

# Questions/réponses à propos du moteur de recherche mysearch.epfl.ch

Francis.Lapique@epfl.ch, e-pfl

Un moteur de recherche Inktomi est en exploitation depuis août 2001 sur le site Web de l'EPFL (voir l'article [Choix d'un moteur de recherche pour le site EPFL, Flash informatique 7/01](#)). Le but de cet article est de donner un certain nombre de recommandations et d'informations pour, d'une part comprendre la réaction de notre moteur Inktomi et, d'autre part optimiser la pertinence des réponses. Nous allons aborder ce sujet sous la forme de questions-réponses, les plus concrètes et les plus complètes possible.

## Comment inclure ou exclure certaines pages de l'indexation?

Vous avez trois solutions:

la plus connue: le fichier *robots.txt* qui permet d'exclure des pages ou des répertoires

**Attention:** le fichier robots.txt prend un **s**, s'écrit en minuscule et est toujours placé à la racine du site: [www.monsite.com/robots.txt](http://www.monsite.com/robots.txt). Je le signale car je ne compte plus les appels s'indignant qu'Inktomi ne suit pas les consignes de robots.txt. Inktomi tient compte de robots.txt si vous le mettez au bon endroit et si vous en respectez la syntaxe. Syntaxe simple car c'est un fichier texte, constitué d'une liste d'instructions destinées aux moteurs de recherche. Il n'existe que deux commandes qui soient reconnues par tous les moteurs:

- la commande **User-agent** qui permet de s'adresser à un moteur en particulier, ou bien à tous en utilisant le caractère \* ;
- la commande **Disallow** qui permet d'interdire à un moteur ou à tous, un fichier précis ou un répertoire désignés par leurs urls relatives.

En l'absence d'interdiction, tout fichier présent sur un site Web est considéré par défaut comme indexable. Le fichier robots.txt peut contenir des commentaires, s'ils sont précédés du caractère "#".

Exemples de fichiers robots.txt:

```
User-agent: *
Disallow: /fichier.html
interdit à tous les moteurs la page fichier.html située à la racine
```

```
User-agent: *
Disallow: /dossier/
interdit à tous les moteurs le répertoire /dossier
```

```
# go away
User-agent: *
Disallow: /
interdit tout le site à tous les moteurs
```

```
User-agent: Inktomi Search
Disallow:
User-agent: *
Disallow: /
interdit tout le site à tous les moteurs, sauf Inktomi.
```

moins connue car non supportée par tous les robots, l'utilisation du meta tag **robots**

Ce Meta Tag doit être présent dans toutes les pages du site que l'on désire traiter à part. Pour chacune des pages, il est nécessaire d'ajouter le tag **robots** tel que: `<meta name="robots" content="index, follow">`. Dans ce cas, le robot indexera toute la page et tous les liens s'y trouvant.

Les valeurs les plus courantes pour ce tag sont:

- **index** —> pour que le robot indexe la page
- **noindex** —> pour que le robot n'indexe pas la page
- **follow** —> pour que le robot suive les liens de la page
- **nofollow** —> pour que le robot ne suive pas les liens de la page

Ainsi, avec `<meta name="robots" content="index, nofollow">`, le robot indexe la page en cours, mais ne suit pas les liens.

Exemple (présenté à la conférence des webmasters-EPFL du 17/01/02):

Considérez le code source de la page [monserver.epfl.ch/pages/index.html](http://monserver.epfl.ch/pages/index.html)

```
<html>
<head>
<meta name="robots" content="noindex, follow">
</head>
<h1>Un titre</h1>
<a href=/index.html>Page d'accueil</a>
<a href=Oudjat.html><img src=/egyptel.gif
alt="Oudjat" border=0></a>
</html>
```

Cette page n'est pas indexée, par contre la page Oudjat.html l'est. La requête suivante

