

Cours  
Ingénierie des langues  
A. Pappa



# Crawlers ou Robot d'indexation

- Définition :

Le terme de crawler, ou spider, désigne un robot d'indexation. Il s'agit d'un logiciel qui a pour principale mission d'explorer le Web afin d'analyser les contenus ainsi explorés.

L'index informatique sert à répertorier des adresses URL et des contenus de sites Web. L'un des index les plus connus est sans doute celui utilisé par Google pour le référencement des sites Internet et l'affichage des résultats dans son moteur de recherche.

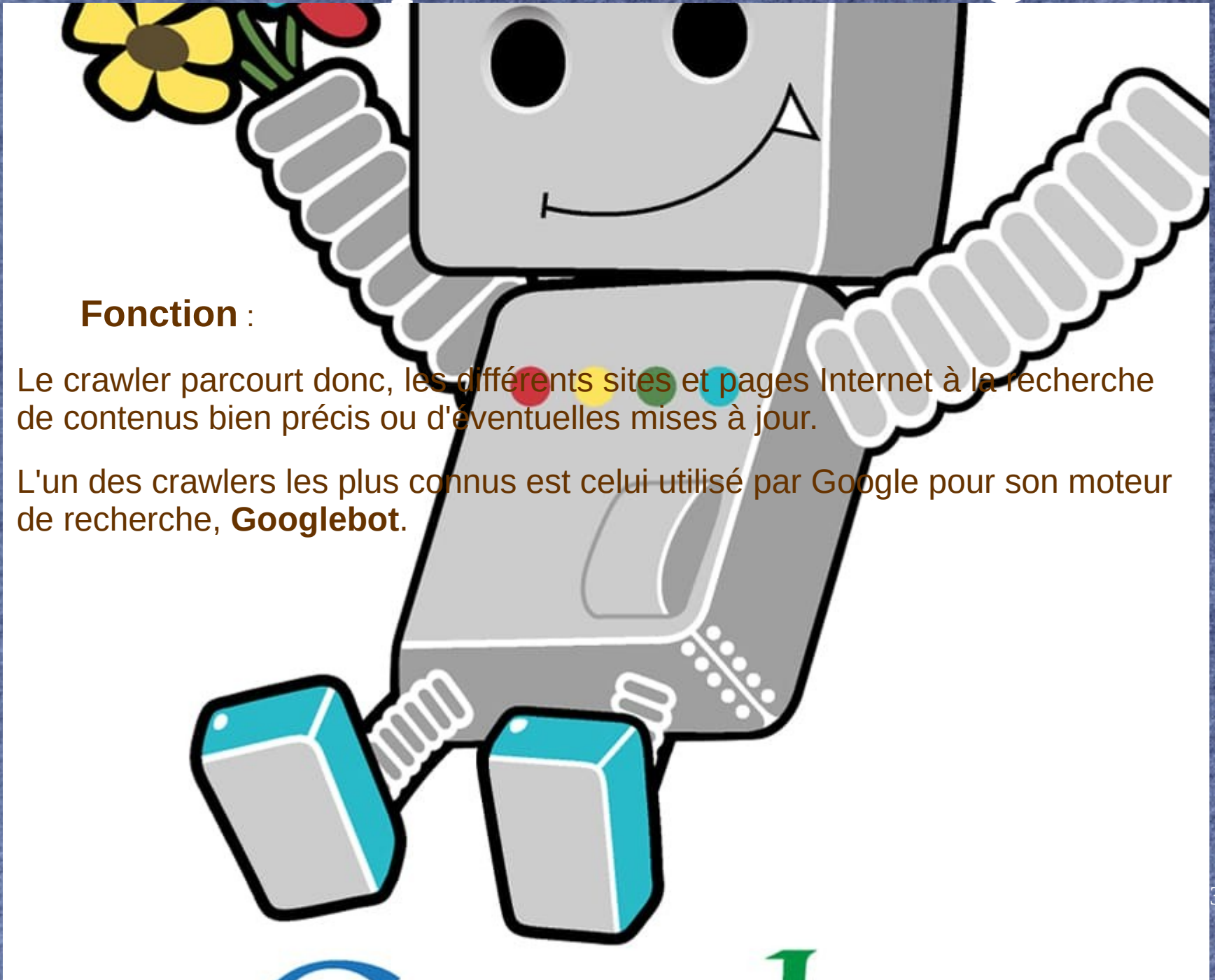


# Le crawler le plus connu : Googlebot

## Fonction :

Le crawler parcourt donc, les différents sites et pages Internet à la recherche de contenus bien précis ou d'éventuelles mises à jour.

