



---

## Création de moteurs de recherche spécialisés

---

En parallèle des moteurs de recherche généralistes (tels que Google ou Yahoo!) qui permettent d'interroger un index de pages web immense mais dont le bruit important empêche souvent d'obtenir des résultats véritablement satisfaisants, il existe, des moteurs de recherche spécialisés : par exemple, en fonction d'un domaine particulier (Ejustice pour le domaine juridique : <http://www.ejustice.fr>), d'une zone géographique (Cooliroo, le moteur de recherche spécialisé Antilles Caraïbes : <http://www.cooliroo.com>) ou bien du type de document recherché (pour les PDF : <http://www.pdf-search-engine.com>).

La solution la plus poussée consiste à créer son propre moteur de recherche. Celui-ci permet de s'adapter au cas par cas aux besoins de l'utilisateur qui dispose alors de la possibilité d'ajuster un contenu à une recherche afin de la rendre beaucoup plus efficace.

Plusieurs outils proposent ces fonctionnalités et notamment :

- **Rollyo** [<http://rollyo.com>]
- **Swicki** [<http://www.eurekster.com/swickibuilder/customize.aspx?mode=wizard>]
- **Google Custom Search Engine** [<http://www.google.fr/coop/cse>]

**Rollyo** est un moteur qui permet de définir des listes, baptisées searchrolls, pouvant aller jusqu'à vingt-cinq sites. Il suffit alors de spécifier à partir de laquelle la recherche sera effectuée et la requête envoyée ne portera que sur les sources désignées. Il s'agit d'un outil basique, intuitif et très simple d'utilisation, idéal pour cibler une recherche en peu de temps. Il ne faut cependant pas oublier que les listes de sites sont publiques, qu'elles ont la possibilité d'être dotées de mots-clé et que, de ce fait, elles peuvent être consultées et utilisées par les autres internautes.

**Swicki** reprend le même type de fonctionnalités mais de façon plus étendue et plus élaborée. Tout d'abord, il permet de lister jusqu'à cinquante pages internet tout en offrant la possibilité d'en exclure d'autres. A chaque moteur créé sont associées une description (rédigée par l'auteur), une indexation dans des catégories prédéfinies ainsi qu'une liste de mots-clé (présentée sous forme de nuage) qui permet, au fil du temps, d'associer à l'index initial d'autres sites traitant du même sujet et considérés comme pertinents. Son interface simple et intuitive peut être, au choix, intégrée sur un site existant (pour faire office de moteur interne ou bien permettre d'effectuer des recherches sur un index défini) ou bien être consultée via une URL spécifique (du type <http://ma-liste-swicki.eurekster.com>). Une option de mise en place d'un flux RSS est également prévue. Cependant, la principale innovation de Swicki est son mode de fonctionnement prévu pour être collaboratif sur le principe du partage et de la participation active. En effet, le moteur évoluera selon les recherches effectuées par la communauté et



fournira ainsi des réponses de plus en plus précises au fil du temps, à condition d'être alimenté et mis à jour très régulièrement.

Contrairement à Rollyo ou Swicki, **Google Custom Search Engine** se base sur un nombre illimité de pages et de flux RSS proposés par l'auteur du moteur. Il offre, en outre, la possibilité de charger cette liste aux formats de structuration et de transfert de données OPML, XML ou TSV. Des possibilités de mise à jour de l'index et d'exclusion de sites sont également prévues. Le moteur peut être intégré à un site (avec possibilité d'adaptation à la charte graphique) ou hébergé sur google.com. Il existe deux possibilités de recherche : soit de façon classique en l'appliquant à la liste de sites prédéfinie ou bien soit en l'effectuant sur tout le web mais en accordant une plus grande importance aux pages contenues dans la liste. Le fonctionnement est collaboratif en permettant l'enrichissement du moteur par des internautes volontaires ou par ceux qui ont, au préalable, été invités. Ils peuvent annoter le contenu avec des labels (ou tags), ce que le système utilise par la suite pour filtrer une recherche (par exemple, en fonction de la langue).

