

Ressoumission
C'est le fait de répéter une ou plusieurs fois l'enregistrement dans les Moteurs de Recherche d'une même page ou d'un même site. Dans certaines circonstances, c'est considéré avec suspicion par les Moteurs de Recherche, car ça peut indiquer qu'une personne expérimente les techniques de Spamming.

Les Moteurs de Recherche d'Infoseek et d'Altavista sont particulièrement vulnérables au Spamming parce qu'ils enregistrent les sites très rapidement. Les deux Moteurs delistent les sites qui sont ressoumis trop souvent et Infoseek, par exemple, ne permet pas plus d'une soumission de la même page en 24 heures. La ressoumission occasionnelle lorsque les pages ont changées n'est pas un problème.

Robot
Tous les programmes de navigation qui suivent les liens hypertexte des pages de Web mais qui ne sont pas directement sous contrôle humain. Les exemples sont les araignée*s des Moteurs de Recherche, les programmes ("harvester") qui extraient les adresses E-mail à partir des pages Web ou groupes de News ainsi que différents programmes de recherche intelligents. Une base de donnée des robots est maintenue par Webcrawler.

robots.txt
C'est un fichier texte déposé dans le répertoire principal de votre site pour interdire l'accès aux robots de certains pages ou sous-répertoires du site. Seuls les robots qui sont conformes à la norme d'exclusion de robots (Robots Exclusion Standard) liront et obéiront aux commandes dans le fichier. Les robots liront ce fichier à chaque visite, de sorte que des pages ou les zones des sites puissent être rendues publiques ou privées à

tout moment en changeant la teneur du fichier robots.txt. L'exemple simple ci-dessous permet d'empêcher tous les robots de visiter le répertoire /secret.

S

Scooter
C'est le petit nom de l'araignée* du Moteur de Recherche d'Altavista.

Search Engine
Cf Moteur de Recherche

Serveur
Un ordinateur, un programme ou un processus qui répond aux demandes d'informations d'un client. Sur l'Internet, toutes les pages web sont stockées sur des serveurs y compris les Moteurs et Répertoires de recherche qui sont accessibles de l'Internet.

Siphonnage
L'utilisation de divers moyens pour voler le trafic d'un autre site. Les techniques utilisées incluent la copie totale d'une page d'un site (la page copiée est modifiée légèrement pour rediriger les visiteurs vers un site différent) et l'utilisation de mots-clés ou de phrases-clé appartenant à d'autres sites Internet.

Slurp
C'est le petit nom de l'araignée* utilisée par Inktomi.

Spamdexing
La création ou modification d'un document avec l'intention de tromper un catalogue ou un système de classement électronique. Toute technique qui a pour objectif d'augmenter la position potentielle d'un site aux dépens de la qualité de la base de données du Moteur de Recherche peut également être considérée comme du spamdexing (également connu sous le nom de spamming ou spoofing).
Le spamming (autrement appelé spam, spamdexing, keyword stuffing et autres) consiste à employer des techniques de fraude dans le

code HTML d'une page Web afin qu'elle soit classée dans les 10 meilleurs résultats des Moteurs de recherche en texte integral pour certains mots clés donnés.

En règle générale, les Moteurs savent sans problème détecter du spamming comme la répétition de mots clés des centaines de fois, en caractères petits, en blanc sur fond blanc ou en clair sur fond blanc, la répétition dans les balises IMG (attribut ALT), dans les faux liens ou les fausses images, les balises <TITLE> repétées, etc... bref toute la panoplie de la fraude basée sur la répétition des mots clés consistant donc à faire augmenter l'indice de densité* d'un mot donné.

D'autres techniques de spam sont également utilisées, notamment l'emploi de la balise Méta Refresh (TRÈS mal vue par les moteurs) ou la technique consistant à repérer les adresses ou les numéros IP des robots des Moteurs pour leur fournir une page différente de celle fournie logiquement aux internautes (technique dite de la substitution de pages). Je passe sur les détails.

Selon les Moteurs, les mesures suivantes peuvent être prises :

Baisse automatique de la note de la page sur un mot clé donné (la page est releguée en fond de classement)
Si des pages identiques et spammées sont trouvées

(technique des pages satellites), la plupart sont enlevées de l'index et une seule occurrence est gardée mais sera mal classée si elle contient du spam
La page est virée de l'index
Le site complet est viré de l'index. Cela pose des problèmes, surtout pour les sites d'hébergement gratuit (imaginez que mygale.org, chez.com ou geocities.com soient virés d'AltaVista ou de HotBot !)