L'un des crawlers les plus connus est celui utilisé par Google pour son moteur de recherche, **Googlebot**.



# Googlebot

**Googlebot** est un robot d'indexation conçu par Google et utilisé pour son moteur de recherche afin de détecter de nouvelles pages web et les mises à jour de pages Web déjà existantes.

## **Googlebot :**

- logiciel capable d'analyser, de façon automatique, le Web et les ressources d'un site Internet en vue d'indexer les pages Web.
- logiciel capable de détecter les sites en suivant les différents liens existant entre les pages.



# Fonctionnement GoogleBot

Le fonctionnement de GoogleBot est basé sur un nombre d'algorithmes et sur des programmes informatiques qui déterminent les pages Web à explorer, à quelle fréquence, et le contenu à extraire de chaque site.

L'algorithme de Google est une suite d'instructions, qui servent à déterminer la position des résultats suivant une recherche.

Au moins **200** critères sont aujourd'hui pris en compte par l'algorithme de Google.



# Lexique

Ancre (Anchor) :

L'ancre de lien est le texte contenu dans un lien hypertexte.

Backlink :

Le **backlink** est un lien hypertexte qui permet d'envoyer un internaute d'un site web X vers un site web Y.

Cross linking :

Le **crosslinking** désigne la mise en place de liens réciproques entre deux sites Web. Dit autrement, il s'agit pour un site A de proposer un lien hypertexte vers un site B, et pour ce site B, de proposer un lien hypertexte vers ce site A.

Google Trends (ex Google Insight) :

**Google Trends** est un outil mis en place par Google pour identifier le nombre de fois où un terme a fait l'objet d'une requête dans son moteur de recherche. Dit autrement, **Google Trends** est un outil qui permet d'analyser la popularité d'un terme sur le moteur de recherche, dans une période de temps déterminée.

PageRank :

Le **PageRank** est un algorithme utilisé par le moteur de recherche américain pour attribuer une note à une page web. Il évalue la popularité de cette page web afin de l'indexer de façon optimale avant de la présenter aux internautes ayant opéré une requête sur Google.



# Lexique suite...

Robots.txt :

le fichier **robots.txt**, est un fichier contenant des commandes destinées aux robots d'indexation des moteurs de recherche. Il participe en ce sens au référencement naturel d'un site web.

Le fichier **robots.txt** apporte des instructions aux robots des moteurs de recherche qui explorent le Web. Il leur indique par exemple de ne pas indexer telle ou telle rubrique d'un site web (inutile que Google indexe une interface d'administration par exemple) et peut même aller jusqu'à empêcher l'indexation d'un site web par les moteurs de recherche.

Sitemap :

Un **sitemap**, est un fichier XML (dans la plupart des cas) qui présente l'architecture générale d'un site web. Il permet de hiérarchiser les ressources et les contenus proposés par le site.

Le **sitemap** a pour principale vocation d'aider les robots d'indexation des moteurs de recherche à retrouver l'ensemble des pages (URL) à indexer. C'est d'ailleurs le géant du Web Google qui est à l'origine du développement du fichier **sitemap**, dès 2005.

Snippet :

Un **snippet**, désigne une partie d'un code source ou une partie d'un texte pouvant être réutilisée. Dans le monde du référencement, le **snippet** définit l'espace dans lequel apparaît la brève description utilisée pour la présentation d'un site web dans les pages de résultats des moteurs de recherche. Soigner le contenu du **snippet** est indispensable si l'on souhaite inciter les internautes à cliquer sur le résultat proposé par le moteur de recherche.



# Lexique suite ...

Netlinking :

Le **netlinking** est une technique de référencement d'un site internet qui consiste à vouloir multiplier le nombre de liens hypertextes, ou "backlinks" pointant vers lui. Objectifs de cette démarche : améliorer la qualité du trafic et la popularité du site web dans le but d'obtenir un meilleur référencement naturel.