<http://mysearch/query.html?col=test&qt=accueil>

ne donne rien, par contre la requête

<http://mysearch/query.html?col=test&qt=défunt>

donne le résultat

<http://msg3.epfl.ch/pages/Oudjat.html>

*Oudjat L'œil est un symbole fondamental dans l'histoire égyptienne, il est représenté sous de multiples formes et dans toutes les situations: L'œil Oudjat est le symbole de l'intégrité du corps et de la préservation de la vie. En effet, dans le ...*

envoyer un email à [webmaster@epfl.ch](mailto:webmaster@epfl.ch) pour des demandes très spécifiques

Inktomi est géré au travers d'une interface d'administration Web. Nous pouvons donner des règles d'exclusion comme par exemple celle-ci: disallow [http://monserveur.epfl.ch/\\*.asp?](http://monserveur.epfl.ch/*.asp?) pour ne pas indexer les pages asp d'un serveur.

## Inktomi n'a pas pris en compte la modification de ma page

A ce jour, Inktomi visite à l'EPFL environ 370 sites et indexe environ 380 000 documents. Ce qui représente à titre indicatif 85% de notre licence. L'ensemble de ce travail prend entre un et deux jours. Nous avons décidé de manière arbitraire de procéder à cette indexation une fois par semaine tous les vendredis. Un rythme plus soutenu peut amener pas mal de *pollution* au niveau des logs des serveurs, une fréquence plus faible risque de provoquer un temps trop important entre deux mises à jour. Donc si vous faites vos modifications un lundi, vos modifications ne seront prises en compte que le lundi suivant.

Là encore, pour des demandes très spécifiques nous pouvons forcer le robot à revisiter tel ou tel site.

## Comment la réponse à une requête est-elle construite ?

Reprenons notre page [monserveur/pages/index.html](http://monserveur/pages/index.html). Par rapport à l'exemple précédent nous avons ajouté 2 autres tags qui sont **description** et **author**:

```
<html>
<head>
<meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
<title>e-pfl: affichage du contenu du tag &lt;title&gt;</title>
<meta name="description" content="META tag &quot;description&quot; C'est une page de test pour la conférence des webmasters">
<meta name="author" content="META tag &quot;author&quot; F.Lapique">
<meta name="robots" content="index, follow">
</head>
<h1>Un titre</h1>Hello!!!
<a href=/index.html>Page d'accueil</a>
<a href=Oudjat.html><img src=/egyptel.gif alt="Oudjat" border=0></a>
</html>
```

Le résultat de la requête

<http://mysearch/query.html?col=test&qt=hello> est le suivant:

Résultats pour: hello

e-pfl: affichage du contenu du tag <title>

META tag "description" C'est une page de test pour la conférence des webmasters

META tag "author" F.Lapique <http://emsg3.epfl.ch/pages/index.html> - 0.6KB

Il a été construit de la façon suivante: le texte de la balise <title> lié à l'url, un descriptif de la page qui n'est rien d'autre que le contenu du Meta tag **description** et l'affichage du contenu du meta tag **author**.

Si nous supprimons la balise <title> et le meta tag **description** nous avons un autre résultat:

<http://emsg3.epfl.ch/pages/index.html>

Un titre Hello!! Page d'accueil

META tag "author" F.Lapique <http://emsg3.epfl.ch/pages/index.html> - 0.5KB

L'information sur l'url est l'url elle-même, quant au descriptif il est construit à partir des premiers éléments trouvés après la balise </head>. Donc si vous voulez éviter un résumé obscur utilisez le meta tag **description**.

## Comment améliorer la pertinence des recherches ?

Vaste problème: Inktomi s'en sort pas si mal, entrez par exemple le mot **electricite** dans [mysearch.epfl.ch](http://mysearch.epfl.ch) et vous verrez qu'il place en premier la home page de l'ancien département d'électricité. Bien, rien d'étonnant et c'est normal diriez-vous. Mais à titre de curiosité, regardez le code source de la page. Je passerai sous silence les exemples où les choses se passent moins bien. Pour éviter les aléas il faut aider Inktomi mais pas de n'importe quelle manière.