L'utilisation d'un moteur de recherche personnalisé permet donc d'effectuer sa recherche d'informations sur un corpus de pages restreint, régulièrement mis à jour et dont la qualité et le thème sont validés par l'internaute, un expert ou une communauté ce qui garantit des résultats beaucoup plus pertinents que ceux obtenus avec un moteur généraliste.

#### Sources :

<http://www.brainsfeed.com/archives/1594-recherche-Trois-outils-de-creation-de-moteur-specialise.html>

<http://www.pandia.com/sew/1053-build-your-own-search-engine.html#more-1053>

<http://www.elanceur.org/Articles/EureksterSwickisUnerecher.html>

[http://www.agoravox.fr/article.php3?id\\_article=11045](http://www.agoravox.fr/article.php3?id_article=11045)

<http://veillepme.blogspot.com/tag/swicki>

<http://www.les-infostrategies.com/article/0704289/creer-son-propre-moteur-de-recherche-avec-google-cse-custom-search-engine>

## Extensions Firefox

Le navigateur Mozilla Firefox dispose de nombreuses extensions ce qui permet de lui ajouter de nouvelles fonctionnalités dont certaines peuvent se révéler très intéressantes dans le cas d'une recherche d'information.

Voici une sélection d'utilisations pratiques pour la veille ou la recherche d'information :

- SurfCanyon [<http://www.surfcanyon.com>]
- Cloudlet [<http://www.getcloudlet.com>]
- Outwit [<http://www.outwit.com>]
- LinkChecker [<https://addons.mozilla.org/fr/firefox/addon/532>]



Une fois installé, **Surfcanyon** permet à l'internaute, après une recherche sur un moteur, d'explorer des résultats n'apparaissant pas en première page pour éviter de manquer une information qui pourrait se révéler intéressante. En effet, la plupart des moteurs classent les résultats en fonction de leur popularité sur plusieurs dizaines de pages et une personne ne consulte, en général, que les deux ou trois premières. Basé sur le traitement sémantique des résultats d'une requête et sur l'analyse en temps réel des liens consultés par l'internaute (et donc identifiés comme pertinents), Surfcanyon remonte des résultats se trouvant dans des pages plus profondes en proposant des contenus proches. De recommandations en recommandations, affinant ainsi la recherche sur plusieurs niveaux, Surfcanyon cible de plus en plus finement les résultats pour proposer, en priorité, une sélection la plus pertinente possible. Cette extension peut être utilisée avec Live Search, Google, Yahoo! et Craigslist.

Google™ intelligence économique Rechercher [Reset recommandations](#) [Recherche avancée](#) [Préférences](#)

Rechercher dans :  Web  Pages francophones  Pages : Luxembourg

Web Résultats 1 - 10 sur un total d'environ 733 000 pour **intelligence**

Cloudlet: [Tags](#) [Sites](#) [Net](#) [Off](#) [Donate to support Search Cloudlet](#)

**Intelligence économique** - Wikipédia

L'**intelligence économique** se distingue de l'espionnage **économique** car elle se développe ouvertement et utilise principalement des moyens légaux. ...  
fr.wikipedia.org/wiki/**Intelligence\_économique** - 147k - [En cache](#) - [Pages similaires](#)

Surf Canyon calculations brought forward 3 search results.  
Categories: [Stratégique](#), [Sur](#)

Plus d'**intelligence économique** = moins d'espionnage ! - Blogs Le ... (from page 2)

21 oct 2008 ... certaines affaires impliquant des pseudos cabinets d'« **intelligence économique** » font peser sur cette pratique des soupçons qui n'ont ...  
blogs.lesechos.fr/article.php?id\_article=2246 - [Pages similaires](#)

**Intelligence économique** territoriale - Wikipédia (from page 2)

24 nov 2008 ... L'**intelligence économique** territoriale (IET) est l'application de l'**intelligence économique** à un territoire ou une région. ...  
fr.wikipedia.org/wiki/**Intelligence\_économique\_territoriale** - 21k -  
[En cache](#) - [Pages similaires](#)  
[Autres résultats, domaine fr.wikipedia.org »](#)

**Cloudlet** est une extension qui associe aux résultats d'une recherche sur Google ou Yahoo! un nuage de tags avec une variation de la taille de la police en fonction de l'importance de chaque mot-clé. Un clic permet de l'insérer dans la requête pour l'affiner. Il existe également d'autres onglets qui mettent en évidence les sites les plus pertinents ou bien les domaines de premier niveau pour limiter la recherche à un pays (par exemples : .lu, .be, etc.). Cloudlet propose également le tri des sources sur Google actualités ou des auteurs sur Google Blogs.