SEA (Search Engine Advertising) :

Le **SEA**, pour Search Engine Advertising, est une forme de référencement qui permet à des sites web d'améliorer leur visibilité dans les pages de résultats proposées par les moteurs de recherche comme Google. À la différence du SEO (Search Engine Optimization), aussi appelé le référencement naturel, il s'agit ici d'une technique de référencement payante. Autrement dit, on ne s'embarrasse pas avec les techniques SEO qui doivent séduire Google sans le rémunérer, mais on accepte de payer la bonne visibilité de son site web dans les moteurs de recherche.

SERP (Search Engine Result Page) :

L'acronyme **SERP**, pour Search Engine Result Page, désigne la page web générée par un moteur de recherche en fonction des mots clés renseignés par un internaute. La SERP propose alors à ce dernier une sélection de liens qu'elle juge pertinents par rapport à sa requête. Il s'agit de la page web qui s'affiche lorsqu'un internaute effectue une recherche sur un moteur de recherche comme Google.

Selon les moteurs de recherche, la **SERP** affiche plusieurs informations comme l'adresse URL, une description brève du contenu, un aperçu, la date de la dernière indexation de la ressource, etc.



# Lexique encore !

SMO (Social Media Optimization) :

Le **SMO**, pour Social Media Optimization, désigne tout un ensemble de méthodes utilisées afin d'augmenter le nombre de visiteurs sur une page web. Le principe du **SMO** consiste à aller recruter ces nouveaux visiteurs sur des médias ou réseaux sociaux. Il peut être présenté comme le bouche-à-oreille du Web dans la mesure où il s'appuie sur la diffusion d'un contenu à travers des réseaux virtuels pour accroître la réputation d'un site web.

Dans la pratique, le **SMO** profite de la popularité des réseaux sociaux (Facebook) et autres plateformes similaires (YouTube, etc.) pour propager un contenu. Les réseaux ou médias sociaux sont alors utilisés pour relayer une information.

Le **SMO** est un dispositif complémentaire au SEO pour améliorer la popularité d'un site web et son positionnement dans les pages de résultats des moteurs de recherche.

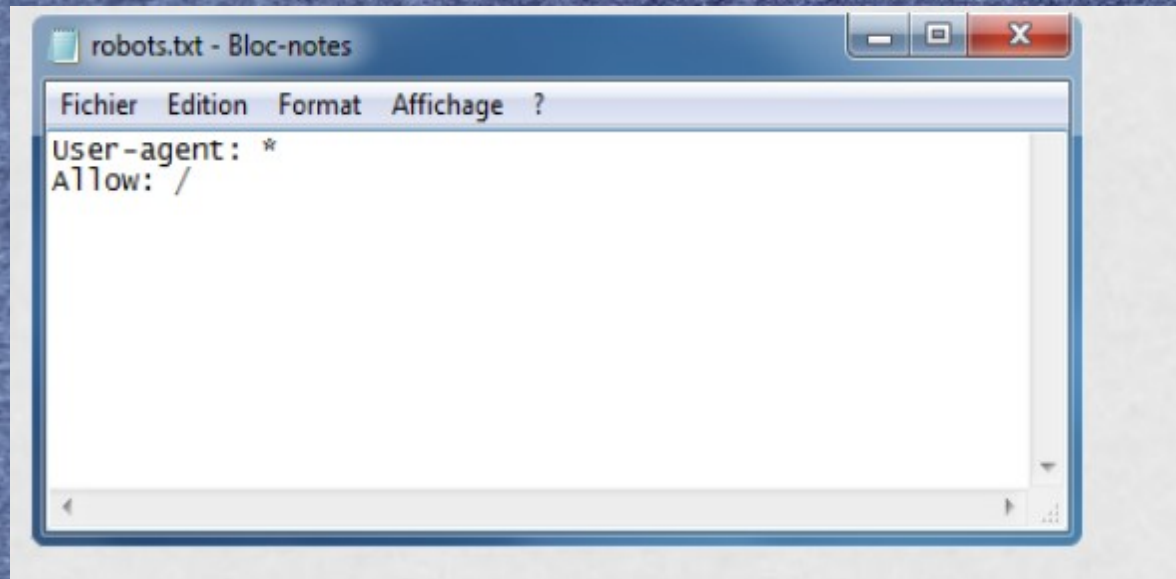
URL rewriting :

L'URL **rewriting** désigne une technique qui consiste à réécrire une URL (ou adresse Web). Le but ? Éviter les adresses web qui ne ressemblent à pas grand-chose et qui sont peu lisibles par les internautes (comme pour les moteurs de recherche).