Spider*, Spyder
Cf Araignée*

Stop Word
Un mot qui est ignoré lors d'une requête auprès d'un moteur de recherche. Le mot est trop souvent utilisé que son utilisation n'améliore en rien la pertinance des résultats. Comme exemples, en anglais, les mots liés au net comme computer, web et des mots plus généralistes comme get, I, me, the, you, etc

Submission Service
Différents agents qui soumettent votre site a plusieurs moteurs de recherche et annuaires. Pratique pour s'enregistrer dans plusieurs moteurs de recherches d'importance moindre, mais il ne faut pas faire confiance à ce genre de programmes pour être repertorié dans les principaux moteurs. Plusieurs de ces services sont automatiques et fonctionnels à partir de sites web et d'autres fonctionnent hors-ligne. Certains de ces services sont gratuits ! Attention à ne faites pas trop confiance aux supposés FFA (Free For All - Gratuit Pour Tous) en donnnt votre adresse électronique, vous risquez de vous retrouver victime de spams !

Substitution de Pages
C'est un processus permettant d'envoyer les robots des Moteurs de Recherche vers une page spécifique, pour diriger ensuite vos visiteurs vers ce que vous voulez qu'ils voient. C'est fait en utilisant les server side includes (ou d'autres méthodes de contenu dynamique). Le SSI, par exemple, permet de délivrer un contenu différent au client selon la valeur de son HTTP_USER_AGENT. La plupart des navigateurs normaux ont une chaîne de caractères agent qui commence par "Mozilla" (créé à partir des termes Mosaic et Godzilla). La plupart des robots des Moteurs de Recherche ont un nom d'agent spécifique comme "Gulliver", "Infoseek sidewinder", "Lycos spider*" et "Scooter".

En changeant la valeur de votre HTTP_USER_AGENT (par un processus appelé agent detection), différentes pages peuvent être affichées à la même URL, de sorte que vos visiteurs ne voient jamais la page soumise aux Moteurs de Recherche (et vice versa).
C'est théoriquement quelquechose d'assez simple. Certains Moteurs de Recherche prétendent être de "vrais navigateurs mozilla" afin, justement, d'éviter ces substitutions. L'utilisation des substitutions de pages peut se réveler difficile et même ne pas fonctionner.
Comment découvrir une substitution au travail ? C'est assez difficile puisque les possesseurs de substitutions arrivent à contrôler ce que vous voyez ! Vous pouvez être capable de deviner qu'une page utilise cette technique si elle est, par exemple, indexée incorrectement ou le titre ou la description ne correspondent pas à ce que vous voyez; mais la raison peut être beaucoup plus simple, la page a changé depuis que le Moteur l'a indexé. La meilleure solution consiste à écrire un petit programme (avec PERL, par exemple) et de récuperer plusieurs fois la page index en mettant un HTTP_USER_AGENT différent à chaque fois. Si une substitution de pages est utilisé, une ou plusieurs pages devraient se révéler différentes des autres !

T

Technique de Positionnement
C'est le fait de modifier sa page web de sorte que les Moteurs de Recherche (ou un Moteur de Recherche en particulier) traitent la page comme la plus appropriée à une requête spécifique (ou à un ensemble de requêtes).

Texte Caché
C'est du texte sur une page qui est visible pour les robots des Moteurs de Recherche mais invisible aux visiteurs. C'est généralement parce que le texte a la même couleur que le fond, ou bien parce que des

tags multiples de <TITRE> ont été utilisés ou alors parce que le texte est un commentaire en HTML. Le texte caché est souvent utilisé pour le spamdexing. Beaucoup de Moteurs de Recherche peuvent maintenant détecter l'utilisation du texte caché et souvent retirent les pages de leur base de données ou abaissent la position des pages.
Le texte peut être caché en utilisant la substitution de pages (agent name delivery en anglais) ou l'IP delivery, ce qui permet de presenterun texte différent aux différents robots des moteurs de recherches ou de cacher les vrais sources HTML aux concurrents. Le "Stealth META Tag CGI Script" utilise probablement cette technique et est disponible à http://www.OutRank.com/stealth.shtml. Un autre logiciel permettant de cacher le source HTML s'appelle Psyral Phobia et est disponible à http://www.merlesworld.com/software.htm.

Titre
C'est le texte compris entre les deux balises HTML <title> et </title>. Ce texte est associé (et non pas affiché) avec la page web qui contient ces balises et affiché dans une place spéciale de votre navigateur (souvent le haut de la fenêtre).
Le texte du titre est important car c'est bien souvent le lien avec votre site qui s'affiche au niveau des listes de résultats des Moteurs de recherche. De même, lors de l'indexation de votre site, les Moteurs font particulièrement attention au titre.

Toile
Cf Web

Trames
C'est une technique de programmation en HTML pour combiner deux documents ou plus séparés dans une fenêtre de navigateur web. Les documents agissants l'un sur l'autre à travers différentes sous-fenêtres, le site peut être

créé dans l'objectif d'être plus interactif, plus pertinent.