Google™ intelligence économique   [Reset recommandations](#)  
[Recherche avancée](#)  
[Préférences](#)

Rechercher dans :  Web  Pages francophones  Pages : Luxembourg

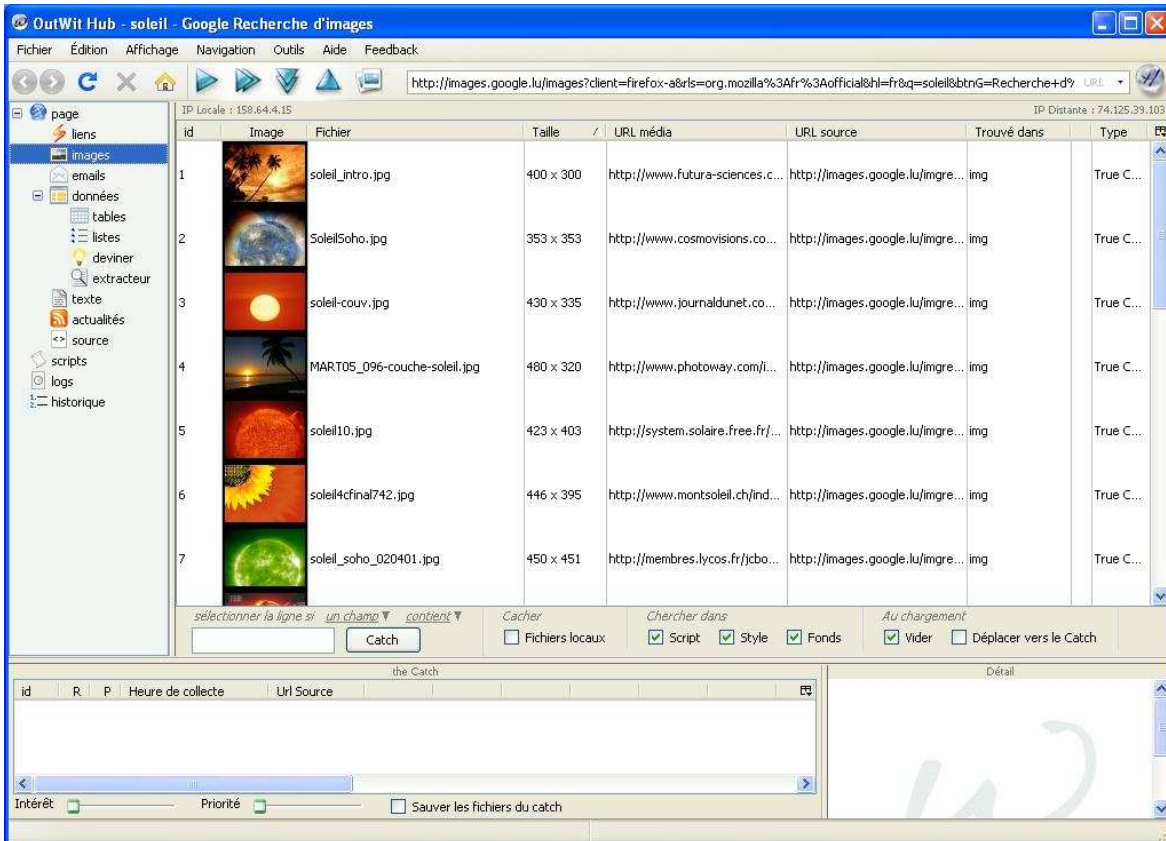
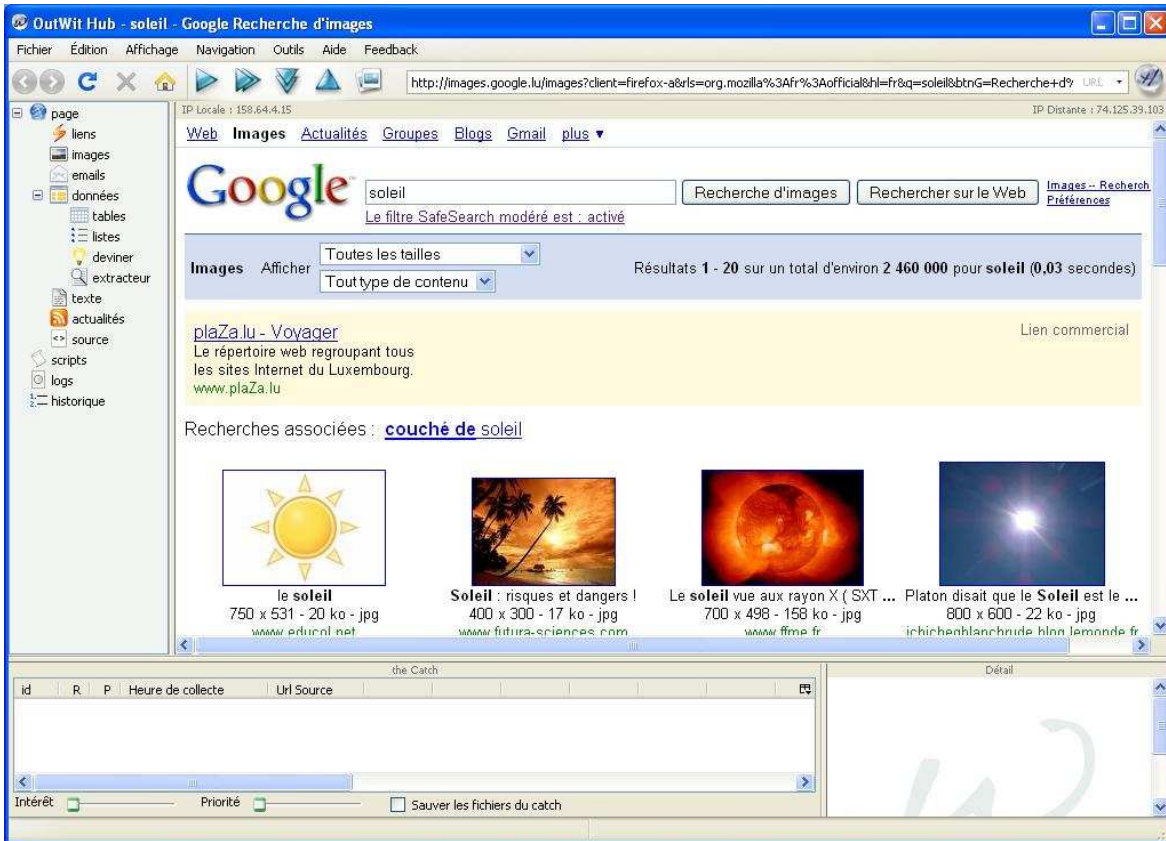
**Web** Résultats 1 - 10 sur un total d'environ 754 000