L'objectif consiste alors à mettre en place des URL simplifiées et mieux optimisées pour le référencement naturel et pour les internautes. Ainsi, une adresse web contenant des caractères comme =, id, ?, &, etc. pourra être épurée pour une meilleure lisibilité.



# Le file “robots.txt”



Les mots-clés pour ce file :

**User-agent**

**Disallow**

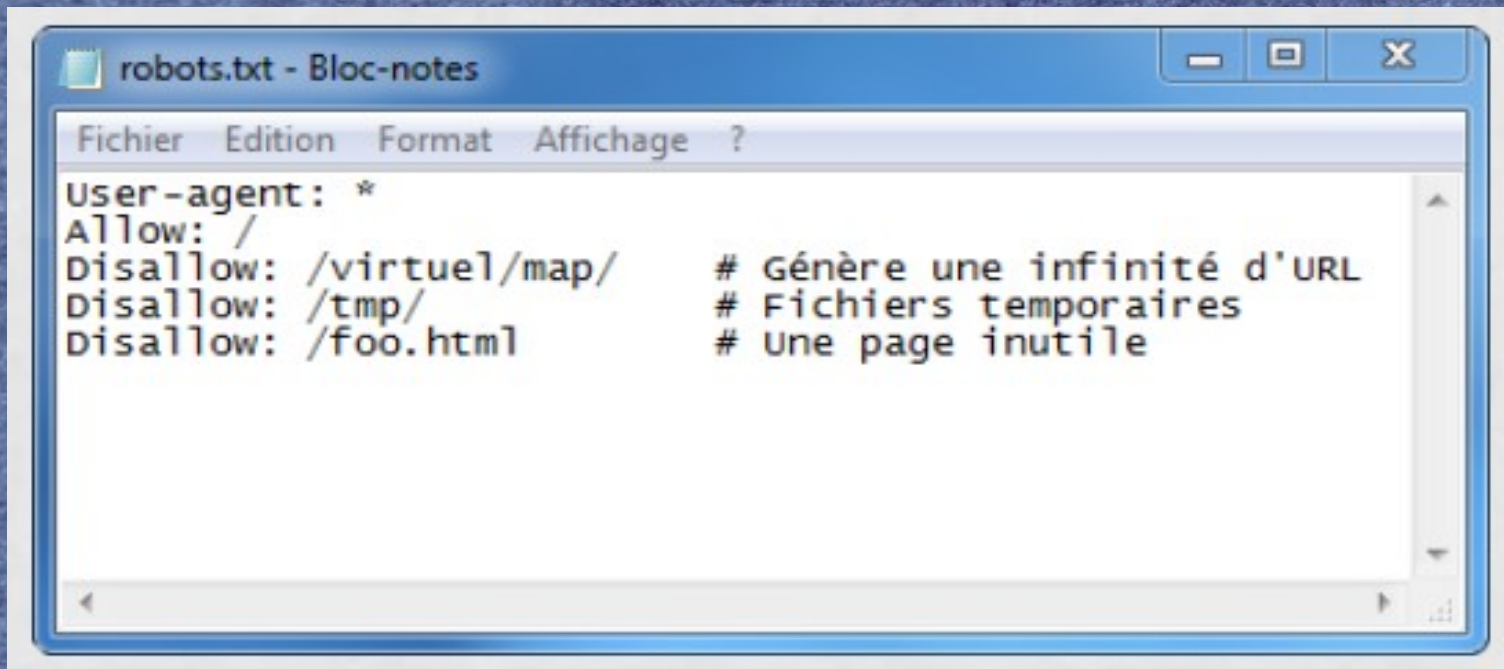
**Allow**

**Sitemap**

**#**



# Comment créer le fichier “robots.txt”



```
robots.txt - Bloc-notes
Fichier  Edition  Format  Affichage  ?
User-agent: *
Allow: /
Disallow: /virtuel/map/      # Génère une infinité d'URL
Disallow: /tmp/              # Fichiers temporaires
Disallow: /foo.html          # Une page inutile
```



# Liste des robots les plus populaires

Nom du moteur	User-Agent
Alta Vista	Scooter
Excite	ArchitextSpider
Google	Googlebot
HotBot	Slurp
InfoSeek	InfoSeek Sidewinder
Lycos	T-Rex
Voilà	Echo