Au moment de l'indexation Inktomi va affecter aux mots des poids différents suivant qu'ils se trouvent par exemple dans un tag <a>, <title> ou non. Ces poids sont numérotés de 0 (plus faible) à 10 (plus fort). La pondération en exploitation aujourd'hui est la suivante:

Title 8  
Description 4  
Keywords 4  
Alt 1  
Remote anchors 4

Il est inutile de répéter un mot 10 fois. Cela serait contre productif car il y a détecteur de spam.

Faisons la requête suivante:

<http://mysearch/query.html?col=test&qt=oeil>

Comme précédemment, reprenons notre page [monserveur.epfl.ch/pages/index.html](http://monserveur.epfl.ch/pages/index.html), à laquelle on a ajouté le tag **keywords** et notre page Oudjat.html où on compte plus de 10 fois le mot **œil**:

```
<html>
<head>
<meta http-equiv="content-type" content="text/html; charset=iso-8859-1">
<title>e-pfl: affichage du contenu du tag
```

```
<title></title>
<meta name="description"
content="META tag &quot;description&quot; C'est
une page de test pour la conférence des
webmasters">
<meta name="keywords" content="egypte, oeil ,
e-pfl">
<meta name="author" content="META tag
&quot;author&quot; F.Lapique"><meta
name="robots" content="index,follow">
</head>
<h1>Un titre</h1>Hello!!!
<a href=/index.html>Page d'accueil</a>
<a href=Oudjat.html><img src=/egyptel.gif
alt="Oudjat" border=0</a>
</html>
```

Nous avons le résultat suivant avec un score de 56% pour le premier et de 45% pour le second

**Recommandations:** tag **<title>** obligatoire, attention aux termes qui apparaissent dans **description** du fait de leur poids important.

#### Résultats pour: oeil

2 résultats trouvés, triés par thème    trier par date    dont la synthèse est cachée    1-2

<a href="http://emsg3.epfl.ch/pages/Oudjat.html">http://emsg3.epfl.ch/pages/Oudjat.html</a> Oudjat L'œil est un symbole fondamental dans l'histoire égyptienne, il est représenté sous de multiples formes et dans toutes les situations : L'œil Oudjat est le symbole de l'intégrité du corps et de la préservation de la vie. En effet, dans le ... <small>http://emsg3.epfl.ch/pages/Oudjat.html - 1.0KB</small>	56% 10 Jan 02 <a href="#">Trouver un terme similaire</a>
<b>e-pfl: affichage du contenu du tag &lt;title&gt;</b> META tag "description" C'est une page de test pour la conférence des webmasters META tag "author" F.Lapique <small>http://emsg3.epfl.ch/pages/index.html - 0.0KB</small>	45% 10 Jan 02 <a href="#">Trouver un terme similaire</a>

Faites cette requête sur le Web entier.    trier par date    dont la synthèse est cachée    1-2

## Peut-on indexer des informations contenues dans des bases de données ?

Comme je l'ai indiqué dans l'article cité plus haut sur le choix d'un moteur de recherche, la réponse est oui pour Oracle et MySQL sur lequel nous avons fait des essais.

## Quand je rentre l'url mysearch.epfl.ch que représente le bouton "epfl" ?

Inktomi est organisé autour de la notion de **collection**. A chaque collection sont associés un serveur **root**, des règles et des propriétés. Les exemples ci-dessus ont été faits sur une collection **test** ayant comme serveur root **monserver.epfl.ch**.

Actuellement, le moteur travaille sur une seule collection **epfl**. On peut regrouper un ensemble de collections dans une seule. Nous allons peut-être dans le futur cons-

truire la collection **epfl** à partir des collections **enac**, **fsb**, etc. Le problème que nous devons résoudre dans ce cas est l'affectation pertinente des 370 serveurs à chacune de ces entités. ■