Cloudlet: **Tags** [Sites](#) [Net](#) [Off](#) [Write a review](#) about [Search Cloudlet](#)

académie accueil acteurs **actualites** alain alp ambition analyse annuaire aux avec ayant besoins bilingue  
 blog car cas chargé christophe collaborateurs collecte commandements commencer communauté  
 consultant cycle dans diffusion distingue doit droit décideurs **démarche** développe echos ecole  
**economique** emploi entreprises français ifie **informations**  
 institut internet offres recherche service site stratégique **veille**

**Intelligence économique** - Wikipédia   
 L'**intelligence économique** se distingue de l'espionnage **économique** car elle se développe  
 ouvertement et utilise principalement des moyens légaux. ...  
[fr.wikipedia.org/wiki/Intelligence\\_économique](http://fr.wikipedia.org/wiki/Intelligence_économique) - 147k - [En cache](#) - [Pages similaires](#)

L'extension **Outwit** est un outil de collecte et de structuration de l'information : il permet de récupérer des données hétéroclites sur des pages web et de les agencer selon leur typologie (images, adresses de courrier électronique, liens, données intégrées dans un tableau, etc.). Il est ensuite beaucoup plus facile de les transférer dans des outils d'analyse afin de les exploiter.





Enfin, **LinkChecker** est un petit outil qui se révèle très pratique lors de la consultation de pages web contenant une grande quantité de liens car il signale par un code de couleur ceux qui sont valides ou brisés ainsi que les redirections.

### Relative Links

- Link to a **sub-folder**
- We all have our **roots**
- Backing out to **another folder**
- Stay in **the current folder**
- **Two slashes** are better than one (thanks Slashdot)

### Absolute Links

- An **absolute link**
- Gotta try a **dead absolute link**
- Don't forget **secure servers**
- **FTP addresses** are addresses too

### Other Links

- Link to a **non-existent domain name**
- A **dead link** and a **good link** to a redirect script
- There's an anchor link just above here that should be ignored.
- Viewing **this directory** is forbidden

### Skipped Links

- Send someone an **email**
- All **the news** that new and approved
- Skip **the 'script**
- It's not the web, it's **your hard drive**
- Login **jetnet style**

LinkChecker at work on a test page  
Image 1 of 2

#### Sources :

<http://veille.aurigance.fr/?p=591>  
<http://www.outilsfroids.net/news/lancement-officiel-de-surfcanyon>  
<https://addons.mozilla.org/fr/firefox/addon/6549>  
[http://www.bulletins-electroniques.com/ti/147\\_02.htm](http://www.bulletins-electroniques.com/ti/147_02.htm)  
<http://www.jeanmariegall.com/2008/10/23/surf-canyon-disponibilit-de-la-nouvelle-version-1101-mj>  
<http://www.activeille.net/index.php/archives/2008/12/20/cloudlet-ajoute-un-nuage-de-mots-cles-aux-resultats-yahoo-et-google>  
<http://sylvaindrapau.com/web/cloudlet-extension-firefox-pour-une-recherche-precise>  
<https://addons.mozilla.org/fr/firefox/addon/9943>  
<http://scidem.emse.fr/index.php?post/2008/12/18/Cloudlet-est-un-plugin-pour-Firefox-qui-va-ajouter-un-nuage-de-tags-aux-r%C3%A9sultats-des-recherches-que-vous-effectuez-sur-Google-et-Yahoo!>  
<http://www.mambro.it/fr/search-cloudlet-comodo-plugin-per-firefox-che-velocizza-le-ricerche-su-google>  
<http://www.sizlopedia.com/2008/12/18/modify-focus-searches-on-google-with-search-cloudlet/fr>  
<http://www.outilsfroids.net/news/outwit-un-plugin-firefox-pour-collecter-et-structurer-les-donnees-du-web>  
<http://www.oezratty.net/wordpress/2008/moissonner-le-web-avec-outwit>  
<http://www.oezratty.net/wordpress/2008/tutorial-outwit-rcupration-dimages>  
<https://addons.mozilla.org/fr/firefox/addon/532>



---

## Traduction de flux RSS

---

S'abonner à des flux RSS permet de suivre facilement l'actualité de sites détectés comme potentiellement intéressants, par exemple pour une veille, d'autant que les nouveaux outils de traduction à la volée permettent d'abolir les limites linguistiques et de juger de la pertinence d'une information même dans une langue étrangère.

Depuis quelques mois, **Google Reader** [[http:// www.google.fr/reader](http://www.google.fr/reader)] intègre dans ses paramètres la possibilité de traduire directement le contenu des flux dans la langue de l'utilisateur sans changer d'interface. L'option, qui peut facilement être décochée pour retrouver le flux initial, est particulièrement utile pour les langues que l'utilisateur ne parle pas du tout (le russe ou le chinois par exemple).

**Mloovi** [<http://mloovi.com>] est un outil en ligne qui traduit les flux RSS de et vers trente-cinq langues. Contrairement à Google Reader, il n'est pas obligatoire de posséder un compte pour procéder à une traduction mais cette option est, bien sur, également disponible. Les sites possédant un flux peuvent, par ailleurs, y intégrer directement Mloovi pour en faciliter la compréhension aux internautes étrangers.

Sources :

<http://www.clubic.com/actualite-177192-google-reader-traduction-flux-rss-volee.html>

<http://veille.aurigance.fr/?p=571>

<http://mloovi.com/pages/about>