Un site tramé pose souvent de grands problèmes pour les Moteurs de Recherche et ne peut pas être classé correctement. Les Moteurs de Recherche classeront souvent seulement la partie du site encadrée dans la section <NOFRAMES>. Ainsi assurez-vous que la section <NOFRAMES> contient bien le texte approprié pour être repertorié par les robots. Si votre site utilise des trames, pensez à fournir une page passerelle ou ajoutez des liens dans les pages tramées. Soumettez la page principale - celle contenant le <FRAMESET> - aux Moteurs de Recherche et si vous utilisez une page passerelle, soumettez celle-ci séparément.

Tromperie sur Fontes & Fond
Il existait diverses techniques pour placer du texte invisible dans une page Web afin d'améliorer le positionnement sans en affecter l'aspect. La plupart du temps, elles sont basées sur une couleur similaire pour les fontes et celle du fond (par exemple blanc). Aujourdhui, la plupart des Moteurs de Recherche détecte ces tours.

U

URL
Universal Resource Locator - Une adresse qui peut indiquer n'importe quelle ressource spécifique à Internet. Le début de l'adresse indique le type de ressource - par exemple http: pour des pages Web, ftp: pour des transferts de fichier, mailto: pour des adresses E-mail, etc

V

Visiteur Unique
Un "vrai" visiteur sur votre site. Les serveurs web enregistrent les adresses IP de chaque visiteur. Ces enregistrements peuvent être utilisés pour mesurer le taux de vrais visiteurs sur votre site web.

Si par exemple une personne visite 20 pages de votre site, le serveur ne comptabilisera qu'un unique visiteur (car toutes les pages accédées sont associées à la même adresse IP) mais vingt accès de page.

W

Web
ou Toile. Le web designe plus largement l'ensemble du réseau de sites où vous pouvez naviguer avec votre navigateur.

TABLE DES LIENS

Dans l'ordre d'apparition :

http://solutions.journaldunet.com
http://www.hummingbird.com/
http://www.verity.com/
http://www.arisem.com/
http://www.sinequa.com/
http://www.convera.com/
http://www.inktomi.com/
http://www.atomz.com/
http://www.triplehop.com/
http://www.autonomy.com/
http://www.instranet.com/
http://www.lexiquest.com/
http://www.albert.com/
http://www.alogic.fr/
http://www.auracom.fr/
http://solutions.journaldunet.com/0108/www.mediapps.com
http://www.microsoft.com/
http://www.reef.com/
http://askjeeves.com
http://www.infoclic.fr/
http://maximini.net/
http://dublincore.org
http://catalogs.google.com
http://www.ltutech.com
http://www.ipea.com
http://www.google.com
http://www.journaldunet.com/dossiers/elections/020205tags.shtml
http://www.panaf.org
http://www.mediametrie.com
http://www.idf.net
http://fr.yahoo.com/
http://fr.dir.yahoo.com/Sante/
http://www.hotbot.lycos.fr
http://search.ke.voila.fr

TABLE DES FIGURES et TABLEAUX

Figure a : Le site tel que vous le voyez 7
Figure b : le code source 8
Figure 4 : L'utilisateur des annuaires peut formuler sa requête sous forme de mots 20
Figure 5 : Ou en se promenant dans les rubriques de la Home Page 21
Figure 6 : IXQuick, métamoteur 22
Figure 7 : Exemple d'un moteur en langage naturel 25
Figure 8 : Détail d'ASKJEEVES 25
Figure 9 : exemple d'une page toute simple en HTML 26
Figure 10: Page d'un site en cache de Google 35
Figure 11: page telle qu'elle est actuellement 35

Tableau 1: Progression du nombre d'internautes 36
Tableau 2: Equipement des foyers 36

BIBLIOGRAPHIE ET BOOKMARKS

lecture et re-lecture

- Documents papiers

  - ✓ *Promouvoir son site web, référencement, campagne de pub, échange de liens, communiqué de presse*, FRAYSSE Emmanuel, édition Microsoft press 2001.
    ISBN :2-84082-888-X

- Sites Internet

Référencement

- ✓ *http://www.indicateur.com*
- ✓ http://www.enfin.com
- ✓ *http://www.abondance.com*
- ✓ *http://www.estat.fr*
- ✓ http://actu.ladoc.net
- ✓ http://www.searchengineworld.com

Langue

- ✓ Dictionnaire des synonymes : *http://elsap1.unicaen.fr/dicosyn.html*
- ✓ Dictionnaire : *http://www.francophonie.hachette-livre.fr/*
- ✓ Annuaires des dictionnaires : *http://www.dicorama.com/*

Autres

- ✓ *http://dublincore.org*

Outils

- ✓ *http://www.ranks.nl*
- ✓ *http://agentwebranking.com/*

- News letters
  - ✓ *Actu.moteur.com*
  - ✓ *Abondance*
  - ✓ *Search days*
- Travaux divers
  - ✓ *http://membres.lycos.fr/ddiaz/*